Etude

Le TDM dans l'e-infrastructure

Visa TM

Vers une infrastructure de services avancés de text mining

2017 / 2019

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR,
DE LA RECHERCHE
ET DE L'INNOVATION

Focus sur les traitements de fouille de textes dans l'e-infrastructure

Livrable Etude – partie 4

I Dresser une cartographie des outils de traitement de fouille de textes et analyser le dispositif d'animation de la communauté académique dans le cadre d'une e-infrastructure I

Description du Document

Focus sur les traitements de fouille de textes dans l'e-infrastructure

Lot	Etude
Participants	MaIAGE (INRA) INIST (CNRS)
Date de livraison	31/10/2019
Nature : Rapport	Version : 1.0

Contributeurs

	Nom	Organisation
Rédaction	Claire Nédellec	MaIAGE (INRA)
	Frank Arnould	INIST (CNRS)
	Mouhamadou Ba	MaIAGE (INRA)
	Fabienne Kettani	INIST (CNRS)
Coordination	Mouhamadou Ba	MaIAGE (INRA)
Relecture	Robert Bossy	MaIAGE (INRA)
	Claire Nédellec	MaIAGE (INRA)
	Frank Arnould	INIST (CNRS)
	Fabienne Kettani	INIST (CNRS)

SOMMAIRE

Avertissement

Ce document contient des descriptions des résultats du projet Visa TM. Certaines parties peuvent être soumises à des droits de propriété intellectuelle. Avant réutilisation du contenu, il est nécessaire de contacter le consortium pour approbation.

Acronymes et sigles

TDM	Text and Data Mining
OMTD	OpenMinTeD
TAL	Traitement Automatique des Langues
API	Interface de Programmation d'Application
ATALA	Association pour le Traitement Automatique des Langues
AFIA	Association Française pour l'Intelligence Artificielle

Résumé publiable

La compréhension de l'environnement des outils et de leurs différentes fonctions ainsi que la mise en place de cadres d'animation autour des ressources de *text mining* (fouille de textes) relève d'une importance capitale dans le futur dispositif Visa TM. Ce document analyse l'écosystème des outils de traitement et d'assistance du text mining et leur intégration dans l'e-infrastructure OpenMinTeD. Il propose un recensement des outils de text mining et s'intéresse à la sélection des outils dans le cadre d'une e-infrastructure. Le document montre ainsi la diversité et la richesse des outils développés par la recherche et l'ingénierie en text mining. Il propose ensuite une analyse du cadre à mettre en place pour l'animation de la communauté académique qui doit permettre de maintenir le futur dispositif à l'état de l'art par rapport aux outils et ressources.

Introduction

Les outils logiciels de traitement ou d'assistance qui permettent d'assurer les fonctions de fouille de textes occupent une place prépondérante dans la mise en place du dispositif Visa TM, comme c'est déjà le cas dans le projet OpenMinTeD. Nous utilisons dans ce document le terme « outils » pour désigner les composants, systèmes ou plateformes logicielles offrant des fonctions de traitement de fouille de textes ou d'assistance aux tâches de fouille de textes. Dans le projet OpenMinTeD comme dans le projet VisaTM l'objectif est de proposer non seulement un catalogue riche d'outils (et de fonctions) de fouille de textes mais également de garantir leur utilisation à travers un environnement adapté aux utilisateurs. Le dispositif doit permettre l'accès aux ressources bibliographiques, des traitements, la curation, la recherche et la visualisation, etc. Ce besoin requiert des outils de nature et de rôle divers qu'il est nécessaire de comprendre pour la mise en place du futur dispositif Visa TM. Au-delà du rôle des outils, il est aussi important de s'intéresser aux moyens permettant d'identifier les outils de fouille de textes les plus appropriés aux besoins de la future plateforme et aux moyens de les mettre à disposition.

Ce livrable propose ainsi un focus sur les outils logiciels impliqués dans le processus de mise en œuvre de la fouille de textes. Il s'intéresse aux outils de traitement et d'assistance au text mining étudiés dans le cadre du projet OpenMinTeD. Il s'intéresse également au dispositif d'animation permettant de poser un cadre favorable à l'accès et au partage des outils, en particulier, et des ressources, en général. Ce document a pour objet de fournir des pistes d'amélioration de l'infrastructure OpenMinTeD dans le cadre du projet Visa TM.

Nous commençons par un recensement des outils de traitement et d'assistance au text mining disponibles dans l'écosystème d'OpenMinTeD accompagné d'un recensement ouvert des outils de text mining existants par ailleurs. Nous poursuivons par un focus sur la sélection des outils dans le projet OpenMinTeD et des pistes d'amélioration sur cet aspect dans le projet VisaTM. Nous terminons par les enjeux et les leviers autour de l'animation de la communauté académique.

Recensement d'outils de text mining

Le travail de recensement des outils logiciels de fouille de textes a pour but d'identifier ceux qui pourront être utilisés dans le futur dispositif étudié par le projet Visa TM. Il se décline en (1) une revue d'ensemble des outils de l'écosystème d'OpenMinTeD et (2) une revue ouverte d'outils existants par ailleurs.

1.1 Bilan des outils dans l'écosystème d'OpenMinTeD

Face à la fragmentation des solutions de fouille de textes et de données, à l'inadéquation des solutions proposées aux utilisateurs finaux et à l'opacité de la réglementation d'accès et d'utilisation des ressources, le projet OpenMinTeD s'est donné pour ambition "d'offrir une infrastructure ouverte et durable de fouille de textes et de données à travers laquelle des chercheurs de différents domaines peuvent collaborer pour créer, découvrir, partager et réutiliser de manière éclairée et transparente des connaissances d'un large éventail de sources textuelles"[1]. Le projet OpenMinTeD, comme nous allons le voir, a ainsi fait appel à un ensemble d'outils préexistants dans la communauté. Le travail est basé sur les rapports de projet et analyse de l'e-infrastructure OpenMinTeD.

La plateforme du projet OpenMinTeD n'est pas la seule qui propose des services de fouille de textes à la communauté. Il y a plusieurs autres plateformes dans le domaine et dans des domaines proches. Nous donnons ici des exemples significatifs de plateformes qui partagent certains des objectifs du projet OpenMinTeD :

- **Galaxy**[2] est une plateforme dont il existe de nombreuses instances à travers le monde. Elle est développée par la communauté bioinformatique pour offrir des services et des pipelines d'analyse de données. Elle est portée par une large communauté open source à travers le monde et est aujourd'hui utilisée dans plusieurs autres domaines comme le text mining;
- **The Language Application (LAPPS) Grid**[3] est une plateforme de collaboration entre le Department of Computer Science de Vassar College, le Department of Computer Science de Brandeis University, le Language Technology Institute de Carnegie Mellon University, et le Linguistic Data Consortium de University of Pennsylvania. C'est une plateforme basée sur Galaxy, qui permet l'accès à des outils de traitements automatiques des langues et à des ressources linguistiques, et qui permet d'utiliser des pipelines pour construire des applications. LAPPS est une plateforme libre destinée à la communauté scientifique et elle entretient des relations avec des membres de la communauté OpenMinTeD;
- **Alveo Virtual Laboratory**[4] est une plateforme collaborative de recherche en sciences de la communication humaine localisée en Australie. C'est le produit de plusieurs

[1] http://openminted.eu/

[2] https://galaxyproject.org/

[3] https://www.lappsgrid.org/

[4] http://alveo.edu.au/

partenaires à travers le monde (University of Western Sydney, RMIT, Macquarie University, Intersect, University of Melbourne, Australian National University, University of Western Australia, University of Sydney, University of New England, University of Canberra, Flinders University, University of New South Wales, La Trobe University, University of Tasmania, ASSTA, AusNC Inc. NICTA). Le but de la plateforme est d'offrir un accès à une variété de bases de données et à des outils de traitement dans les domaines de la linguistique, du traitement automatique des langues, des sciences de la parole, de la psychologie, du traitement de la musique et de l'acoustique. Comme LAPPS, Alveo utilise également le framework Galaxy. Elle intègre des outils d'autres plateformes telles que NLTK et UIMA;

- **CLARINO Language Analysis Portal (LAP)**[5] est une autre plateforme basée aussi sur Galaxy portée par l'université d'Oslo. Comme les plateformes précédentes, LAP donne accès à des outils de traitement automatique des langues naturelles (NLP) à destination de la communauté des chercheurs. Elle permet l'utilisation de workflows et des calculs intensifs pour des expériences en lien avec le traitement des langues;
- Plusieurs autres plateformes (que nous n'allons pas présenter) existent dans le public comme dans le privé. L'annexe 3 contient une liste de plusieurs plateformes basées sur des frameworks ou outils du marché payants ou libres comme Gate, Weka, R, RapidMiner, Argo, etc.

Par rapport aux autres projets du domaine, la particularité du projet OpenMinTeD réside dans le fait qu'il cherche à mutualiser et à élargir la gamme de services et de fonctions à offrir aux utilisateurs. Il couvre des besoins qui vont au-delà de l'accès aux traitements de text mining à travers des services et des workflows; il cherche aussi à offrir des services transversaux qui contribuent au text mining, par exemple des services d'annotation, des services d'accès aux contenus, de recherche, de visualisation, de stockage, etc. Le projet OpenMinTeD s'applique également à prendre en compte les questions importantes liées à la réglementation autour des ressources de text mining et à favoriser la recherche interdisciplinaire en attirant les chercheurs de différents domaines d'application. Pour y parvenir, la communauté OpenMinTeD a adopté une stratégie qui consiste à s'appuyer sur les technologies déjà existantes dans le domaine pour construire son infrastructure. Cette partie du rapport propose ainsi de faire un bilan de ces outils logiciels de l'écosystème d'OpenMinTeD, de comprendre leurs rôles et d'identifier des points d'amélioration pour le projet Visa TM.

L'écosystème des outils d'OpenMinTeD comporte des outils effectivement intégrés à l'e-infrastructure OpenMinTeD telle que décrite dans le livrable "Architecture OpenMinTeD" mais également des outils non intégrés mais interagissant et utilisés en dehors de l'e-infrastructure OpenMinTeD. Nous n'aborderons pas directement toutes les technologies et ressources, nous allons ici nous limiter aux outils qui assurent les fonctions suivantes dans le processus de text mining : accès au contenu bibliographique, configuration de workflows, traitement de text mining, annotation et curation, recherche et visualisation. Nous avons choisi ces fonctions parce qu'elles représentent les fonctions de base dans le processus de text mining dans le projet OpenMinTeD qui rendent l'infrastructure utilisable par des non-spécialistes.

[5] https://galaxyproject.org/use/lap/

1.1.1 Vue d'ensemble des outils

Les outils logiciels de l'écosystème d'OpenMinTeD se divisent en deux ensembles. Le premier concerne les outils ou plateformes de traitement de text mining qui offrent des composants internes autonomes (qu'on appelle modules) pour assurer des fonctions spécifiques d'extraction d'information (*splitting*, *tagging*, *parsing*, etc.). Le deuxième concerne les outils ou plateformes d'assistance pour assurer des fonctions à procédures complexes telles que l'annotation de corpus, la configuration de workflows, la recherche et visualisation de données ou l'accès aux contenus. Nous ne traitons pas ici de l'architecture logicielle d'OpenMinTeD qui est décrite dans le livrable "Architecture OpenMinTeD".

Les différences entre les deux groupes d'outils portent sur les types de fonctions assurées, la granularité des fonctionnalités et les types d'interfaces offertes. Les outils de traitement de text mining offrent des fonctionnalités qui se présentent sous une forme plus décomposable, c'est-à-dire que les traitements qui les composent sont isolables et utilisables comme modules autonomes. Ces outils offrent une grande flexibilité mais nécessitent des compétences techniques spécifiques. Ils permettent une utilisation de bas niveau, par exemple la ligne de commande ou API en mode *batch*. Ces outils sont intégrés de façon à pouvoir échanger automatiquement des données selon un format et un schéma uniforme. Les outils d'assistance du text mining ont en général une forme monolithique qui permet des fonctionnalités accessibles à travers des interfaces graphiques. Leur utilisation se fait par des interactions avec des utilisateurs. Ils doivent être intégrés de façon à faire profiter pleinement les utilisateurs de leur ergonomie.

Nous nous basons sur cette différenciation pour caractériser et mieux comprendre pour chaque type d'outils le processus d'intégration dans l'e-infrastructure OpenMinTeD ou envisageable dans le futur dispositif Visa TM. Le tableau 1 donne un aperçu des outils que nous classons selon les groupes. La liste n'est pas exhaustive mais elle couvre les fonctions du processus de text mining. Nous nous intéressons dans les sections suivantes à chaque catégorie d'outils.

VUE D'ENSEMBLE DES OUTILS DE TRAITEMENT ET D'ASSISTANCE DU TEXT MINING DANS L'ECOSYSTEME D'OPENMINTED

(*) outils non intégrés actuellement à l'e-infrastructure OpenMinTeD

<table>
<tr><th colspan="3">Fonctions de base par groupes d'outils</th><th>Outils</th></tr>
<tr><td rowspan="5">Traitements de fouille de textes (voir Section 1.2. Outils des traitements du text mining)</td><td>frameworks qui offrent des modules de traitement TDMd, je m</td><td></td><td>> DKPro Core
> UIMA
> AlvisNLP
> Gate
> Argo (*)
> TermSuite
> PubRunner
> VineSum
> MLPLA
> - OGER</td></tr>
<tr><td>modules pour effectuer des traitements TDM unitaires</td><td></td><td>> plus de 400 modules basés sur les outils de l'état de l'art (Stanford NER/Parser,POS-tagger, TermSuite, Tree tagger, GeniaTagger, Yatea, Contes) (*)
> - dont 50 intégrés à OMTD</td></tr>
<tr><td>applications pour un usage direct dans des domaines d'application</td><td></td><td>> 5 applications pour l'agriculture/ biodiversité
> 5 applications pour "Scholarly Communication"
> 3 applications pour les sciences du vivant
> - 1 application pour les sciences sociales</td></tr>
<tr><td rowspan="2">plateformes qui offrent du contenu</td><td>connecteurs qui sont les composants d'accès aux contenus</td><td>> AlvisCrawler (*)
> OpenAire connector
> Core connector
> Istex Connector (*)
> - AgroPortal Connector (*)</td></tr>
<tr><td rowspan="2">Tâches spécifiques d'assistance contribuant au processus de fouille texte (voir Section 1.3. Outils d'assistance du text mining)</td><td>plateformes sources proposant l'accès aux contenus</td><td>> OpenAIRE
> CORE
> ISTEX (*)
> PubMed (*)
> PubAnnotation (*)
> - AgroPortal (*)</td></tr>
<tr><td>outils qui proposent un moteur de workflows</td><td></td><td>> Galaxy
> Argo (*)
> LAPPS Grid (*)
> Gate (*)
> AlvisNLP (*)
> - DKPro Core (*)</td></tr>
</table>

	outils permettant l'édition d'annotations	> AlvisAE > WebAnno (*) > Inception > - AnnoMarket (*)
	outils qui offrent la recherche et la visualisation de ressources TDM	> OMTD Registry

Tableau 1. Vue d'ensemble des outils de traitement et d'assistance du Text Mining dans l'écosystème OpenMinTeD

1.1.2 Outils des traitements du Text Mining

Dans l'écosystème des outils d'OpenMinTeD nous distinguons trois niveaux d'outils chargés des traitements de base du text mining. Ce sont les modules de text mining, les plateformes qui les contiennent et les applications.

- Les modules sont des composants de traitement logiciel qui assurent des fonctions de fouille de textes spécifiques telles que la création de représentations pour des entités de textes, l'analyse syntaxique, la reconnaissance d'entités nommées, le résumé automatique de texte, la classification, etc. Ils assurent également des opérations de type support comme la lecture, la conversion, la sérialisation ou la visualisation de données.
- Les plateformes, appelées aussi *frameworks ou* systèmes, contiennent une bibliothèque de modules et fournissent des mécanismes pour exploiter de manière uniforme les modules. Par exemple, AlvisNLP est un *framework* qui contient un ensemble de modules de fouille de textes. Les modules peuvent être nativement développés dans la plateforme qui les contient ou intégrés au *framework* par la ré-encapsulation d'outils tiers.
- Les applications dans OpenMinTeD correspondent à une configuration opérationnelle d'un ou plusieurs modules qui répond à un besoin réel. On utilise les workflows pour représenter les configurations, notamment quand plusieurs modules doivent fonctionner successivement ou ensemble pour réaliser une tâche.

L'e-infrastructure OpenMinTeD est basée sur ces trois niveaux d'outils pour offrir des traitements de fouille de textes exploitables à l'aide workflows configurables et (ré)utilisables. Pour offrir les traitements de text mining qu'on trouve dans la plateforme, aucun module ni *framework* n'est développé dans le projet 0penMinTeD. Celui-ci réutilise des outils qui existent et fonctionnent.

Plateformes de Text Mining

L'e-infrastructure OpenMinTeD a été développée à partir de quelques plateformes initiales qui sont DKPro Core, UIMA, Gate et AlvisNLP. Elle a accueilli de nouvelles plateformes par un *tender call (*appel à proposition d'outils et de ressources*)*.

- **AlvisNLP,** portée par l'INRA, contient environ 100 modules dont certains sont natifs et d'autres sont une encapsulation d'outils de l'état de l'art. Les modules ont subi le même mode d'intégration dans AlvisNLP et ils partagent une structure de données interne unique. La plateforme AlvisNLP offre les moyens d'utiliser des chaînes de traitement de text mining.
- **DKPro-Core**, portée par l'université de Darmstadt, est basée sur le framework UIMA. Elle est similaire à AlvisNLP dans les éléments qui la composent mais utilise son propre schéma d'intégration et de gestion de ses modules. DKPro-Core propose plusieurs représentations internes des données basées sur un méta-modèle de référence (les CAS UIMA) et offre une API qui permet de programmer des chaînes de traitement.
- **Gate**, portée par l'université de Sheffield, suit la même logique que les deux autres plateformes. Elle intègre des outils tiers de l'état de l'art comme modules. La plateforme Gate ne dispose pas de représentation de données partagées. Elle propose cependant une interface graphique interne pour configurer des workflows.
- plateformes additionnelles apportées via le tender call (**TermSuite**, **PubRunner**, **VineSum**, **UIMA**, **MLPLA**, **AgroPortal** et **OGER**) par des partenaires externes (CNRS Délégation Centre Est, German Research Center for Artificial Intelligence, Universidade de Lisboa, Universitat Pompeu Fabra, BC Cancer Agency, Science For You" NPC - SciFY, SC INEOSOFT S.R.L., Manchester Metropolitan University, University of Zurich, University of Montpellier).

Les plateformes internes DKPro Core, UIMA, Gate et AlvisNLP sont à la base de la prise en charge des workflows. En plus des modules offerts, ce sont les solutions techniques de ces plateformes sur la gestion des modules et l'utilisation des workflows de modules qui ont inspiré les solutions d'OpenMinTeD. Le projet OpenMinTeD a aligné toutes les plateformes avec une solution d'intégration commune et a proposé une réponse au problème d'interopérabilité. Ce travail était nécessaire pour utiliser les plateformes de manière uniforme. Le projet OpenMinTeD a ainsi proposé une liste de spécifications[6] auxquels les plateformes doivent se conformer. Les spécifications définissent l'état et le fonctionnement attendus des plateformes. Elles sont définies de manière consensuelle en analysant les plateformes et les besoins. Elles couvrent des aspects tels que l'encapsulation et la documentation à associer aux plateformes, la mise à disposition des ressources associées aux plateformes, mais également des contraintes pour rendre les plateformes opérationnelles dans le contexte d'OpenMinTeD. Le travail collectif a également abouti à la mise en place d'un schéma de description des ressources, OMTD-SHARE[7], et aux solutions techniques d'interopérabilité, d'encapsulation et de déploiement qui sont décrites dans les *guidelines*

[6] https://github.com/openminted/interoperability-spec

[7] https://guidelines.openminted.eu/the_omtd-share_metadata_schema.html

d'OpenMinTeD[8]. Les plateformes se sont conformées aux spécifications selon les besoins et spécificités de chacun.

Dans le projet VisaTM nous pensons que le futur dispositif gagnera à poursuivre les efforts dans l'intégration des plateformes. Certaines spécifications ont été définies en s'adaptant aux contraintes des plateformes impliquées dans le projet, ce qui représente une faiblesse pour la généralisation et peut être une limite pour l'adoption par des plateformes tierces. Des efforts peuvent être menés pour étendre et implémenter des spécifications plus adaptées, notamment en ce qui concerne la production de la documentation (métadonnées, documentation technique et utilisateur) et l'interopérabilité (échange de données entre plateformes). Des efforts ont été menés par le projet OpenMinTeD pour produire la documentation des plateformes, le projet Visa TM recommande que le futur dispositif poursuive ces efforts en fonction des cibles (utilisateurs, développeurs, experts, etc.). Par ailleurs, étant donné que les solutions proposées pour l'échange de données entre les plateformes couvrent actuellement le niveau syntaxique (la structure des données), le niveau sémantique (la nature des données) pourrait être mieux pris en compte dans certains cas pour faciliter l'intégration. Pour aborder toutes ces questions, une ouverture vers des plateformes externes peut permettre de mieux définir les spécifications et trouver des solutions plus standardisées. Il serait utile d'éviter autant que possible les technologies *ad hoc* et privilégier la réutilisation des technologies existantes même si elles nécessitent des efforts de compréhension et de prise en main non négligeables. Voici quelques exemples d'outils (tirés de l'annexe 3) qui peuvent servir de base pour le transfert de technologies et de pratiques.

- **Apache OpenNLP** : boîte à outils basée sur l'apprentissage automatique pour le traitement de texte en langage naturel. https://opennlp.apache.org/
- **Stanford CoreNLP** : ensemble d'outils d'analyse du langage naturel. http://stanfordnlp.github.io/CoreNLP/
- **NLTK (Natural Language Toolkit)** : plateforme pour la construction de programmes Python pour travailler avec des données en langage humain. http://www.nltk.org/
- **FreeLing** : bibliothèque C++ fournissant des fonctionnalités d'analyse des langues. http://nlp.lsi.upc.edu/freeling/
- **Weka** : collection d'algorithmes d'apprentissage automatique pour les tâches de fouille de données. www.cs.waikato.ac.nz/ml/weka/
- **Scikit-learn** : outils de fouille de données et d'analyse de données. http://scikit-learn.org
- **MLlib** : Bibliothèque d'apprentissage machine d'Apache Spark. http://spark.apache.org/mllib/

Modules de Text Mining

Le projet OpenMinTeD s'est intéressé aux modules parce qu'ils représentent les briques pour les chaînes de traitement de fouille de textes avec pour objectif de permettre la gestion (créer,

[8] https://guidelines.openminted.eu/

exécuter, partager, etc.) de workflows à base de modules unitaires. Le projet OpenMinTeD a ainsi initialement répertorié plus 400 modules qui sont proposés uniquement par les partenaires internes au projet. Ce sont des modules disponibles dans les plateformes Gate, Alvis et DKPro Core/UIMA présentés plus haut. Ce sont assez majoritairement des segmenteurs, parseurs, lemmatiseurs, normalisateurs, détecteurs d'entités nommées, avec une grande proportion de modules qui assurent des fonctions support telles que la lecture, la conversion, la sérialisation et la visualisation des données. Une vingtaine de modules supplémentaires ont été obtenus des plateformes ayant participé au tender call.

Les plus de 400 modules répertoriés présentent une grande variété. Beaucoup de modules existent en plusieurs versions. Plusieurs sont des encapsulations d'outils bien connus dans le domaine et plusieurs reposent sur les mêmes outils de l'état de l'art (CGG Parser[9], Genia tagger[10], Stanford NLP suite[11], TreeTagger[12], Yatea[13]). La grande majorité des modules sont sous licence libre, même si certains (par exemple ceux basés sur TreeTagger) ont des restrictions d'usage. Certains modules sont assez bien documentés même si la documentation reste dans certains cas dispersée et différente d'une plateforme à une autre. Par ailleurs, les modules sont encapsulés de manière différente selon les plateformes et totalisent plus de 50 groupes de modèles de données qui impliquent un grand nombre de formats. Les domaines couverts par les modules sont majoritairement ceux du biomédical avec l'extraction d'information fine de texte concernant des organismes ou organes du vivant (bactérie, cerveau) et des substances ou processus de la biologie moléculaire (ADN, ARN, métabolisme). La plupart des modules s'appliquent à la langue anglaise. Quelques exceptions sont multi-langues. Beaucoup de modules proposent des fonctions génériques de segmentation et de *tokenization* et des fonctions pour la lemmatisation, l'étiquetage morphosyntaxique et l'extraction de dépendances syntaxiques. Quelques modules sont proposés pour l'extraction d'information liée au journalisme (personne, organisation, lieu, date) et pour la classification de documents.

Pour prendre en charge les modules dans OpenMinTeD, la procédure d'intégration évoquée à la section précédente (décrite par les guidelines OpenMinTeD) est utilisée. Elle est basée sur les spécifications mentionnées également dans la partie précédente. Cette procédure repose sur l'idée qui consiste à considérer les composants de traitement de text mining (des plateformes) comme étant des modules autonomes avec pour chaque module une fonction identifiable. Chaque module, avec ses ressources associées, est alors décrit à l'aide du schéma global OMTD-SHARE. L'utilisation d'un système de types basé sur les Cas UIMA[14] est

[9] http://groups.inf.ed.ac.uk/ccg/software.html

[10] http://www.nactem.ac.uk/GENIA/tagger/

[11] https://nlp.stanford.edu/software/

[12] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[13] https://perso.limsi.fr/hamon/YaTeA/

[14] https://dkpro.github.io/dkpro-core/releases/1.9.0/docs/typesystem-reference.html

recommandée pour faciliter l'échange de données entre modules. Il est en plus proposé d'encapsuler chaque module fonctionnel sous l'une des solutions suivantes :

- une image docker qui est disponible dans un dépôt docker;
- des dépendances Maven accessibles via un dépôt Maven;
- ou un service web déployé chez le fournisseur.

Les solutions proposées dans le cadre du projet OpenMinTeD ont permis d'intégrer des modules des plateformes internes. Elles sont utilisées et validées par l'intégration de différents modules externes obtenus par le *tender call*. Elles peuvent cependant être améliorées dans le futur dispositif étudié par Visa TM. En effet, au-delà de la phase d'intégration, les phases d'utilisation effective des modules et de leur maintenance pourraient être mieux prises en charge. Il est aujourd'hui encore difficile pour un utilisateur de comprendre les mécanismes pour exploiter un module individuel ou des modules interconnectés. On remarque par exemple une simplification de l'interface des modules qui, même si elle a l'avantage de faciliter l'usage, a pour inconvénient d'empêcher l'accès à certaines options ou paramètres importants des modules (comme le passage d'une ontologie en paramètre d'un module). Le cycle de vie des modules (intégration, évolution, suivi, etc…) dans la plateforme reste encore mal maîtrisé. Ces questions importantes pourront être prises en charge dans le futur dispositif proposé par Visa TM. Par ailleurs, sur l'ensemble des modules répertoriés dans le projet, une cinquantaine est effectivement intégrée à l'e-infrastructure OpenMinTeD. Des efforts peuvent être faits dans le sens de l'ajout de plus de modules. On note cependant que l'absence d'intégration de modules supplémentaires n'est pas liée à la faisabilité technique. Les fournisseurs partenaires ont ajouté des modules qu'ils ont jugés utiles durant le projet. Les nouveaux modules à ajouter dans le futur dispositif pourront apporter plus de diversité et couvrir plus de domaines. Des modules plus génériques ou populaires (ex: Word2Vec, TensorFlow, Keras, PyTorch, Theano, etc.) qui correspondent à un besoin partagé seraient utiles. Le futur dispositif pourra pour cela exploiter le recensement d'outils présenté au chapitre 1.2.

Applications de Text Mining

Les applications configurées dans OpenMinTeD sont les suivantes (pour plus de détails sur les applications voir les livrables d'OpenMinTeD[15] :

- cinq applications pour l'agriculture et la biodiversité,
- cinq applications pour "Scholarly Communication",
- trois applications pour les sciences du vivant
- une application pour les sciences sociales.
- Et chacun des participants au *tender call* a proposé une application.

On recense actuellement 40 applications proposées sur l'e-infrastructure OpenMinTeD. La spécification de ces applications a suivi une logique qui a permis de les matérialiser sous forme de workflows. Les workflows modélisent et représentent les applications sous forme de

[15] http://openminted.eu/deliverables/

traitements successifs matérialisés par des modules et de données compatibles. La partie exécutable de l'application est rendue disponible en mode *batch* et détachée d'une interface qui permet l'ajout des données, le lancement de l'application et l'accès aux résultats (se référer aux livrables de Visa TM « Application IST », « Application Scientifique ») et d'OpenMinTeD[16](D9.2 « Community driven applications design report ») .

L'e-infrastructure OpenMinTeD propose une interface graphique, partie intégrante du moteur de workflows, pour matérialiser les applications. L'interface permet de configurer des workflows en choisissant les modules dans une bibliothèque et en les interconnectant. Les workflows obtenus sont ensuite enregistrés avec des métadonnées pour devenir des applications qu'un utilisateur peut lancer avec des données préenregistrées sur la plateforme. Les applications les plus simples sont des workflows composés de deux modules où un premier module se charge de la lecture des données et un autre se charge d'appliquer sur les données un traitement de fouille de textes spécifique. Bien que des workflows plus complexes soient envisageables, les applications actuellement dans l'e-infrastructure OpenMinTeD sont en majorité basées sur deux modules. La configuration d'applications reste encore hors de portée des utilisateurs finaux de la plateforme. Les pré-configurations nécessaires pour utiliser une application restent encore complexes. Le futur dispositif pourra prendre en charge ces questions et mieux étudier les solutions à proposer en prenant en compte des utilisateurs ciblés. Il pourrait s'inspirer des solutions à disposition dans les outils comme Galaxy, Argo[17], NextFlow[18], Snakemake[19]

1.1.3 Outils d'assistance au Text Mining

Moteurs de workflows

Les moteurs de workflows permettent d'assurer deux fonctions importantes : la création et l'exécution des workflows. Les outils Argo et Galaxy avaient initialement été étudiés par le projet OpenMinTeD comme moteurs de workflows pour assurer les fonctions citées. Après une étude comparative[20] entre Argo et Galaxy, la solution Galaxy a été finalement retenue. Galaxy convient parce que c'est un outil conçu spécialement pour le traitement intensif de données dans le domaine de la recherche biomédicale. C'est un outil libre qui est soutenu par une large communauté. Il offre des interfaces adaptées aux utilisateurs pour assurer la création et l'exécution de workflows. Il répond aussi aux critères concernant la distribution et la pérennité.

[16] http://openminted.eu/deliverables/
[17] http://argo.nactem.ac.uk/
[18] https://www.nextflow.io/
[19] https://snakemake.readthedocs.io
[20] https://bit.ly/2oJieMP

Le projet OpenMinTeD a levé plusieurs verrous pour l'intégration de Galaxy. Il a réussi à prendre en main Galaxy qui est une solution externe à la communauté de text mining. Il a isolé les parties pertinentes de Galaxy pour l'e-infrastructure OpenMinTeD. Il a également adapté ses ressources pour les rendre compatibles à Galaxy. Deux instances de Galaxy ont ainsi été intégrées, l'une pour l'exécution de workflows et l'autre pour l'édition de workflows. L'API Blend4J est utilisée pour communiquer avec les deux instances. Les modules à base de workflows sont encapsulés sous le principe de Galaxy avec les *wrappers* spéciaux de Galaxy.

L'intégration de Galaxy peut se poursuivre dans le futur dispositif étudié par le projet Visa TM, notamment pour résoudre les problèmes de passage à l'échelle. Comme mentionné dans la partie précédente, l'adéquation de l'éditeur de workflows aux utilisateurs devra être mieux analysée. L'éditeur actuel ne permet pas une expressivité satisfaisante pour les ingénieurs et n'est pas assez simple pour un utilisateur final. La construction de workflows repose aujourd'hui sur un intermédiaire humain qui configure les workflows pour les utilisateurs finaux. Un bon niveau d'abstraction et de simplification permettrait aux utilisateurs finaux de le prendre en main facilement. Par ailleurs, il faudrait prendre en compte des besoins spécifiques de certains outils, par exemple la composition de réseaux de neurones. Sans être exhaustifs, nous listons quelques solutions en rapport avec la gestion de workflows qui peuvent inspirer ou être exploités par le futur dispositif de Visa TM:

- **CLARINO Language Analysis Portal**, https://galaxyproject.org/use/lap/
- **LAPPS Grid**, https://www.lappsgrid.org
- **Alveo**, https://galaxyproject.org/use/alveo/
- **Snakemake**, système de gestion de workflows destiné à l'analyse de données reproductibles et évolutives. https://snakemake.readthedocs.io/en/stable/#
- **Keppler** : application pour des workflows scientifiques. https://kepler-project.org/
- **TextGrid** : infrastructure pour un environnement de recherche virtuel en sciences humaines. https://textgrid.de/en
- **WebLicht** : environnement d'exécution pour l'annotation automatique des corpus de texte. http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page
- **DKPro Core** : collection de composants logiciels pour le traitement du langage naturel (NLP) basée sur le framework Apache UIMA. https://dkpro.github.io/dkpro-core/
- **NewsReader** : index d'événements structurés de grands volumes de données financières et économiques pour la prise de décision dans diverses langues. http://www.newsreader-project.eu/the-project
- **TextFlows** : plateforme de composition, d'exécution et de partage de workflows en fouille de textes et de traitement du langage naturel. http://textflows.org/
- **Panacea** : plateforme d'annotation normalisée et d'acquisition de ressources langagières pour les technologies du langage humain. http://www.panacea-lr.eu/
- **U-Compare** : système intégré de fouille de textes / langage naturel basé sur le framework UIMA. http://u-compare.org/
- **Taverna** : système *open source* et générique pour la conception et l'exécution de workflows. http://www.taverna.org.uk/

- **Heart of Gold** : architecture middleware pour l'intégration de composants de traitement de langage naturel http://heartofgold.dfki.de/
- **Vistrails** : système open source de gestion de workflows scientifiques et de la provenance. https://www.vistrails.org

Systèmes d'édition d'annotations

Les éditeurs d'annotations permettent l'annotation manuelle de corpus de documents par des experts. Les éditeurs permettent de créer des corpus annotés utilisés, notamment dans certaines méthodes d'extraction d'information. Les corpus annotés sont des ressources très précieuses car coûteuses à produire. Les outils proposés dans OpenMinTeD pour assurer la fonction d'édition d'annotations sont AlvisAE[21], WebAnno[22] et Inception[23]. AlvisAE est un éditeur d'annotation développé à l'INRA pour gérer les campagnes d'annotation avec des experts. Les éditeurs WebAnno et Inception sont développés par UKP-TUDA. Les trois éditeurs offrent des interfaces graphiques aux utilisateurs. AlvisAE offre une grande expressivité aux annotateurs et gère le cycle de vie des campagnes d'annotation (ajout, assignation, annotation, adjudication). Il permet en plus d'utiliser des ontologies pour l'annotation et la possibilité d'éditer à la volée une ontologie. WebAnno et Inception offrent moins de fonctionnalités qu'AlvisAE mais ces deux outils ont l'avantage de bénéficier régulièrement d'extensions.

Pour la prise en charge des éditeurs d'annotation, le projet OpenMinTeD a proposé le protocole d'échange de données AERO[24]. Ce protocole permet, à partir d'un éditeur, de tirer les données à annoter d'OpenMinTeD et d'envoyer les données annotées vers OpenMinTeD. Les éditeurs qui implémentent le protocole restent utilisables avec toutes leurs fonctionnalités à l'extérieur de la plateforme. Actuellement, le protocole est implémenté par AlvisAE et WebAnno. Les efforts doivent cependant se poursuivre par une extension du protocole et la mise en place d'interfaces dans l'e-infrastructure OpenMinTeD et dans les éditeurs pour sécuriser et simplifier les échanges d'information liées aux annotations.

Moteurs de recherche

Les moteurs de recherche offrent les fonctions pour accéder et consulter les différentes ressources. L'e-infrastructure OpenMinTeD contient plusieurs types de ressources qui couvrent des articles, des dictionnaires, des ontologies, des modèles, etc., et des outils logiciels (modules, workflows, applications). Pour permettre la recherche à travers ces ressources, la plateforme OpenMinTeD propose des fonctionnalités de base pour la recherche et la visualisation à travers *OMTD register*. Des systèmes externes tels qu'AlvisIR[25], qui est un moteur de recherche sémantique proposée par l'INRA, sont utilisés par des applications de la

[21] Papazian, et al 2012
[22] https://webanno.github.io
[23] https://inception-project.github.io
[24] https://github.com/openminted/omtd-aero
[25] https://github.com/Bibliome/alvisir

plateforme pour compléter les besoins notamment pour la recherche et la visualisation à travers des corpus annotés. La solution d'OpenMinTeD a l'avantage de répondre directement aux besoins du projet, elle reste cependant moins riche en fonction que les systèmes spécialisés tels que AlvisIR. Aucune dynamique n'a été réellement initiée pour intégrer les systèmes externes de recherche et de visualisation à OpenMinTeD. Cette question pourrait être prise en charge dans le futur dispositif Visa TM. La recherche et visualisation de données sont incontournables pour la plupart des utilisateurs, elles mériteraient une attention particulière.

Plateformes de contenu

Les plateformes de contenu gèrent et fournissent le contenu de la littérature scientifique. OpenAire[26] et CORE[27], portés par des partenaires internes (ARC et OU), sont les fournisseurs principaux de l'e-infrastructure OpenMinTeD. OpenAire, comme CORE, gère une quantité importante de publications et tous deux fournissent des services pour l'accès aux documents. L'e-infrastructure OpenMinTeD a été ouverte à d'autres fournisseurs tels que ISTEX[28] qui est une plateforme portée par l'INIST/CNRS qui gère plusieurs millions de documents de la littérature scientifique pour l'enseignement supérieur et la recherche en France. Le projet OpenMinTeD élargit l'accès aux ressources de tiers en facilitant la connexion de nouvelles plateformes. Il propose une API de base (content-connector-api[29]) pour intégrer des plateformes externes. L'API définit les opérations qu'une plateforme de contenus doit implémenter. Ces opérations sont de simples modules pour la recherche et le transfert (à distance) de contenus. L'API est accompagnée de guidelines qui décrivent le processus pour connecter une plateforme de contenu.

Les partenaires fournisseurs d'OpenAIRE, CORE, ISTEX et AgroPortal[30] ont implémenté l'API d'OpenMinTeD. Si l'intégration a été finalisée pour OpenAIRE et CORE, elle ne l'est que partiellement pour ISTEX et AgroPortal (voir le livrable « Bilan technique »). Il peut mieux intégrer le travail réalisé sur la compatibilité des licences d'accès aux données[31]. Les solutions peuvent être plus élaborées pour couvrir plus de bibliothèques spécialisées de la littérature scientifique et des plateformes de partage de corpus annotés telles que PubAnnotation[32].

1.2 Recensement ouvert d'outils

Dans le cadre du volet Étude du projet VisaTM, nous avons réalisé un recensement d'outils de fouille de textes. Le premier objectif de ce travail est de faire le point sur les moyens logiciels

[26] https://www.openaire.eu/
[27] https://core.ac.uk/
[28] https://www.istex.fr/
[29] https://github.com/openminted/content-connector-api
[30] http://agroportal.lirmm.fr
[31] https://openminted.github.io/releases/license-matrix/
[32] http://pubannotation.org

issus de l'Intelligence artificielle, du Traitement automatique du langage naturel (TAL) et des statistiques qui sont disponibles pour la découverte de connaissances à partir du traitement informatique de corpus textuels. Le second objectif est de disposer d'un référentiel d'outils pour l'enrichissement de la plateforme de fouille de textes dont le projet VisaTM évalue la faisabilité. Autrement dit, il s'agit de rassembler en un point unique des informations souvent disponibles sur des sources dispersées. Bien évidemment, ce recensement n'a pas prétention à être exhaustif. Il s'agit d'une liste ouverte qui est complétée régulièrement[33].

1.2.1 Méthode

Plusieurs types de sources ont servi au recensement :

- Répertoires d'outils sur le web ;
  - MultiTAL
    http://multital.inalco.fr/
  - TAPoR
    http://tapor.ca/home
  - KDnuggets
    https://www.kdnuggets.com/software/text.html
  - Predictive analytics today
    https://www.predictiveanalyticstoday.com/top-software-for-text-analysis-text-mining-text-analytics/
  - Wikipedia
    Outils de TAL
    (https://en.wikipedia.org/wiki/Outline_of_natural_language_processing#Natural_language_processing_toolkits) ;
    Apprentissage profond
    (https://en.wikipedia.org/wiki/Comparison_of_deep-learning_software) ;
  - Lingua Greca
    (https://linguagreca.com/blog/2018/03/nine-terminology-extraction-tools-are-they-useful-for-translators/) ;
- Tutoriels en TAL et fouille de textes ;
- Articles scientifiques et communications en TAL et fouilles de textes ;
- Liste des composants OpenMinTeD
  - https://www.futuretdm.eu/tool-list/ ;
  - https://openminted.github.io/releases/interop-spec/1.0.0/components/ ;
- Expertise des auteurs.

La sélection des outils a mis l'accent sur les outils purement TAL et fouille de textes. N'ont donc pas été retenus, par exemple, les environnements de développement, les langages de

[33] https://visatm.inist.fr/2019/04/11/recensement-doutils-de-fouille-de-textes/

programmation, les outils OCR, de conversion de fichiers, d'encodage de textes, etc. Pour les outils généralistes, comme R ou RapidMiner, les packages et extensions spécialisés dans la fouille de textes ne sont pas indiqués (sauf s'ils existent par ailleurs comme outils autonomes). Pour R, une carte mentale des packages disponibles est accessible à l'adresse suivante :

> http://www.bnosac.be/index.php/blog/87-an-overview-of-the-nlp-ecosystem-in-r-nlproc-textasdata

Dans le document de recensement à l'annexe 3, les outils sont présentés de la manière suivante :

Nom de l'outil ;

> **Description** : description générale de l'outil, sauf exception, de son site web ;
> **Licence de l'outil** ;
> **Tâche(s)** : tâches principales réalisées par l'outil, catégorisées, sauf exception, à l'aide de l'ontologie OpenMinTeD
  - (http://www.meta-share.org/ontologies/omtd-share/omtd-share-ontology.owl/documentation/doc/index-en.html) ;
  - Le terme d'annotation dans le document renvoie à l'annotation sémantique ;
> **Source** : lien vers le site web de l'outil ;
> **OMTD** : présence ou non de l'outil dans l'e-infrastructure OpenMinTeD; Les outils présents dans la liste https://openminted.github.io/releases/interop-spec/1.0.0/components/ ont été considérés comme « présents », même quand leur intégration dans la plateforme n'est pas encore effective ;
> **Pays** : pays d'origine du développement de l'outil.

1.2.2 Résultats

Actuellement, 300 outils de TAL et de fouille de textes sont recensés. La liste est disponible dans l'annexe 3. La figure 1 présente la répartition des outils en fonction du type de licence. Les outils qui proposent des licences mixtes (libres et commerciales en fonction des fonctionnalités mises à disposition) et ceux dont la licence n'a pas pu être déterminée avec exactitude sont regroupés dans la catégorie Autres. La majorité des outils recensés sont sous licence libre.

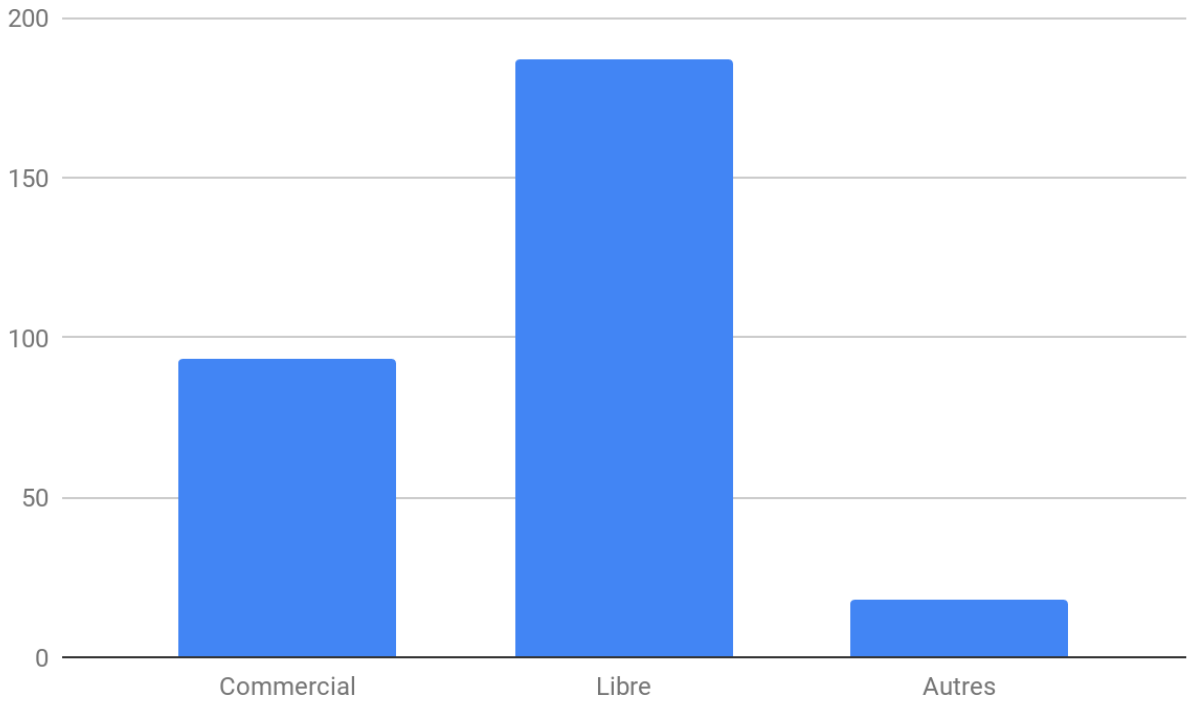


Figure 1. Licence des outils

La figure 2 présente la répartition en pourcentage des outils en fonction de leur origine géographique (du moins pour celles qui ont pu être identifiées). Les pays ont été regroupés dans cinq catégories : Amérique du Nord, Asie, Europe, Océanie et International. Cette dernière regroupe des outils issus de la collaboration d'un pays de l'une des catégories restantes avec au moins un autre pays de ces mêmes catégories (par exemple, États-Unis et Israël). Les applications françaises représentent 17 % des outils recensés. On note une répartition équitable entre l'Amérique du Nord et l'Europe (44 % et 46 %, respectivement). Ces deux régions géographiques produisent l'essentiel des outils. Peu d'outils sont le résultat d'une collaboration internationale.

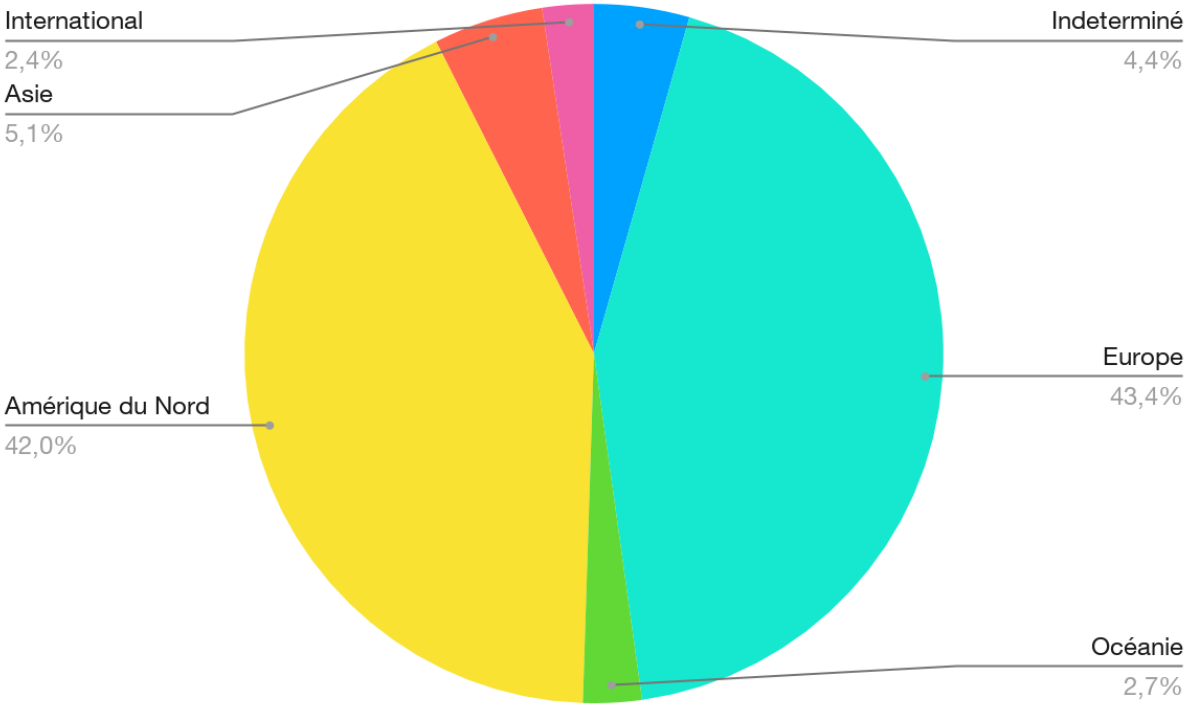


Figure 2. Répartition des outils par origine géographique

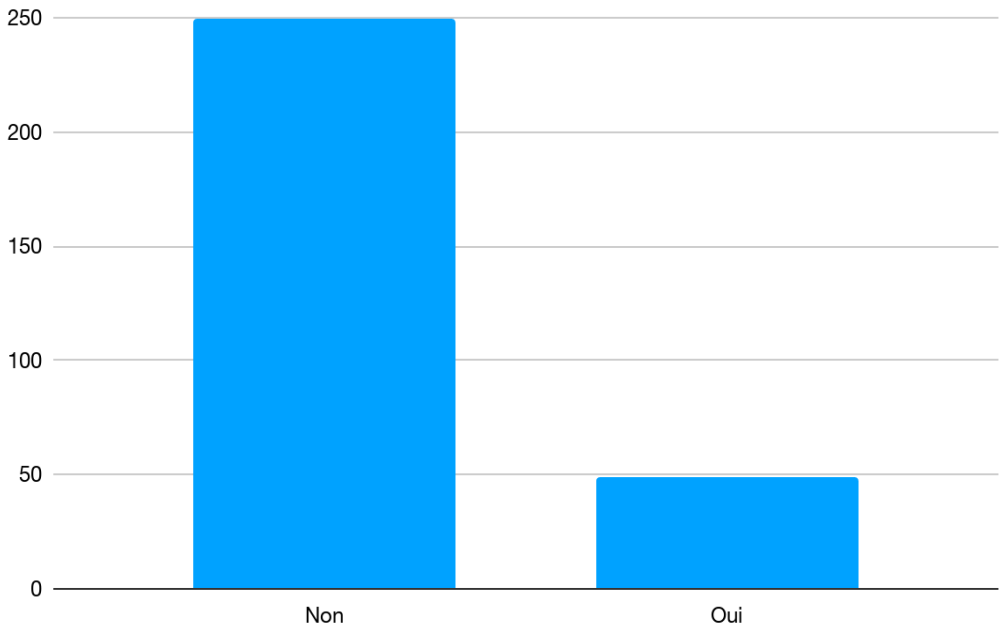


Figure 3. Présence ou non des outils dans OpenMinTeD

La figure 3 représente la répartition des outils en fonction de leur présence ou absence dans l'e-infrastructure OpenMinTeD. La majorité d'entre eux y sont évidemment absents.

1.2.3 Conclusions

La figure 4 représente la répartition des composants OpenMinTeD en fonction de leurs tâches principales : TAL, fouille de textes et de données (TDM), autres (conversion de fichiers, outils de développement, etc.). Cette répartition a été réalisée à partir de la liste des composants rapportés dans la liste disponible à l'adresse suivante : https://openminted.github.io/releases/interop-spec/1.0.0/components/.

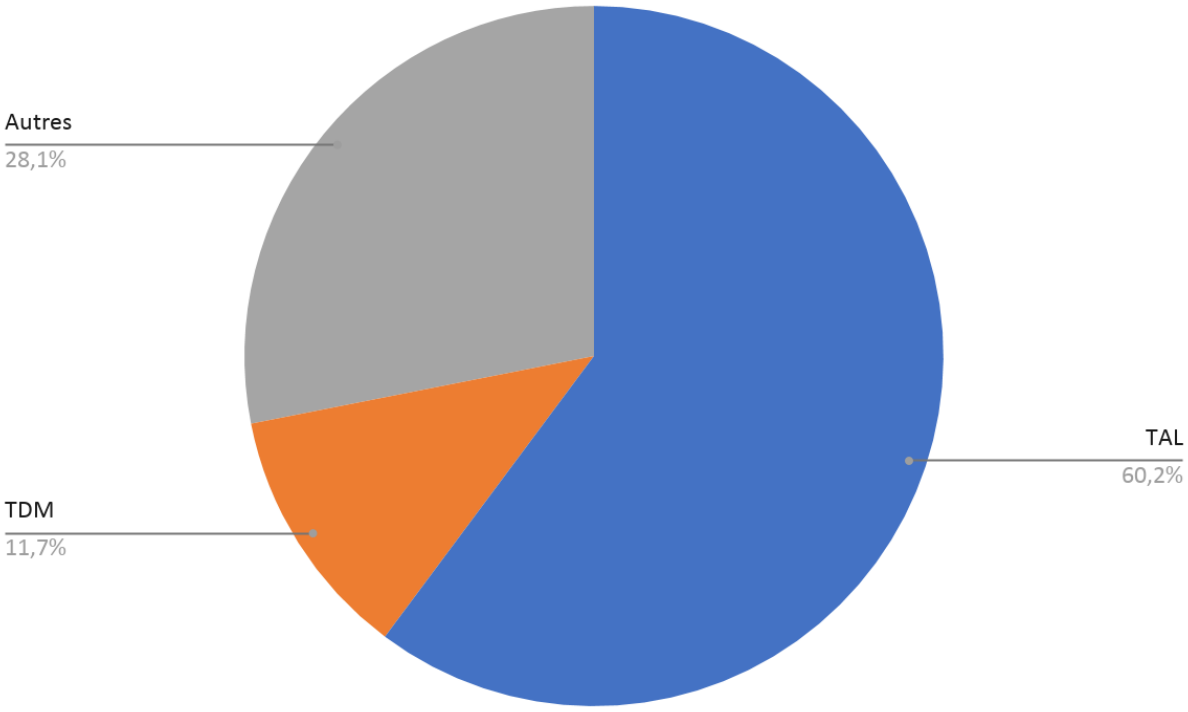


Figure 4. Répartition des composants OMTD en fonction de leur tâche principale

Le graphique indique que les outils TAL, c'est-à-dire, d'analyse linguistique des textes, sont bien représentés dans la plateforme (tokenisation, étiquetage morpho-syntaxique, parsing, etc.). En revanche, les outils de TDM (compris comme des applications pour la découverte de connaissances dans des masses de données textuelles) sont moins bien présents. L'effort d'enrichissement de la plateforme est peut-être à réaliser sur cet aspect.

Sélection d'outils

La partie précédente a mis l'accent sur le recensement des outils de fouille de textes. Elle présente les outils logiciels présents dans OpenMinTeD et a ouvert la voie à la découverte et à l'utilisation d'outils supplémentaires. Les outils existants dans les communautés vont continuer à croître, à s'enrichir et à se diversifier. Il est alors important de poser un cadre qui puisse permettre une prise en charge efficace de la sélection des outils. L'analyse du processus de sélection des outils par le projet OpenMinTeD et de la prise en compte de cette question dans le futur dispositif de VisaTM est l'objet de cette partie.

2.1 Objectifs

Face au très grand nombre d'outils et à leur diversité, il est difficile de repérer et de choisir les outils adaptés à mettre à disposition des utilisateurs de la plateforme. Un certain nombre de critères sont à prendre en compte pour arriver à une sélection rigoureuse des outils. Certains de ces critères sont liés à la localisation des outils (par les réseaux, communautés) et à la collecte (appels à proposition, collecte proactive), d'autres à la conformité des outils à différents niveaux stratégique, technique, fonctionnel, etc. S'ajoutent à cela les ressources matérielles et humaines à mobiliser ainsi que l'organisation à mettre en place pour s'assurer d'un processus fonctionnel.

Tout cela demande une gestion organisée de la sélection des outils à mettre à disposition au niveau de la plateforme. Dans des plateformes comme OpenMinTeD où on espère un très grand nombre d'outils, cette gestion est nécessaire pour offrir des outils de qualité de manière transparente et durable. Dans la suite, nous allons voir comment le projet OpenMinTeD a procédé pour attirer de nouveaux outils sur la plateforme. Nous donnons ensuite quelques pistes d'amélioration concernant la sélection des outils futurs.

2.2 Bilan de la sélection d'outils dans OpenMinTeD

La sélection d'outils dans OpenMinTeD s'est faite de manière ponctuelle lors d'un appel à proposition d'outils (*tender call)*. OpenMinTeD a sélectionné d'autres outils durant les phases de développement de la plateforme, mais cela s'est fait sur la base des outils répertoriés au montage du projet. Le but de cette partie est de revenir sur la procédure de sélection utilisée durant le *tender call* (voir D2.5 "Open Call Programme Implementation Report"). Elle décrit ainsi la procédure et les critères utilisés pour sélectionner les outils.

2.2.1 Procédures de sélection

Le projet OpenMinTeD a évalué et sélectionné des outils à partir des propositions de fournisseurs ayant répondu au *tender call*. La communauté OpenMinTeD s'est basée sur une liste de critères définis à cet effet pour examiner et comparer les propositions ayant répondu

à l'appel. Un comité a conduit le processus d'évaluation qui s'est déroulé en deux étapes successives :

- Le comité technique réalise l'évaluation technique initiale de chaque proposition. Cela a consisté à vérifier si la proposition remplit les spécifications techniques et à identifier les propositions qui ne sont pas éligibles techniquement. Le comité technique s'est ainsi basé sur une *checklist* et a considéré spécialement les éléments techniques suivants : vérifier si l'outil proposé est compatible avec les systèmes UIMA et GATE et s'il est possible de l'encapsuler selon les principes d'OpenMinTeD.
- Chaque membre du comité d'évaluation vérifie chaque proposition en fonction de chaque critère de la liste de critères d'OpenMinTeD. Il attribue une note à chaque proposition sur la base du nombre de points associé. L'évaluation se termine ensuite par un classement des propositions selon les moyennes des scores.

La procédure admet une étape préliminaire qui assure la vérification de la faisabilité technique des propositions. Cette étape peut être généralisée et complétée par la vérification d'autres aspects importants aux niveaux stratégique, opérationnel ou prévisionnel, par exemple. Plus important, la procédure étant destinée à l'évaluation de proposition d'outils, elle pourrait être généralisée pour la sélection et l'évaluation d'outils qui arriveront par des canaux différents (*tender calls*, recherche proactive, etc.).

2.2.2 Critères d'évaluation

Le projet OpenMinTeD a défini une liste de critères et un score associé à chaque critère pour la sélection des outils proposés dans le cadre d'un *tender call*. Les critères sont proposés par les membres du projet selon les sujets. Ils ont été soumis à l'appréciation d'un comité habilité à analyser, adapter, accepter ou refuser un critère. L'annexe 1 présente la liste des critères. Les critères ont été utilisés selon le processus de sélection décrit à la section précédente.

Les critères ont été très utiles pour l'évaluation et la sélection des propositions d'outils du *tender call*. Il faut poursuivre les efforts pour utiliser les critères d'OpenMinTeD pour la sélection de manière générale. Les critères d'OpenMinTeD sont conçus pour répondre à un besoin ponctuel (les *tender call*). Ils ont été davantage utilisés pour évaluer des propositions d'outils déjà collectés que pour une collecte d'outils.

2.3 Sélection d'outils dans Visa TM

2.3.1 Procédure de sélection

La procédure de sélection d'outils adoptée par le projet OpenMinTeD est une procédure définie de manière ponctuelle et destinée à des outils proposés durant un *tender call*. Dans la future plateforme, il peut convenir d'adopter une approche plus générale et de définir une politique globale de sélection des outils. Pour cela plusieurs aspects vont entrer en ligne de compte. Parmi ces aspects, la stratégie de collecte d'outils : elle peut être réalisée à travers

des appels à proposition ponctuels ou permanents ou à travers une recherche proactive d'outils pris en charge par les acteurs internes. Selon les cas la politique à mettre en place en amont et en aval doit être adaptée. Par exemple, lorsqu'il s'agit d'un appel, on peut considérer qu'il faut impliquer les communautés à travers une animation et mettre en place les mécanismes permettant d'accueillir les propositions. S'il s'agit d'une recherche proactive, on peut imaginer la mise en place et l'enrichissement continu de référentiels d'outils accompagnés d'une politique de veille. Lorsqu'on dispose de référentiels ou de propositions d'outils, il faut ensuite savoir quels sont les outils qui satisfont aux besoins avant une éventuelle mise à disposition sur la plateforme. Les étapes que nous venons de décrire impliquent des acteurs, des ressources et une organisation qui reste à mettre en place. La gestion du cycle de vie des outils au niveau de la plateforme n'est pas non plus évidente compte tenu de la dynamique autour des outils qui naissent, grandissent, vieillissent, changent, disparaissent, etc. Le projet VisaTM n'est pas en mesure de formaliser de manière précise la sélection et l'intégration d'outils à cet stade mais peut d'ores et déjà s'appuyer sur l'expérience du projet OpenMinTeD et commencer à identifier et améliorer les éléments importants de ce processus. On identifie déjà pour cela les spécifications du *tender call*, les critères de sélection ainsi que d'autres ressources (état de l'art des outils, sources des outils) et les acteurs clés du processus de sélection d'outils (fournisseurs, intégrateurs, utilisateurs, etc.).

2.3.2 Critères de sélection pour Visa TM

Les critères de sélection doivent permettre de caractériser les outils de fouille de textes et de données afin de mieux cibler et sélectionner les outils qui sont les plus adaptés. Les critères définissent une liste de propriétés pertinentes à considérer lorsqu'on collecte des informations sur les outils lors d'une campagne de sélection ou d'une collecte d'outils. Les critères proposés dans l'annexe 2 sont un complément aux critères d'OpenMinTeD. Ils sont définis avec l'idée de considérer la sélection et l'évaluation des outils de manière moins ponctuelle. Ils considèrent les mêmes aspects que dans OpenMinTeD avec une tentative de catégorisation et une précision du rôle de chaque critère. Contrairement aux critères d'OpenMinTeD, ces critères n'intègrent pas un poids prédéfini, le poids à donner à chaque critère est une variable à adapter aux situations.

La liste de critères représente une initiative qui est amenée à être complétée et enrichie. Cette initiative devra s'enrichir avec une plateforme opérationnelle et permettre de réfléchir sur le processus d'acquisition d'outils de manière organisée et durable en s'appuyant sur les communautés.

Animation de la communauté TDM

Les sections précédentes ont mis en évidence la diversité et la richesse des outils développés par la recherche et l'ingénierie en TAL et text mining. Cette production est très évolutive et maintenir la future plateforme à l'état de l'art nécessite plus qu'une veille suivie, elle nécessite une participation proactive au transfert des résultats de la recherche vers des services. La communauté de recherche et de développement en text mining est une composante importante du dispositif de l'e-infrastructure. Elle est à la fois (1) fournisseuse de composants nécessaires, mais aussi (2) utilisatrice des services. Pour que la bibliothèque logicielle de la plateforme et les services proposés soient également appropriés aux deux types de besoins, il faut mettre en place une animation efficace et souple partant des besoins et des stratégies des acteurs et reposant sur les structures et les moyens existants. Dans la suite du document, nous nous concentrons sur la communauté académique spécialisée en fouille de textes dont nous allons esquisser les contours dans un premier temps et nous verrons ensuite quels peuvent être les enjeux et les leviers d'une dynamisation de son animation.

3.1 Le paysage français des laboratoires de recherche autour du TAL et de la fouille de textes

Nous avons réalisé un recensement des différents laboratoires académiques impliqués dans des activités autour du traitement automatique du langage ou de la fouille de textes en France. Il est consultable dans les annexes de ce livrable sous le titre « Recensement de laboratoires spécialisés en fouille de textes et TAL ». Ce recensement, qui n'a de loin pas la prétention d'être exhaustif, repose sur les réseaux et connaissances personnelles des partenaires du projet VisaTM, sur des recherches en ligne, sur la consultation de sites d'associations autour de ces thématiques (telles que l'ATALA : Association pour le Traitement Automatique des Langues). A noter que l'AFIA (Association Française pour l'Intelligence Artificielle)[34] prépare un numéro thématique de son bulletin pour janvier 2020 sur le sujet.

Nous avons reporté la répartition de ces laboratoires sur une carte de France et le résultat (Figure 5) met en lumière une couverture assez homogène du territoire français avec des concentrations plus prononcées dans certaines régions françaises: Ile de France évidemment mais aussi Grand Est, Auvergne Rhône-Alpes, Occitanie et Provence-Alpes-Côte d'Azur correspondant à de grands centres universitaires. L'ensemble de ces laboratoires ne couvre pas nécessairement les mêmes thématiques de recherche : certains sont plus tournés vers le TAL et l'apprentissage automatique, d'autres plutôt vers les statistiques. Leurs domaines d'application peuvent également être très différents et on notera que tous ont des coopérations en cours avec d'autres laboratoires situés en Europe (Allemagne, Italie, Luxembourg, Royaume-Uni, Irlande, Espagne, Suisse, Belgique, Pays- Bas, Roumanie, Pologne,

[34] https://afia.asso.fr/

Hongrie, Bulgarie, Grèce, Suède, Danemark, Islande, Finlande,) ou à l'international (Etats-Unis, Canada, Japon, Brésil, Mexique, Russie, Tunisie).

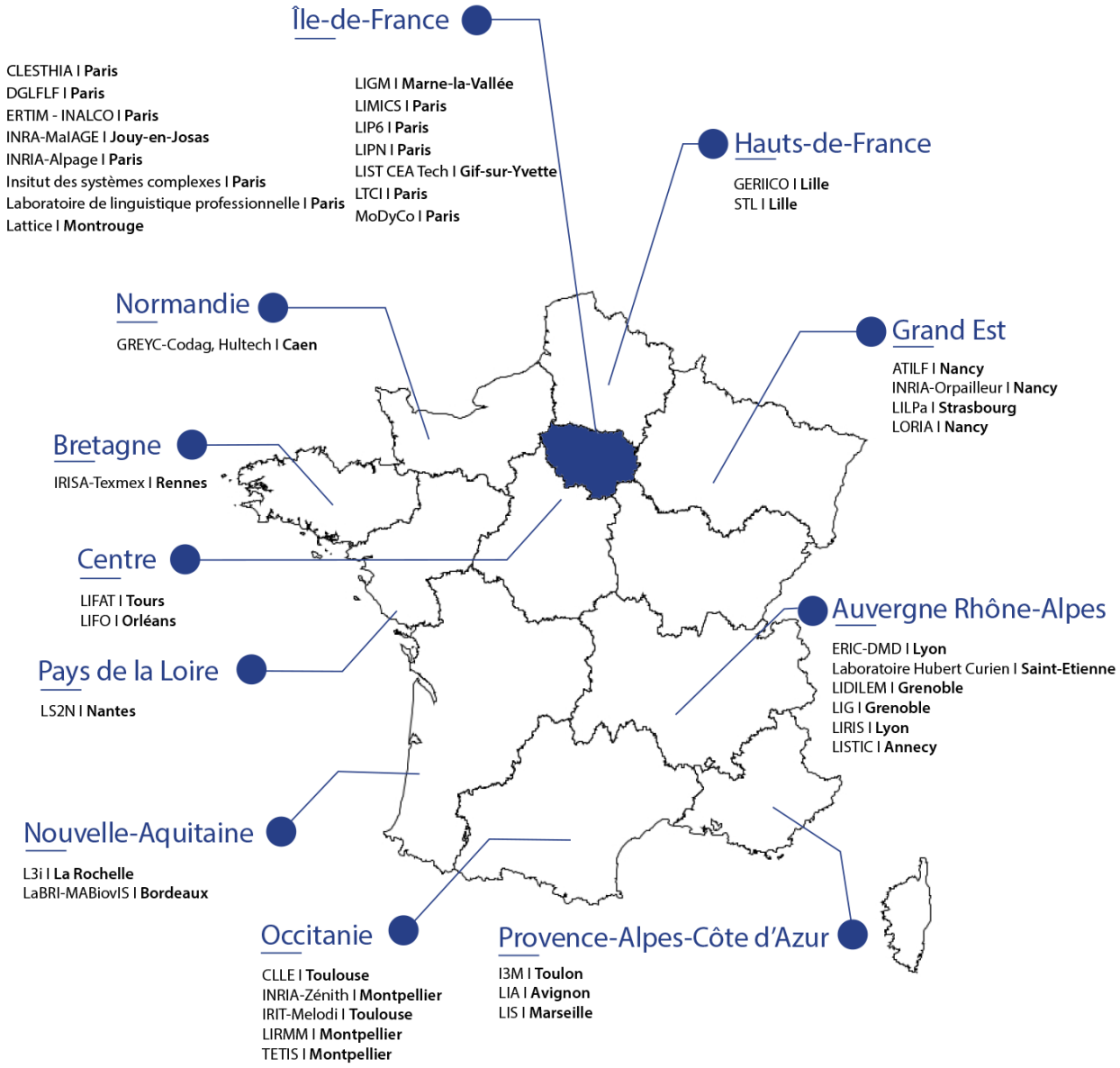


Figure 5. Répartition géographique française de laboratoires de fouille de textes et TAL

Le domaine de la recherche en fouille de textes est bien couvert au niveau national mais les résultats de ces recherches, bien que présentés dans des journaux, colloques et congrès, ne font pas nécessairement l'objet aujourd'hui de coopérations et partages systématiques ce qui met en lumière tout l'intérêt d'actions d'échanges, de valorisation et de transfert autour de communautés de pratiques.

3.2 Une dynamique d'animation des communautés de recherche autour du TAL et de la fouille de textes

3.2.1 Objectifs et freins

L'enjeu est pour la future plateforme de mettre en place un cercle vertueux de telle sorte que l'utilisation des services de la plateforme soit incitative pour le transfert des composants logiciels et corpus, ce qui rendra à leur tour les services plus attractifs. Les services attendus par les chercheurs et ingénieurs en fouille de textes diffèrent en partie des services aux autres communautés de recherche. La question des solutions informatiques et d'accompagnement humain pour répondre à ces besoins spécifiques reste en partie à traiter. Par exemple, l'expérimentation et l'évaluation jouent un rôle très important et elles devraient être facilitées par une architecture et des modes d'exécution et d'analyse des processus adaptés aux besoins. L'analyse de ces besoins est détaillée dans les livrables (D6.3 "Platform Architectural Specification III", D9.2 "Community Driven Applications Design Report (2nd edition)", D6.6 "Platform UI Specification") du projet OpenMinTeD.

Un deuxième axe d'incitation à la participation au dispositif relève de la formation et du partage d'expérience en text mining. La capacité à créer autour du dispositif des lieux d'animation et d'échange sur les pratiques du text mining basées sur l'exploitation des moyens logiciels, de contenus et de ressources mis à disposition est un enjeu important.

Dans la communauté du text mining, le dispositif pourra également s'appuyer sur les chercheurs et ingénieurs à l'interface de questions de recherche interdisciplinaires. Ils sont spécialistes des questions de text mining et compétents pour instruire et traduire les besoins de disciplines spécifiques en solutions techniques. Le dispositif devrait proposer des lieux attractifs pour favoriser les rencontres entre l'expression des besoins disciplinaires et ces chercheurs et ingénieurs pour la création et le renforcement de collaboration et de projets. Les structures d'animation et les plateformes interdisciplinaires existantes (ex. SHS, bioinformatique) pourront servir de relais.

3.2.2 Transfert et valorisation

Le transfert et la valorisation de logiciels ou de corpus de fouille de textes sur la plateforme nécessitent que les logiciels soient réutilisables et puissent éventuellement interopérer. Du point de vue pratique le transfert et la valorisation des logiciels et des ressources associés demandent un effort important à plusieurs niveaux. Cela nécessite par exemple des compétences d'ingénierie pointues sur la plateforme, de la documentation, ainsi que la formation et l'accompagnement. Au niveau plus technique, la question de l'interopérabilité est, par exemple, à traiter particulièrement dans un contexte où différentes ressources provenant de sources différentes sont transférées et assemblées pour offrir des services adaptés aux utilisateurs.

Cette charge est un frein au transfert par les chercheurs ou les ingénieurs des unités de recherche si elle n'est pas suffisamment soutenue, éventuellement financée, valorisée comme produit de la recherche au même titre que les publications, ou contrebalancée par l'avantage retiré des services mis à disposition par la plateforme. L'expérience des projets Istex, OpenMinTeD et Visa TM montre que l'emploi de personnel temporaire de façon ponctuelle ne répond pas au besoin d'expertise approfondie sur la durée et les difficultés de recrutement en ingénierie informatique pour de courtes durées et des salaires modestes rendent cette solution impraticable.

Ces difficultés bien que réelles n'empêche pas le transfert d'être un enjeu important. Il permet principalement d'augmenter l'impact et la visibilité, d'élargir la communauté d'utilisateurs, et éventuellement de mettre en commun l'accompagnement humain. Ainsi, l'organisation et les comités en charge des prises de décision doivent favoriser ces transferts en mettant en place les conditions d'une mutualisation à la fois exigeante sur la qualité et la documentation, et respectueuse des travaux des contributeurs. Des mécanismes d'automatisation devraient pouvoir permettre de faciliter le travail.

3.2.3 Utilisation

L'utilisation typique de la future e-infrastructure pour des chercheurs ingénieurs en text mining porte sur la réutilisation de composants logiciels combinés avec leurs propres composants. L'expérience des "supporting resources" dans les compétitions qui incitent à la réutilisation de composants standards montre que les participants sont plus enclins à réutiliser des traitements très simples à mettre en place (ex. segmentation, entités nommées) ou des composants très populaires (Mc Closky Parser, Word2Vec), ou des composants rares (langues spécifiques) ou des composants qui associés à des corpus de référence, permettent d'évaluer la qualité des prédictions de leurs solutions, et de comparer à des prédictions obtenues par d'autres. Ces résultats d'évaluation et de comparaison sont nécessaires à la publication des recherches en fouille de textes et en TAL. Le dispositif devrait donc mettre à disposition prioritairement des corpus de références, des outils populaires et des outils d'évaluation et de comparaison sur des tâches populaires.

L'opportunité de collaborations scientifiques disciplinaires ou interdisciplinaires serait également un moteur de participation aux activités du dispositif. Des actions d'animation entre utilisateurs pourraient viser le partage de compétences entre spécialistes de la fouille de textes, mais aussi la rencontre entre des porteurs de besoin dans des disciplines et des spécialistes de la fouille de textes. Les domaines des sciences humaines et sociales et des sciences du vivant sont exemplaires à cet égard.

3.3 Les moyens d'animation

Le dispositif devra s'appuyer sur les nombreux moyens d'animation existants des communautés nationales, européennes et internationales. Sans prétention d'exhaustivité, nous en donnons ici quelques exemples importants.

3.3.1 Structures d'animation

Pré GdR TAL

En France, le pré GDR TAL[35] (*Groupement de Recherche pour le Traitement Automatique de la Langue*) financé par le CNRS a pour objectif de réfléchir à la mise en place d'un GdR pour "l'animation scientifique de la communauté [académique] en vue d'améliorer sa stratégie scientifique, son attractivité et sa visibilité". Deux des axes sont particulièrement pertinents pour l'élaboration d'une stratégie partagée avec le dispositif :

- Accès à l'information et fouille de textes (animateurs T. Charnois (LIPN) et Jean-Pierre Chevalleret (LIG))
- Ressources linguistiques (animateurs Gilles Ada (LIMSI TLP) Philippe Muller (IRIT))

Les échanges préliminaires ont permis de dégager des pistes. L'intérêt du pré GdR pourrait porter dans un premier temps sur la mise en valeur de quelques beaux exemples de réussite d'applications pilotes, où le dispositif servirait de "vitrine du TAL" et amorcerait la mise en place d'un transfert plus systématique. Une autre direction intéressante pour le GdR serait que la plateforme propose un ensemble de bons outils interconnectés pour le français : étiqueteurs morpho-syntaxiques, calculs de dépendances syntaxiques, extracteurs de termes, de manière similaire par exemple à Standford University de manière à promouvoir l'utilisation de la plateforme par la communauté TAL. La facilité d'intégration des outils est un enjeu important ici. Une stratégie raisonnable consiste à initier le processus avec des gens motivés de manière à mettre au point l'accompagnement et la documentation. Les outils d'animation (ci-dessous) tels que les *hackhatons* / formation et *shared tasks* sont des moyens éprouvés pour mobiliser ce public.

ATALA est l'association française pour le Traitement Automatique des Langues. Elle rassemble et fédère tous les acteurs de la communauté du TAL francophone. Elle organise la conférence nationale TALN (Traitement Automatique de Langues Naturelles) qui est un lieu de rencontre et d'échange central de la communauté TAL française. L'ATALA peut être un interlocuteur du dispositif pour l'organisation d'événements de type journées dédiées, ou journées de formation.

ARIA, l'association francophone de Recherche d'Information et Applications organise Coria, la COnférence Recherche d'Information et Applications. Les conférences jointes Coria et Taln

[35] https://pregdr-tal.ls2n.fr

sont un lieu intéressant pour l'organisation de tutoriels comme le tutoriel du projet OpenMinTeD organisé en mai 2018 par l'INRA.

Ressources

OrtoLang (*Outils et Ressources pour un Traitement Optimisé de la LANGue*) est un EquipEx, il a pour objectif d'offrir une bibliothèque de données (corpus, lexiques, dictionnaires, etc.) et d'outils sur la langue et son traitement pour le français. Il promeut l'utilisation de format standard d'entrée et sortie qui facilitent la réutilisation et l'intégration dans des workflows. En septembre 2019, elle contient plus de 420 ressources déclarées et téléchargeables dont 32 corpus écrits et des outils de type indexation, analyse flexionnelle et analyse syntaxique. Une partie de ces ressources n'est pas destinée à une réutilisation de type plateforme. La bibliothèque d'OrtoLang est déversée dans celle de CLARIN dont Ortolang est le nœud français. OrtoLang est coordonné avec la plateforme Huma-Num (ci-dessous). Le taux d'utilisation des ressources publiées et l'analyse de l'impact d'OrtoLang seront des éléments intéressants pour le choix des outils les plus pertinents pour le dispositif et leur promotion pour leur utilisation effective par le public visé.

CorLI (*Corpus, Langues, Interactions*)[36] est un consortium français qui "réunit des chercheurs et enseignants-chercheurs en linguistique, et se donne pour objectif de fédérer les équipes et laboratoires, les chercheurs, enseignants-chercheurs, ou ingénieurs engagés dans la production et le traitement de corpus numériques écrits et oraux, quels que soient la langue et/ou le système d'écriture considérés.". CorLI maintient une liste des outils et corpus. CorLI s'adresse aux chercheurs en linguistique intéressés et producteurs d'outils particuliers de type concordancier, outil d'annotation[37], ou textométrique et de ressources de type corpus[38]. Il interagit avec les bibliothèques de ressource OrtoLang[39] au niveau français et CLARIN au niveau européen. Serait à réfléchir par le dispositif visé la mise en place de modalités spécifiques de transfert et l'adaptation aux besoins particuliers de cette communauté.

CLARIN (*European Research Infrastructure for Language Resources and Technology*)[40] est une infrastructure européenne financée par les états ou les centres associés. La communauté de contributeurs de CLARIN est essentiellement celle du TAL en SHS. CLARIN met en place les services pour assurer le partage, l'utilisation et la pérennité des données et des outils linguistiques (écrit, oral ou multimodal) pour la recherche en sciences humaines et sociales. Un des objectifs de CLARIN est de proposer des outils avancés pour découvrir, explorer, exploiter, annoter, analyser ou combiner ces ensembles de données, où qu'ils se trouvent. L'accès est rendu possible par la fédération en réseau des dépôts de données linguistiques,

[36] https://corli.huma-num.fr
[37] https://corli.huma-num.fr/outils/
[38] https://corli.huma-num.fr/inventaire-des-corpus-ecrits/
[39] https://www.ortolang.fr
[40] https://www.clarin.eu

centres de services et centres de connaissances, avec un accès unique pour tous les membres de la communauté universitaire dans tous les pays participants. L'ambition de rendre les outils et les données provenant de différents centres interopérables, c'est-à-dire enchaînés pour effectuer des opérations complexes est partagée par le projet OpenMinTeD et le futur dispositif. Mais l'action de CLARIN aujourd'hui sur ce point est limitée à la standardisation des métadonnées de description et la promotion de format standard pour les entrées/sorties. De façon très complémentaire, OpenMinTeD fournit effectivement les outils permettant de composer puis d'exécuter les workflows par des non- spécialistes. De même qu'OpenMinTeD, le futur dispositif national devrait pouvoir exploiter de façon la plus automatisée possible les ressources de CLARIN. La coordination qui le permettrait est encore à créer en s'appuyant sur OrtoLang, le nœud français de Clarin.

ELRA (*European Language Resources Association*) assure la distribution de ressources linguistiques vocales, écrites et terminologiques pour la technologie du langage humain selon un modèle économique payant. ELRA participe également à la production, ou à la mise en service de la production, de ressources linguistiques par le biais d'un certain nombre d'initiatives également activement engagées dans l'évaluation des outils d'ingénierie linguistique ainsi que dans l'identification de nouvelles ressources [Source Wikipedia[41]]. Tous les deux ans, l'ELRA organise la grande conférence LREC, *l'International Language Resources and Evaluation Conference* qui est un lieu incontournable de rencontre européen pour la communauté de production, d'utilisation et d'évaluation de ressources. La conférence principale LREC est associée à de nombreux workshops et tutoriels satellites qui pourraient être des lieux d'animation privilégiés du futur dispositif au niveau européen en ce qu'ils permettent de rassembler un grand nombre de personnes intéressées, par la promotion et la réutilisation de ressources au-delà de la recherche.

DARIAH (*Digital Research Infrastructure for Arts and Humanities*) ERIC (European Research Infrastructure Consortium), DARIAH est une infrastructure européenne pour les chercheurs en arts et en sciences humaines travaillant avec des méthodes informatiques. Elle vise à améliorer et à soutenir la recherche et l'enseignement dans ces domaines au moyen du numérique. Le Groupe de travail sur *l'analyse des textes et des données* (Text and Data Analytics) a pour objectif d'élaborer et de démontrer des méthodologies pour l'application de l'état de l'art en TAL aux questions de recherche en sciences humaines. Pour ce faire, il éduque les chercheurs aux différentes technologies, outils et résultats. Les principales activités du GT sont de fournir des indications sur les meilleures pratiques, inventorier les applications réussies d'outils de premier plan et de démontrer des exemples spécifiques de services qui peuvent être utilisés dans la recherche en sciences humaines numériques [d'après le site web du projet[42]]. Ce groupe de travail devrait être un interlocuteur naturel du futur dispositif pour

[41] https://en.wikipedia.org/wiki/European_Language_Resources_Association

[42] https://www.dariah.eu

l'animation de la communauté et le partage des expériences et évaluations avec l'appui des instruments nationaux, dont Huma-Num en premier lieu.

Plateformes

La TGIR Huma-Num[43] est une TGIR : très grande infrastructure de recherche en SHS et financée par le Ministère de l'Enseignement supérieur et de la Recherche. Elle est portée par l'unité mixte de service associant CNRS, Université Aix-Marseille et Campus Condorcet [Wikipédia[44]]. Elle met à disposition un ensemble de services pour le stockage, le traitement, l'exposition, le signalement, la diffusion et la conservation sur le long terme des données numériques de la recherche en sciences humaines et sociales [site web]. Elle s'appuie sur un dispositif humain et technologique (services numériques pérennes) à l'échelle nationale et européenne avec un réseau de partenaires et d'opérateurs. Elle coordonne la participation française à DARIAH et CLARIN. Une partie de l'activité d'Huma-Num relève du text mining pour lequel elle s'appuie en partie sur les technologies développées par l'unité CNRS LSIS (Laboratoire des Sciences de l'Information et des Systèmes). Elle développe à la demande des applications exploitant les données disponibles sur la plateforme Isidore.

Un partenariat entre le futur dispositif national de text mining et Huma-Num serait naturel étant donné la complémentarité des missions. Huma-Num pourrait y trouver les moyens de rationaliser, systématiser et partager les développements en text mining. Elle pourrait faire bénéficier de son expertise en text mining pour les humanités numériques et mutualiser les technologies développées et retours d'expérience.

IFB, l'Institut Français de Bioinformatique est l'infrastructure nationale de service en bioinformatique créée dans le cadre du programme national des «Investissements d'Avenir». Elle mutualise, soutient et coordonne le développement des ressources et des activités de support à la recherche de plateformes de bioinformatique dépendant d'organismes publics de recherche. Le text mining pour la biologie est dans le périmètre de l'IFB, mais cet axe n'est pas développé au niveau français, partiellement au niveau européen dans ELIXIR dont l'IFB est le nœud français.

ELIXIR[45] (*the European life-sciences Infrastructure for biological Information*) a pour objectif de permettre aux laboratoires des sciences de la vie à travers l'Europe de partager et de stocker leurs données de recherche dans le cadre d'un réseau organisé. Le réseau réunit les principales organisations européennes en sciences de la vie pour gérer le volume croissant des données issues de la recherche publique. Elle coordonne, intègre et soutient les ressources bioinformatiques de ses États membres et permet aux utilisateurs du monde universitaire et de l'industrie d'accéder à des services essentiels pour leurs recherches [d'après le site web[20]].

[43] https://www.huma-num.fr

[44] https://fr.wikipedia.org/wiki/Huma-Num

[45] https://elixir-europe.org

Elixir a lancé des initiatives soutenues par le CNIO (nœud espagnol) de services de text mining pour contribuer en particulier à la curation des ressources biologiques comme SciLite[46] avec PMC Europe. Elixir a également contribué dans OpenMinTeD à l'alignement des ontologies des métadonnées de description des données et outils (OMTD-Share[47]/EDAM[48]). ELIXIR et l'IFB sont des interlocuteurs privilégiés pour le dialogue avec les plateformes de bioinformatique européennes. Le faible développement de service de text mining au regard de la multiplication d'applications, dispersées à ce jour, et des besoins, nécessitera une stratégie de coordination, de formation et d'identification des besoins spécifiques à ce domaine.

3.3.2 Activités

La communauté de recherche et développement en text mining se rassemble autour d'activités qui favorisent la mutualisation la réutilisation et les transferts d'outils, leur évaluation et le partage d'expérience. Elles sont des moyens puissants d'impliquer la communauté dans le développement du dispositif national.

Ateliers scientifiques

En premier lieu les ateliers réunissant des scientifiques produisant des ressources, avec des utilisateurs de ces ressources pourront (i) contribuer à la formation d'une communauté partageant un intérêt commun pour le transfert et le développement d'applications avec des scientifiques non-informaticiens, et en particulier de domaines de spécialité (ii) préciser les questions de recherche auxquelles ces ressources contribuent (iii) identifier les ressources dont la maturité permet l'intégration dans la plateforme et (iv) analyser collectivement les facteurs facilitant le transfert.

Ils pourraient être complétés par d'autres ateliers avec l'objectif de mener une analyse réflexive de la démarche, de communiquer les résultats et de définir les instruments à mettre en place pour pérenniser la dynamique de mutualisation et de co-construction, dans la perspective de l'exploitation de l'infrastructure en France, et au-delà.

Accompagnement

Le soutien aux participants pourra se décliner en différents instruments : des outils classiques de type courrier électronique, forum et l'organisation d'un *hackathon* sur une journée ou deux destiné à accompagner techniquement l'intégration et la documentation des nouvelles ressources dans l'e-infrastructure OpenMinTeD. Nous pourrons utiliser (pour des cas d'usages

[46] https://europepmc.org/annotations
[47] https://openminted.github.io/releases/omtd-share-schema/
[48] http://edamontology.org/page

agro/agri/food, la série des AgroHackathon (http://www.agrohackathon.com) mise en place en 2016 et 2017 au LIRMM, pour inscrire nos actions dans la durée.

Compétitions scientifiques

L'organisation de campagnes d'évaluation basées sur des tâches partagées, comme outils d'évaluation et d'animation scientifique pour la fouille de textes, est un instrument très intéressant dans la perspective du développement de la plateforme. Ces campagnes favorisent l'émergence d'outils de bonne maturité technologique, la standardisation des entrées-sorties des outils et facilitent la communication et l'évaluation. La contribution de la plateforme à des campagnes pérennes est un moyen d'incitation à travailler sur des applications finalisées à travers les données spécialisées. Elles promeuvent une approche modulaire et la composition de workflow avec une forte motivation de réutilisation (à travers les *supporting resources*) et de reproductibilité. La plateforme aura un rôle dans l'incitation et l'accompagnement du passage de l'évaluation à la réutilisation (suivant le modèle BeCalm[49] dans BioCreative[50] ou de PubAnnotation[51] dans BioNLP-OST[52]). A l'issue des compétitions, l'organisation d'ateliers de restitution avec sélection d'articles et présentations orales suivant le modèle de BioNLP-ST ou CLEF[53] (Conference and Labs of the Evaluation Forum) est un moyen reconnu de construire une analyse partagée pour tirer un bilan sur les technologies et ouvrir de nouvelles perspectives de recherche et développement.

Au niveau français certaines campagnes sont pressenties pour une coordination avec la future plateforme : BioNLP-OST27, compétition internationale historique en extraction d'information dont l'INRA-MaIAGE, partenaire de Visa TM est co-organisateur (ii) CLEF28 dans laquelle le LSIS (Laboratoire des Sciences de l'Information et des Systèmes) est impliqué depuis une dizaine d'années, pour des services orientés recherche d'information ou recommandation, navigation, résumé-synthèse, (iii) DEFT[54] pour des corpus en français et dont le LIMSI (Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur) est le principal organisateur, (iv) SemEval[55] (International Workshop on Semantic Evaluation), notamment pour du text mining orienté vers l'extraction d'information, mais aussi l'analyse de sentiments (par ex. pour la détection de prises de position et d'opinions dans les articles scientifiques).

[49] https://jcheminf.biomedcentral.com/articles/10.1186/s13321-019-0363-6
[50] https://en.wikipedia.org/wiki/BioCreative
[51] http://pubannotation.org
[52] https://2019.bionlp-ost.org
[53] http://www.clef-initiative.eu
[54] https://deft.limsi.fr/2019/
[55] https://en.wikipedia.org/wiki/SemEval

Appel incitatif à contribution pour l'intégration de nouvelles ressources (Tender Call)

L'objectif de tels appels est de pallier au manque d'ingénieurs de transfert par un soutien financier en ciblant prioritairement des personnes engagées dans la démarche pour tester les conditions humaines, matérielles et logicielles du transfert et en tirer des leçons. Nous identifions plusieurs facteurs conjoints basés sur notre analyse des expériences de *tender call* du projet OpenMinTeD et des chantiers d'usage du projet Istex : (1) les ressources doivent répondre à des critères d'intégrabilité sur les plans techniques et légaux et à des critères d'impact prédéfinis et évalués par un comité technique et scientifique (voir D2.3 "OpenCalls Specification") (2) une équipe de support technique et de formation accompagne les fournisseurs dans l'intégration technique et la documentation du produit (métadonnées, tutoriels), (3) un crédit incitatif, conditionné par l'évaluation de la production finale (fonctionnalité, documentation, analyse de l'impact) soutient financièrement l'ingénierie de développement. L'objectif est triple : augmenter significativement les services clef en main disponibles en déléguant leurs développements à leurs inventeurs, identifier les évolutions techniques et organisationnelles qui faciliteront la mutualisation et élargir la base des participants et promoteurs du projet.

Hackathon

L'accompagnement et la formation devra se décliner en différents instruments selon les besoins : des outils classiques de type courrier électronique, forum, forges. Ponctuellement des actions de formation (*webinar*, tutoriel) sont à prévoir. Les *hackathons* et RAMP[56] (Rapid Analytics and Model Prototyping) sont des instruments très intéressants. Sur une journée ou deux, ils seraient destinés à accompagner techniquement l'intégration et la documentation des nouvelles ressources dans l'e-infrastructure OpenMinTeD et inversement à travailler collectivement à la réutilisation. Ils sont une forme de formation participative qui contribue fortement à créer des collectifs mêlant ingénieurs et scientifiques. L'exemple en agro/agri/food, la série des AgroHackathon[57] mise en place par le LIRMM en 2016 et 2017 est intéressant. On pourra également s'inspirer du modèle de la série des BLAH (Biomedical Linked Annotation Hackathon)[58] dans le domaine du text mining qui rassemble annuellement un grand nombre de participants réguliers du monde entier. Les RAMP organisés par le CDS[59] (Center for Data Science) donnent un exemple complémentaire de *hackathon* en mode compétition. De la même manière qu'un défi de text mining, le principe en est le suivant : le fournisseur de données arrive avec un problème de prédiction et un ensemble de données correspondant. Un informaticien expérimenté préalablement nettoie et conserve les

[56] https://ramp.studio
[57] http://www.agrohackathon.com
[58] http://blah5.linkedannotation.org/home
[59] https://www.datascience-paris-saclay.fr

données, formalise le problème et installe les ressources et outils et accès nécessaires. Lorsque le problème de text mining nécessite la maîtrise d'un outil spécifique, l'événement peut être précédé d'un Sprint[60] d'entraînement à l'utilisation des outils. C'est typiquement le cadre approprié pour promouvoir l'utilisation d'une plateforme avec ses logiciels spécifiques telle qu'OpenMinTeD et faire se rencontrer les personnes pour favoriser les échanges futurs.

3.3.3 Conclusion

L'implication de la communauté de recherche et d'ingénierie en TAL et text mining est critique pour le renouvellement des technologies de la future plateforme. Nous avons présenté ici un certain nombre de structures nationales et européennes susceptibles de contribuer à cette dynamique. Comme nous l'avons vu, la plateforme pourra également s'impliquer dans des activités plus spécifiques à ses objectifs susceptibles de contribuer à créer une communauté.

[60] https://fr.wikipedia.org/wiki/Sprint_(développement_logiciel)

Conclusion

Identifiés comme les éléments prépondérants dans le processus de mise en œuvre de la fouille de textes, les outils logiciels offrant des fonctions de traitement de fouille de textes ou d'assistance aux tâches de fouille de textes ont fait l'objet d'une attention particulière dans ce document. Nous avons proposé une revue des outils de l'écosystème d'OpenMinTeD avec des pistes d'amélioration de leur prise en charge. Nous identifions aussi l'importance de se doter de mécanismes pour la sélection des outils et d'un dispositif d'animation de la communauté. Ce travail dégage une feuille de route pour l'amélioration de l'infrastructure OpenMinTeD et pour l'intégration d'outils de text mining supplémentaires dans le dispositif en partant de l'existant et en impliquant les communautés.

Bien que les ressources de manière générale, et en particulier les ressources utilisées à travers les outils (corpus, grammaires, ontologies, etc.), n'aient pas bénéficié d'une attention dans ce livrable, elles n'en restent pas moins importantes dans le processus de mise en œuvre du text mining. Leur utilisation dans OpenMinTeD, leur recensement ainsi que leur gestion dans VisaTM sont autant de questions qui méritent un travail supplémentaire que nous mettons en perspective.

Index des figures

Index des tableaux

Annexes

Annexe 1 : Critères de sélection des outils du tender call dans OpenMinTeD

C1	**Alignment with OpenMinTeD interoperability standards for software components and/or for knowledge resources.**	**3 points**
Licensing: the software is distributed under a perpetual, worldwide, no-charge, royalty-free copyright/patent licence that permits unrestricted use and allows unlimited redistribution		
The proposal explicitly adheres to OMTD format specifications: XMI, OpenAnnotation.		
The proposal includes formal description and documentation and detailed information about operation and access (inputs/outputs, executional requirements, annotation schema dependencies, ...		
C2	**Appropriateness , feasibility of methodology**	5 points
The technical description includes enough details about the implementation.		
The option chosen is properly identified & described. (see: [[1]] for software options & [[2]] for knowledge resources options at the bottom of template; raw 78 and 79)		
The proposal clearly defines all necessary tasks and includes a feasible detailed agenda.		
The proposal details the proposed work, rationale, use cases and usage scenario		
The proposal is reasonable in terms of technical complexity and integration into the OMTD platform.		
C3	**Risk assessment and timescales**	2 points
The proposal includes the identification of risks related to the eventual implementation.		
The proposal includes risk solving actions.		

C4	Appropriateness of level of staffing, resources, expertise	2 points
The proposal provides enough details about staff, capabilities and expertise (reference sites, past projects, historical background,...)		
The applicants have/provide the required resources to fulfill the project.		
C5	**Level of innovation**	4 points
The proposal addresses state-of-the-art topic/task		
The proposal provides better solutions that meet new requirements, implies a renewal and enlargement of products		
C6	**Price and value for money**	4 points
The proposal will increase the visibility of OMTD (relevance of the integrated tool/resource, expected users, dissemination plan/activities ...)		
The proposal integrates a high impact component/resource in any of the OMTD domains and there exists an an active user community.		
C7	**Project experience and proven track records planning, management**	2 points
The applicants have extensive experience in the field (reference sites, past projects, historical background, ...)		
The proposal includes planned tasks and management information.		
C8	**Alignment with high relevance tender topics**	10 points
HRT-1. The proposal is relevant to OMTD community use cases, accessible and aligned with the OMTD infrastructure, in particular semi-automated biocuration in large databases, named entity recognition, concept indexing, relation extraction and entity grounding systems.		
HRT-2. The proposal assesses technical aspects, compatibility with OMTD specification and robustness of third party components for integration into the OMTD infrastructure.		
HRT-3. The proposal aligns high impact third party general purpose language processing tools with the OMTD infrastructure.		

HRT-4. The proposal handles OMTD interoperability issues at the level of document representation and widely used standard annotation formats and their evaluation in terms suitability for usage OMTD workflow infrastructures.
HRT-5. The proposal enables alignment with the OMTD infrastructure of widely used third party data mining and machine-learning components.
HRT-6. The proposal enables alignment with the OMTD infrastructure of one of the following language processing systems: Stanford CoreNLP, Apache OpenNLP, NLTK, FreeLing, IXA pipes.
HRT-7. The proposal enables alignment of one of the following processing environments with the OMTD infrastructure: Apache uimaFIT, Kachako, Argo, GATE, Taverna, Heart of Gold, Vistrails, Kepler, ALPE (Automatic Linguistic Processing Environment), TextGrid, WebLicht, DKPro Core, Newsreader, ...
HRT-8. The proposal enables alignment of high impact or community use case provided knowledge bases, ontologies or controlled vocabularies with the OMTD infrastructure (data analysis and data integration of text-derived and knowledge base-derived data).
HRT-9. The proposal enables alignment with the OMTD infrastructure with text meta-annotation systems offering integration and/or providing consensus/harmonized annotations.
HRT-10. The proposal is an adaptation, integration and interoperability of third party software including proprietary text mining platforms as well as standard annotation formats into the OMTD platform.
HRT-11. The proposal use/combines OMTD infrastructure components (in/via workflows) to create innovative services in scientific domains.
HRT-12. The proposal uses the OMTD annotation platform services to generate compliant text mining services and Gold Standard data.

Annexe 2 : Base de critères pour la sélection d'outils TM dans VisaTM

C1	Adéquation d'un outil par rapport aux objectifs de VisaTM	Poids à définir
Types de critères	Politique, Stratégique	
Rôles	Critère permettant d'évaluer l'alignement d'un outil par rapport à des objectifs globaux définis dans VisaTM	
Commentaires	Ce critère permet de traiter des questions d'ordre politique et stratégique. Les caractéristiques des outils correspondant à ces aspects y seront traitées (coopérations, transfert de technologies et de compétences, ouverture, opportunités, risques, prévisions...). Il peut dépendre d'autres critères.	
Exemples de questions	Est-ce que l'outil permet de répondre aux objectifs de VisaTM concernant les outils ?	
Exemples de pré-requis	Définir et fixer des objectifs stratégiques concernant les outils dans VisaTM	
C2	**Adéquation avec un cahier des charges qui évalue un outil par rapport à l'environnement de VisaTM**	**Poids à définir**
Types de critères	Politique, Organisationnel	
Rôles	Critère permettant d'évaluer l'adéquation d'un outil par rapport au cadre et aux besoins de VisaTM	
Commentaires	Les ressources pour exploiter un outil sont déterminantes pour son adoption. Ces ressources peuvent concerner l'environnement pour rendre opérationnel l'outil (système d'exploitation, mémoire, CPU), l'expertise pour maintenir l'outil fonctionnel, le cadre légal, le type de support, etc. On peut aussi être amené à considérer une liste d'autres fonctions (charges, multi-utilisateurs, interactivité).	

Exemples de questions	Quels compétences sont nécessaires pour l'outil ? Quel type d'infrastructure pour l'outil ? Quel cadre légal pour l'exploitation de l'outil ? Quelles fonctions offre l'outil ?	
Exemples de pré-requis	Définir un cahier des charges technique et fonctionnel de l'infrastructure VisaTM	
C3	**Couverture d'un outil par rapport aux thématiques prioritaires de VisaTM**	**Poids à définir**
Types de critères	Politique, Organisationnel	
Rôles	Critère permettant d'évaluer les thématiques couverts par un outil en comparaison avec des thématiques fixés dans VisaTM	
Commentaires	Les outils peuvent être spécialisés à un domaine d'application, de portée générale, spécialisés pour un ou plusieurs tâches de haut (classification, résumé, traduction) ou bas niveau (détection d'entités nommées, normalisation)	
Exemples de questions	Quels sont les thématiques couvertes par l'outil ? Comment les thématiques de l'outil se comparent aux thématiques fixés dans VisaTM ?	
Exemples de pré-requis	Disposer de listes des thématiques prioritaires de VisaTM	
C4	**Qualité méthodologique d'un outil**	**Poids à définir**
Types de critères	Scientifique, Méthodologique	
Rôles	Critère pour évaluer les approches implémentées par un outil	

Commentaires	Les qualités méthodologiques peuvent être difficiles à estimer, cependant elles peuvent être déterminantes pour évaluer les outils
Exemples de questions	Quelles catégories d'approches concerne l'outil ? Quelle est la position de l'outil par rapport à l'état de l'art ? Comment est évaluer/valider les approches de l'outil ?
Exemples de pré-requis	Définir une méthodologie d'évaluation des approches implémentées par les outils

C5	**Couverture d'un outil (par rapport aux langues, domaines, richesse fonctionnelle)**	**Poids à définir**
Types de critères	Technique, Fonctionnel	
Rôles	Critère pour évaluer la couverture d'un outil	
Commentaires	Les outils couvrent plus ou moins bien divers aspects tels que la langue, la spécialisation, etc. Par exemple, certaines outils tels que AlvisNLP offrent des modules indépendants adaptées à un domaine, TermSuite supporte plusieurs langues, TensorFlow propose des fonctions qui sont du domaine général (Word2Vec). Il convient ainsi de faire un parallèle entre les besoins et les caractéristiques des outils	
Exemples de questions	Est-ce que l'outil prend en charge les langues désirées? Est-ce que l'outil est spécialisé pour une langue, une tâche, une fonction, ou un domaine ? Est-ce que l'outil est adaptable, évolutif ?	
Exemples de pré-requis	Etablir la liste des aspects obligatoires, optionnelles à considérer pour comparer les outils	

C6	**Types de contenus traités par un outil**	**Poids à définir**
Types de critères	Technique, Fonctionnel	

Rôles	Critère permettant de caractériser les types de ressources qui sont traités par un outil
Commentaires	Les contenus, les sources, les ressources annexes liés à un outil peuvent apporter des renseignements importants
Exemples de questions	L'outil est-il spécialisé dans un/des type(s) de contenus ? L'outil dépend-il de ressources annexes spécifiques ? Existe-t-il des sources, des caractéristiques spécifiques pour les données que traite l'outil ?
Exemples de pré-requis	

C7	**Qualité d'un outil permettant de garantir une utilisation réelle dans VisaTM**	**Poids à définir**
Types de critères	Technique, Opérationnel	
Rôles	Critère permettant de donner des indications sur la qualité au niveau opérationnel d'un outil	
Commentaires	Des indicateurs mesurables sur la performance et le plan de gestion de l'outil peuvent être utilisés	
Exemples de questions	Quel niveau global de performance, de fiabilité, de portabilité, de sécurité ? Existe-t-il un cadre actif pour la gestion des releases, bugs ? Existe-t-il une documentation technique et fonctionnelle pour les développeurs et les utilisateurs ?	
Exemples de pré-requis	Définir une liste d'indicateurs pour juger de la qualité des outils	

C8	**Modularité et clarté des usages prévus pour un outil**	**Poids à définir**
Types de critères	Technique, Opérationnel	

Rôles	Critère permettant d'évaluer les interfaces d'exploitation d'un outil
Commentaires	Critère pouvant intervenir pour aider à anticiper l'intégration et l'évolution des outils, important pour les intégrateurs d'outils
Exemples de questions	Quels sont les modules identifiables de l'outil ? Quels sont les usages prévus pour l'outil ? Quelles sont les interfaces d'accès ?
Exemples de pré-requis	

C9	**Intégration et interopérabilité**	**Poids à définir**
Types de critères	Technique, Opérationnel	
Rôles	Critère permettant de donner des indications sur les problématiques d'intégration et d'interopérabilité d'un outil	
Commentaires	Question centrale pour une bonne intégration et un bon fonctionnement d'un outil dans une plateforme qui évolue avec d'autres ressources	
Exemples de questions	Quel niveau de FAIRization ? Quel(s) sont les types d'entrée/sortie, formats, hyper-paramètres, ressources auxiliaires de l'outil ? Quelle est la nature de l'outil (Service Web, API, platform en ligne, client lourd, containers) ? Quelles relations entre l'outil et les ressources existantes dans la plateforme VisaTM ?	
Exemples de pré-requis	Définir les exigences concrètes de la plateforme VisaTM en termes d'interopérabilité et d'intégration	

C10	**Prise en main, degré d'expertise requis**	**Poids à définir**
Types de	Technique, Opérationnel	

critères	
Rôles	Critère permettant d'évaluer la prise en main d'un outil
Commentaires	Certains outils peuvent être plus faciles à utiliser que d'autres ayant des fonctions similaires s'ils sont accompagnés de certaines ressources (tutoriels, démos, donnée de test/évaluation).
Exemples de questions	Quelle interface d'accès pour l'outil ? Quels niveaux de compétences sont requis pour l'utilisation de l'outil ? Existe-t-il des ressources pour aider à utiliser l'outil ? Qui sont les utilisateurs habituels de l'outil ?
Exemples de pré-requis	Connaître le cadre d'usage des outils et les acteurs de la plateforme VisaTM

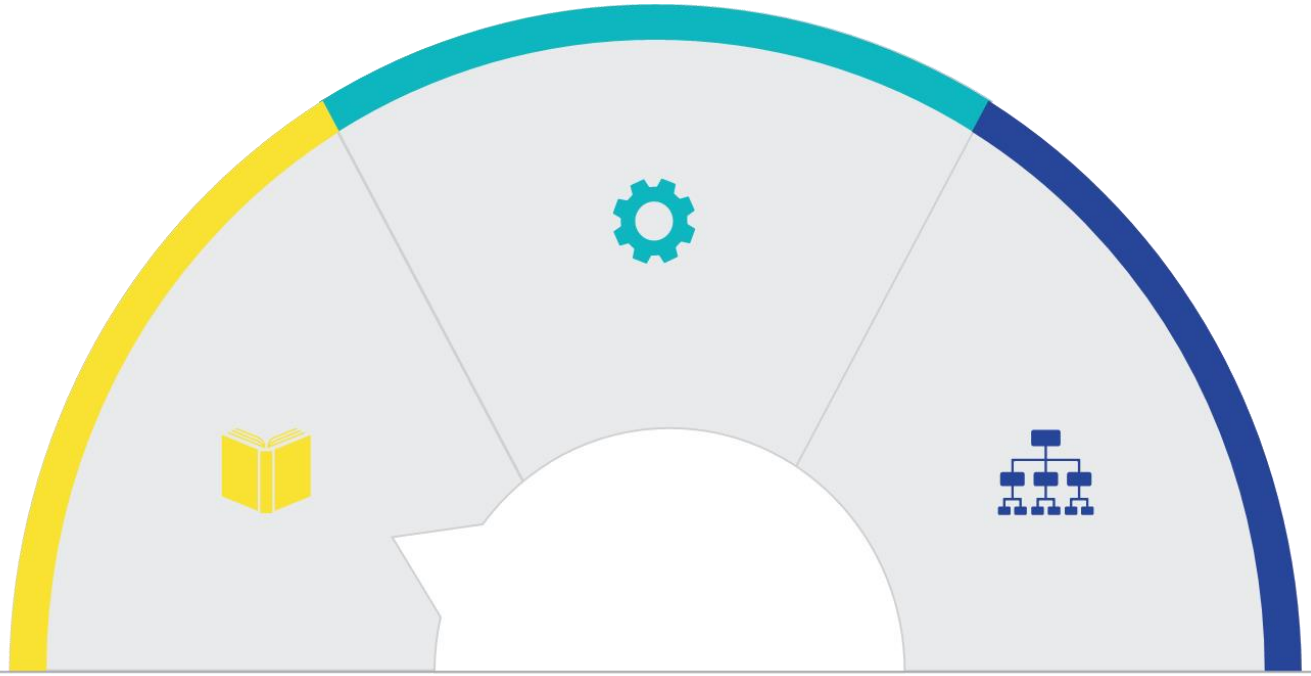
C11	**Existence d'une communauté autour de l'outil, popularité de l'outil**	**Poids à définir**
Types de critères	Communauté	
Rôles	Critére permettant de mesurer la portée d'un outil	
Commentaires	L'existence, la portée et le type de(s) communauté(s) autour d'un outil peut apporter des renseignements	
Exemples de questions	Existe-t-il une communauté Open source autour de l'outil ? Quelle est la nature des activités autour de l'outil ? Quelle est de manière globale la dynamique autour de l'outil ?	
Exemples de pré-requis		

C12	**Méthode et plan de diffusion de l'outil**	**Poids à définir**
Types de	Prévisionnel	

critères		
Rôles	Critére permettant d'évaluer l'accessibilité d'un outil	
Commentaires	Il y a des outils sous licences commerciales, avec des restrictions fonctionnelles ou techniques ou un support payant. Il peut être utile de savoir	
Exemples de questions	Quel est la politique de diffusion de l'outil ? Existe-t-il d'autres implémentations communautaires, des versions pédagogiques, des options payantes ?	
Exemples de pré-requis		
C13	**Expérience et importance des fournisseurs**	**Poids à définir**
Types de critères	Prévisionnel	
Rôles	Critère permettant d'évaluer la position des fournisseurs d'un outil dans la communauté	
Commentaires	Ce critère peut être important, peut permettre de gagner du temps dans l'évaluation	
Exemples de questions	Qui sont les fournisseurs ? Sont-ils fournisseurs d'outils déjà intégrés dans VisaTM ?	
Exemples de pré-requis	Une connaissance des fournisseurs d'outils dans le domaine	
C14	**Historique, avenir de l'outil**	**Poids à définir**
Types de critères	Prévisionnel	
Rôles	Critère permettant de recueillir des informations sur l'historique, l'avenir et les dépendances d'un outil	
Commentaires	L'historique et les relations existantes entres les outils peuvent être utiles	

Exemples de questions	Comment l'outil a évolué depuis sa création ? Es-ce que l'outil dépend d'autres outils ? Existe-il des outils dépendants déjà intégrés dans VisaTM ? Quelles évolutions pour l'outil ?
Exemples de pré-requis	

Annexe 3 : Recensement des outils de fouille de textes et de données



VisaTM étude

Recensement des outils de fouille de textes et de données

Mise à jour du 28/08/19

INIST-CNRS / F. Arnould

ABBYY Solutions	ABBYY technologies and platforms for document recognition, data capture, and language processing.	Commercial	Classification de textes ; Reconnaissance d'entités nommées ; Découverte de connaissances ; Traduction automatique ; Recherche d'information ;	Non	https://www.abbyy.com/en-eu/	Russie ;
ABNER	ABNER is a software tool for molecular biology text analysis.	Libre	Reconnaissance d'entités nommées ;	Oui	http://pages.cs.wisc.edu/~bsettles/abner/	États-Unis ;
Abzooba	Social media and text analytics software	Commercial	Analyse de sentiments ; Classification ; Reconnaissance d'entités nommées ;	Non	http://www.abzooba.com/	États-Unis ;
ADAM	Data Mining and Image Processing Toolkits	Libre	Classification ; Clustering ; Reconnaissance de forme ; Règles d'association ; Optimisation ; Traitement d'images ;	Non	http://projects.itsc.uah.edu/datamining/adam/	États-Unis ;
AdaMSoft	ADaMSoft is a free and Open-Source System for Data Management, Data and Web Mining, statistical Analyis and more.	Libre	Classification ; Clustering ; Analyse de régression ;	Non	http://adamsoft.sourceforge.net/	Italie ;
ai-one	Transforms big data into opportunity using machine learning that mimics the biological brain's ability to find patterns and relationships.	Commercial	Apprentissage automatique ;	Non	http://www.ai-one.com/	États-Unis ; Suisse ; Allemagne ;
Aika	Java library that automatically extracts and annotates semantic information into text	Libre	Annotation ; Désambiguïsation ; Catégorisation de textes ; Reconnaissance d'entités nommées ; Extraction d'information ;	Non	http://www.aika-software.org	Allemagne ;
Alceste	logiciel d'analyse de données textuelles, ou statistique textuelle	Commercial	Classification ;	Non	http://www.image-zafar.com/Logiciel.html	France ;
AllenNLP	An open-source NLP research library, built on PyTorch.	Libre	Etiquetage sémantique ; Reconnaissance d'entités nommées ; Q&A ; Résolution de coréférence ; Textual entailment ; Constituency parsing ;	Non	http://allennlp.org/	États-Unis ;

Alteryx Project Edition	Predictive for Project Edition	Commercial	Prétraitement ; Apprentissage automatique ;	Non	https://www.alteryx.com/fr/predictive-project-edition	États-Unis ;
Alveo	Alveo connects HCS (Human Communication Science) researchers, their desks, computers, labs, and universities and accelerates HCS research to produce emergent knowledge that comes from novel application of previously unshared tools to analyse previously difficult to access data sets.	?	Recherche d'informations ; Reconnaissance de la parole ; Annotation ;	Non	http://alveo.edu.au/	Australie ;
Alvis NLP	A pipeline framework for Natural Language Processing	Libre	Annotation ; Classification ; Clustering ; Q&A ; Traduction ; Recherche d'information ; Analyse de sentiments ; Analyse d'opinions ; Reconnaissance d'entités nommées ; Racinisation ; PoS tagging ;	Oui	http://www.quaero.org/module_technologique/alvis-nlp-alvis-natural-language-processing/	France ; Allemagne ;
Amazon Comprehend	Détection d'informations et de relations dans un texte	Commercial	Extraction de phrases clés ; Analyse de sentiments ; Analyse syntaxique ; Reconnaissance d'entités nommées ; Extraction d'informations ; Détection de la langue ; Classification ; Topic modeling ;	Non	https://aws.amazon.com/comprehend/	États-Unis ;
AMI	Intégrateur de solutions logicielles de pointe pour la Cybersécurité, la Cyber Intelligence, la Veille Stratégique et le Traitement Automatique de la Parole.	Commercial	Traitement automatique de la parole ; Recherche d'informations ; Extraction de mots clés ; Annotation ;	Non	https://www.bertin-it.com/	France ;
Anaconda	Python data science platform	Libre/Commercial	Apprentissage automatique ;	Non	https://www.anaconda.com/	États-Unis ;
Analec	Annotation et analyse de corpus écrits	Libre	Annotation ;	Non	http://www.lattice.cnrs.fr/Telecharger-Analec	France ;
Annomarket	Cloud-based Text Annotation	Commercial	Annotation ;	Oui	https://github.com/annomarket/	?
AntConc	Freeware text analysis and concordance tool kit	Libre	Concordancier ;	Non	http://www.laurenceanthony.net/software/antconc/	Japon ;
Apache cTAKES	Natural language processing system for extraction of information from electronic medical record clinical free-text.	Libre	Extraction d'information ; Annotation ;	Non	http://ctakes.apache.org/	États-Unis ;

Apache Mahout	Environment for quickly creating scalable performant machine learning applications.	Libre	Architecture logicielle ; Analyse régression ; Clustering ;	Non	http://mahout.apache.org/	États-Unis ;
Apache OpenNLP	The Apache OpenNLP library is a machine learning based toolkit for the processing of natural language text.	Libre	Apprentissage automatique ; Lemmatisation ; Parsing ; Chunking ; Tokenisation ; PoS tagging ; Sentence splitting ; Reconnaissance d'entités nommées ; Résolution de coréférence ; Détection de la langue ; Classification ;	Oui	https://opennlp.apache.org/	États-Unis ;
Apache UIMA	Component software architecture for the development, discovery, composition, and deployment of multi-modal analytics for the analysis of unstructured information and integration with search technologies.	Libre	Architecture logicielle ;	Oui	https://uima.apache.org/	États-Unis ;
Argo	Argo is a workbench for building and running text-analysis solutions. It facilitates the development of custom workflows from a selection of elementary analytics.	Libre	Annotation ; Recherche d'informations ; Reconnaissance d'entités nommées ;	Non	http://argo.nactem.ac.uk/	Royaume-Uni ;
Ascribe	Accelerate ROI via our surveys, vast sample and advanced text analytics	Commercial	Classification ; Analyse d'opinions ; Analyse de sentiments ; Clustering ; Extraction d'informations ;	Non	https://goascribe.com/	États-Unis ;
ats	Regression and clustering analysis	Libre	Clustering ; Analyse de régression ;	Non	http://www.mepx.org/	?
Averbis	Averbis provides leading text mining and machine learning solutions for your business. We convert text into information, automate cognitive processes, and make meaningful predictions.	Commercial	Découverte de connaissances ; Extraction terminologique ; Reconnaissance d'entités nommées ; Classification de textes ; Analyse de sentiments ; Analyse d'opinions ; Recherche d'informations ;	Non	https://averbis.com/en/	Allemagne ;
Aylien	AI-driven content analysis solutions that bring the power of NLP to the masses. We help developers, data scientists, and marketers understand human-generated textual content at scale.	Commercial	Analyse de sentiments ; Classifiation ; Résumé automatique ; Reconnaissance d'entités nommées ; Extraction d'informations ; Annotation ;	Non	https://aylien.com/	Irlande ;
Babel X	Babel X® is a multi-lingual, geo-enabled, text-analytics, social media and web-monitoring platform designed to meet the needs of our customers by fully leveraging publicly available information in this era of overwhelming quantities of geographically diverse, multi-lingual data.	Commercial	Analyse de sentiments ; Clustering ; Reconnaissance d'entités nommées ; Extraction de relations ; Recherche d'informations ;	Non	https://www.babelstreet.com/	États-Unis ;
BANNER	Named entity recognition system	Libre	Reconnaissance d'entités nommées ;	Oui	http://banner.sourceforge.net/	États-Unis ;

BioCreative	BioCreative: Critical Assessment of Information Extraction in Biology is a community-wide effort for evaluating text mining and information extraction systems applied to the biological domain	Libre	Reconnaissance d'entités nommées ; Extraction de relations ; Annotation ;	Oui	http://www.biocreative.org/events/biocreative-v/CFP/	?
BioLemmatizer	The BioLemmatizer is a domain-specific lemmatization tool for the morphological analysis of biomedical literature.	Libre	Lemmatisation ;	Non	http://biolemmatizer.sourceforge.net/	États-Unis ;
BioNLP	BioNLP is an initiative by the Center for Computational Pharmacology at the University of Colorado Denver to create and distribute code, software, and data for applying natural language processing techniques to biomedical texts.	Libre	Parsing ; Lemmatisation ; Annotation ; Classification de textes ; Extraction d'informations ;	Oui	http://bionlp.sourceforge.net/	États-Unis ;
BioTagger	These pages describe briefly Penn's BioTagger software suite. Currently the tagger supports three types of entities: gene entities, genomic variations entities and malignancy type entities.	Libre	Reconnaissance d'entités nommées	Non	https://www.seas.upenn.edu/~strctlrn/BioTagger/BioTagger.html	États-Unis ;
Biotex	Biomedical term extraction	Libre	Extraction terminologique ;	Non	http://tubo.lirmm.fr/biotex/index.jsp	France ;
Bitext	NLP Plateform	Commercial	Lemmatisation ; PoS tagging ; Détection de la langue ; Extraction de phrases ; Reconnaissance d'entités nommées ; Analyse de sentiments ;	Non	https://www.bitext.com/	Espagne ;
Bluima	Natural language processing toolkit for neuroscience	Libre	Reconnaissance d'entités nommées ;	Non	https://github.com/BlueBrain/bluima	Suisse ;
Brainspace	Brainspace creates breakthrough machine learning software that intelligently detects and relates unique phrases in massive unstructured datasets	Commercial	Clustering ;	Non	https://www.brainspace.com/	États-Unis ;
Brat	Online environment for collaborative text annotation.	Libre	Extraction d'information ; Annotation ; Reconnaissance d'entités nommées ;	Oui	http://brat.nlplab.org/	Japon ; Royaume-Uni ;
Bulstem	Stemming for Bulgarian	Libre	Racinisation ;	Oui	https://github.com/peio/PyBulStem	États-Unis ;
Caffe2	A new lightweight, modular, and scalable deep learning framework	Libre	Apprentissage profond ;	Non	https://caffe2.ai	États-Unis ;
Calliope	Logiciel d'analyse des tendances et de "fouille de textes"	Libre/Support technique et formation payants	Extraction terminologique ; Analyse de tendances ; Analyse des co-occurrences ;	Non	https://www.calliope-textmining.com/	France ;

Canopy	Enthought Canopy provides a proven scientific and analytic Python package distribution plus key integrated tools for iterative data analysis, data visualization, and application development. Users have the ability to extend and innovate with scripting and open platform APIs, driving the creation and sharing of innovative workflows, tools, and applications.	Commercial	Appentissage automatique ;	Non	https://store.enthought.com/basket/	États-Unis ;
Carrot2	Carrot2 organizes your search results into topics. With an instant overview of what's available, you will quickly find what you're looking for.	Libre	Clustering ; Recherche d'informations ;	Non	http://search.carrot2.org/stable/search	Pologne ; Royaume-Uni ;
Chemicalize	Chemicalize is a powerful online platform for chemical calculations, search, and text processing.	Libre	Annotation ; Reconnaissance d'entités nommées ;	Non	https://chemicalize.com/welcome	Hongrie ;
CiteSpace	Visualizing Patterns and Trends in Scientific Literature	Libre	Clustering ;	Non	http://cluster.cis.drexel.edu/~cchen/citespace/	États-Unis ;
Clarabridge CX Suite	CX Analytics is the backbone of the world's most complex Customer Experience Management Programs, providing the industry's most accurate Natural Language Processing (NLP), sentiment and data categorization, making issues transparent—and next steps clear. CX social : Social listening, rapid social media engagement, and social media analytics that empower teams of all sizes to wow customers and have a big impact.	Commercial	Clustering ; Analyse de sentiments ; Parsing ; Extraction de relations ;	Non	https://www.clarabridge.com/	États-Unis ;
ClearNLP	The ClearNLP project provides fast and robust NLP components implemented in Java	Libre	Tokenisation ; Sentence splitting ; Parsing ; Etiquetage de rôles sémantiques ; PoS tagging ;	Oui	http://clearnlp.wikispaces.com/	États-Unis ;
ClearTK	ClearTK is a framework for developing machine learning and natural language processing components within the Apache Unstructured Information Management Architecture.	Libre	Classification ; Clustering ; Parsing ; Racinisation ; Tokenisation ; PoS tagging ; Feature extraction ;	Non	https://cleartk.github.io/cleartk/about.html	États-Unis ;
Clementine	Clementine packages a number of tools with a GUI which simplifies the process of performing a data mining project. In particular the Clementine workbench supports a number of data mining algorithms through a simple linked node interface supporting the entire business process of data mining using the CRISP-DM model.	Commercial	Clustering ; Classification ; Analyse de régression ;	Non	http://datamining.togaware.com/survivor/Summary20.html	États-Unis ;

CLUTO	CLUTO is a software package for clustering low- and high-dimensional datasets and for analyzing the characteristics of the various clusters. CLUTO is well-suited for clustering data sets arising in many diverse application areas including information retrieval, customer purchasing transactions, web, GIS, science, and biology.	Libre	Clustering ;	Non	http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview	États-Unis ;
CMSR Data Miner	Integrated environment for predictive modeling, segmentation, data visualization, statistical data analysis, and rule-based model evaluation	Libre	Classification ; Clustering ;	Non	http://www.roselladb.com/starprobe.htm	Australie ;
CogComp-NLP	CogComp-NLP provides a suite of state-of-the-art Natural Language Processing (NLP) tools that allows you to annotate plain text inputs.	Libre	Lemmatisation ; PoS tagging ; Parsing ; Extraction de relations ; Reconnaissance d'entités nommées ; Etiquetage de rôles sémantiques ; Analyse de co-référence ;	Non	http://nlp.cogcomp.org/	États-Unis ;
Cogito	Multilingual text analytics, cognitive technology software that understands the meaning of words in context.	Commercial	Reconnaissance d'entités nommées ; Classification ; Recherce d'informations ; Désambiguisation ; Extraction de mots clés ; Découverte de connaissances ;	Non	http://www.expertsystem.com/fr/	Italie ;
Cognitive Computation Group NLP Tools		Libre	Reconnaissance d'entités nommées ; PoS tagging ; Chunking ; Lemmatisation ; Etiquetage de rôles sémantiques ; Résolution de coréférence ;	Non	http://cogcomp.org/page/software/	États-Unis ;
Coheris SPAD integral	Pour l'analyse de données et le traitement de toute l'information, notamment l'information textuelle.	Commercial	Tokenisation ; Lemmatisation ; Classification ; Clustering ; Arbre de décision	Non	https://www.coheris.com/produits/analytics/logiciel-data-mining/analyse-de-donnees/	France ;
ConcQuest	Concordancier dédié à la recherche d'expressions complexes à travers des corpus monolingues et multilingues alignés.	Libre	Concordancier ; Rercherche d'informations ;	Non	http://olivier.kraif.u-grenoble3.fr/index.php?option=com_content&task=view&id=36&Itemid=55	France ;
Content Annotation Manager	Create relevant metadata based on vocabularies and rules	Commercial	Annotation ; Reconnaissance d'entités nommées ;	Non	http://www.mondeca.com/content-annotation-manager/	France ;
ContentMine	Text and data mining tools	Libre	Extraction de connaissances ;	Non	http://www.contentmine.org/text-and-data-mining-tools/	Royaume-Uni ;
CORICO	Outil de visualisation de données multifactorielles sans équivalent. A partir d'un tableau de données, "L'Iconographie des Corrélations" élimine les "fausses bonnes corrélations" (celles qui sont dues à une tierce variable), et révèle les corrélations "masquées" lorsqu'une variable dépend de plusieurs variables.	Commercial	Découverte de connaissances ;	Non	http://www.coryent.com/coricо.html	France ;
Cortex Manager	CorText proposes a full ecosystem of modeling and exploratory tools for analyzing text corpora.		Reconnaissance d'entités nommées ; Topic modeling ; Extraction terminologique ; Apprentissage profond ; Clustering ; Analyse de sentiment ; Indexation ; Plongement de mots ;	Non	https://www.cortext.net/projects/cortext-manager/	France ;

Databionic ESOM Tools	The Databionic ESOM Tools is a suite of programs to perform data mining tasks like clustering, visualization, and classification with Emergent Self-Organizing Maps	Libre	Classification ; Clustering ;	Non	http://databionic-esom.sourceforge.net/	Allemagne ;
Dataiku	Dataiku DSS is the collaborative data science software platform for teams of data scientists, data analysts, and engineers to explore, prototype, build, and deliver their own data products more efficiently.	Libre/Commercial	Clustering ;	Non	https://www.dataiku.com/	États-Unis ;
DataMelt	Free mathematics software for scientists, engineers and students. It can be used for numeric computation, statistics, symbolic calculations, data analysis and data visualization.	Libre/Commercial	Classification ; Clustering ; Analyse de régression ;	Non	http://jwork.org/dmelt/	Allemagne ;
DataPreparator	DataPreparator is a free software tool designed to assist with common tasks of data preparation (or data preprocessing) in data analysis and data mining.	Libre	Prétraitement ;	Non	http://www.datapreparator.com/	Australie ;
Datumbox	The Datumbox Machine Learning Framework is an open-source framework written in Java which allows the rapid development of Machine Learning and Statistical applications.	Libre	Classification ; Clustering ; Analyse de régression ; Analyse de sentiments ; Détection de la langue ; Extraction de mots clés ; Extraction de textes ;	Non	http://www.datumbox.com/machine-learning-framework/	Grèce ;
DBpedia Spotlight	It is a tool for automatically annotating mentions of DBpedia resources in text, providing a solution for linking unstructured information sources to the Linked Open Data cloud through DBpedia.	Libre	Annotation ;	Non	https://www.dbpedia-spotlight.org	Allemagne ;
Deeplearning4j	Eclipse Deeplearning4j is the first commercial-grade, open-source, distributed deep-learning library written for Java and Scala. Integrated with Hadoop and Apache Spark, DL4J brings AI to business environments for use on distributed GPUs and CPUs.Deep learning	Libre	Apprentissage profond	Non	https://deeplearning4j.org/	États-Unis ;
DeLFT	DeLFT (Deep Learning Framework for Text) is an Open Source Keras framework for text processing, covering sequence labeling (e.g. named entity tagging) and text classification (e.g. comment classification)	Libre	Apprentissage profond ; Reconnaissance d'entités nommées ; Classification ; Analyse de sentiments ; Analyse d'opinions	Non	http://science-miner.com/deep-learning/	France ;
Diction	DICTION 7 is a computer-aided text analysis program for determining the tone of a verbal message	Commercial	Analyse de sentiments ;	Non	https://www.dictionsoftware.com/	États-Unis ;
Digimind	Social media analytics	Commercial	Clustering ; Traduction automatique ; Annotation ; Analyse de sentiments ; Recherche d'informations ;	Non	http://www.digimind.com/fr/	France ;

Discovertext	With dozens of powerful text analytics, data science, human coding, and machine-learning features, including instant access to the Gnip PowerTrack 2.0 for Twitter and the free Twitter Search API, DiscoverText provides cloud-based software tools to quickly evaluate large amounts of text, survey, and Twitter data.	Commercial	Classification ;	Non	https://discovertext.com/	États-Unis ;
DKPro Core	A collection of software components for natural language processing (NLP) based on the Apache UIMA framework.	Libre	PoS tagging : Tokenisation ; Parsing ; Lemmatisation ; Etiquetage des rôles sémantiques ; Segmentation ; Reconnaissance d'entités nommées ; Chunking ; Racinisation ; Identification de langue ; Résolution de coréférence ; Apprentissage profond ; Analyse morphologique ; Clustering ; Annotation ;	Oui	https://dkpro.github.io/dkpro-core/	Allemagne ;
Dlib	Dlib is a modern C++ toolkit containing machine learning algorithms and tools for creating complex software in C++ to solve real world problems.	**Libre**	Apprentissage profond ; Classification ; Clustering ;	**Non**	http://dlib.net/	**États-Unis ;**
Dtm-Vic	Statistique Exploratoire Multidimensionnelle pour données complexes comprenant des données numériques et textuelles.	Libre	Classification ;	Non	http://www.dtmvic.com/05_SoftwareF.html	France ;
Egas	**Collaborative biomedical text annotation.**	**Libre**	**Annotation ;** **Extraction de concepts ;** **Extraction de relations**	**Non**	**https://demo.bmd-software.com/egas/**	**Portugal ;**
ELAN	ELAN is a professional tool for the creation of complex annotations on video and audio resources.	Libre ?	Annotation ;	Non	https://tla.mpi.nl/tools/tla-tools/elan/	Pays-Bas ;
ELKI	**ELKI is an open source (AGPLv3) data mining software written in Java. The focus of ELKI is research in algorithms, with an emphasis on unsupervised methods in cluster analysis and outlier detection.**	**Libre**	**Clustering ;**	**Non**	**https://elki-project.github.io/**	**Allemagne ;** **Danemark ;**
Elsevier Text Mining	Enables the retrieval of highly specified information from unstructured content, providing more meaningful answers to complex research questions.	Commercial	Recherche d'informations ; Reconnaissance d'entités nommées ; Extraction de relations ;	Non	https://www.elsevier.com/solutions/professional-services/text-mining	Pays-Bas ;

Enju	**A deep syntactic parser for English**	**Libre**	**Parsing ;**	**Oui**	**http://www.nactem.ac.uk/enju/index.html**	**Royaume-Uni ;**
EnjuParser	Enju is a syntactic parser for English	Libre	Parsing ;	Oui	http://pubannotation.org/annotators/EnjuParser	Japon ;
Entity fishing	**entity-fishing is an open source tool dedicated to the automatic identification and disambiguation of Wikidata entities in multilingual text and PDF documents. The tool is based on machine-learning techniques (Gradient Tree Boosting, CRF, embeddings) exploiting Wikipedia as training source.**	**Libre**	**Reconnaissance d'entités nommées ; Apprentissage automatique ;**	**Non**	**http://science-miner.com/entity-disambiguation/**	**France ;**
Etuma	Etuma text analysis service turns all your open-ended customer feedback into consistent and actionable information.	Commercial	Classification ;	Non	http://www.etuma.com/home	Finlande ;
EventMine	**Event extraction system for biomedical text**	**Libre**	**Extraction d'évènements ;**	**Non**	**http://nactem.ac.uk/EventMine/**	**Royaume-Uni ;**
Expernova	Expernova utilise des algorithmes sophistiqués, s'appuyant sur le Big Data et le Machine Learning, pour connecter les réseaux d'innovation et dessiner un panorama global.	Commercial	Apprentissage automatique ;	Non	https://fr.expernova.com/	France ;
Fastr	**Automatic indexing**	**Libre**	**Indexation ;**	**Non**	**https://perso.limsi.fr/jacquemi/FASTR/**	**France ;**
FastText	Library for efficient text classification and representation learning	Libre	Classification de textes ; Apprentissage prodond ;	Non	https://fasttext.cc/	États-Unis ;
FreeLing	**An Open-Source Suite of Language Analyzers**	**Libre**	**Reconnaissance d'entités nommées ; Annotation ;**	**Oui**	**http://nlp.lsi.upc.edu/freeling/demo/demo.php**	**Espagne ;**
Galaxy	Galaxy is an open, web-based platform for accessible, reproducible, and transparent computational biomedical research.	Libre	Architecture logicielle ;	Oui	https://galaxyproject.org/	États-Unis ;
Gargantex	**A web platform to explore text-mining.**	**Libre**	**Textométrie ;**	**Non**	**https://gargantext.org/**	**France ;**
GATE	Suite of tools written in Java, used for human language processing, analysis, and information extraction.	Libre	Tokenisation ; Segmentation ; Chunking ; Résolution de coréférence ; Reconnaissance d'entités nommées ; Sentence splitting ; Annotation ; Analyse morphologique ; PoS tagging ; Classification de textes ; Apprentissage automatique ;	Oui	https://gate.ac.uk/	Royaume-Uni ;
Genia tagger	**Part-of-speech tagging, shallow parsing, and named entity recognition for biomedical text**	**Libre**	**PoS tagging ; Reconnaissance d'entités nommées ; Parsing ;**	**Oui**	**http://www.nactem.ac.uk/GENIA/tagger/**	**Royaume-Uni ;**
Gensim	Scalable statistical semantics ; Analyze plain-text documents for semantic structure ; Retrieve semantically similar documents	Libre	Clustering ;	Non	https://radimrehurek.com/gensim/	République Tchèque ;

GibbsLDA	**GibbsLDA++ is a C/C++ implementation of Latent Dirichlet Allocation (LDA) using Gibbs Sampling technique for parameter estimation and inference. It is very fast and is designed to analyze hidden/latent topic structures of large-scale datasets including large collections of text/Web documents**	**Libre**	**Clustering ;**	**Non**	**http://gibbslda.sourceforge.net/**	**Japon ; Vietnam ;**
GLoVe	GloVe is an unsupervised learning algorithm for obtaining vector representations for words.	Libre	Apprentissage profond ; Plongement de mots ;	Non	https://nlp.stanford.edu/projects/glove/	États-Unis ;
Glozz	**Plateforme d'annotation**	**Libre**	**Annotation ;**	**Non**	**http://www.glozz.org/**	**France ;**
GNU PSPP	GNU PSPP is a program for statistical analysis of sampled data	Libre		Non	https://www.gnu.org/software/pspp/	?
Google Cloud Natural Language API	**L'API Google Cloud Natural Language révèle la structure et la signification des textes grâce à des modèles de machine learning puissants, dans une API REST conviviale.**	**Commercial**	**Reconnaissance d'entités nommées ; Analyse de sentiments ; Parsing ; Classification ; Apprentissage profond ;**	**Non**	**https://cloud.google.com/natural-language/**	**États-Unis ;**
Google Prediction API	L'API Prediction de Google propose des fonctionnalités de filtrage par motif et de machine learning.	Commercial	Classification ;	Non	https://cloud.google.com/prediction/	États-Unis ;
GROBID	**GROBID is an Open Source tool for parsing and extracting structured information from technical and scientific documents in raw format like PDF**	**Libre**	**Prétraitement ; Normalisation ; Structuration de documents ; Reconnaissance d'entités nommées ;**	**Non**	**http://science-miner.com/document-engineering/**	**France ;**
Heart of Gold	Middleware architecture for the integration of deep and shallow natural language processing components. It provides a uniform and flexible infrastructure for building applications that use Robust Minimal Recursion Semantics (RMRS) and/or general XML standoff annotation produced by natural language processing components.	Libre	Annotation ;	Non	http://heartofgold.dfki.de/	Allemagne ;
HunPos tagger	**Hunpos is an open source reimplementation of TnT, the well known part-of-speech tagger by Thorsten Brants.**	**Libre**	**PoS tagging ;**	**Oui**	**http://mokk.bme.hu/resources/hunpos/**	**Hongrie ;**
Hyperbase	Logiciel hypertexte pour le traitement documentaire et statistique des corpus textuels	Libre	Textométrie ; Concordancier ; Lemmatisation ; Classification ; Clustering ;	Non	http://bcl.cnrs.fr/article69?redirected_from=www%252eunice%252efr%252fbcl%252farticle69	France ;
IBM SPSS Modeler	**Versatile data and text analytics workbench that helps you build accurate predictive models quickly and intuitively, without programming.**	**Commercial**	**Extraction d'informations ; Reconnaissance d'entités nommées ; Extraction de relations ;**	**Non**	**http://www.spss.com.hk/software/modeler/**	**États-Unis ;**

IBM Watson Natural Language Understanding	Analyze text to extract meta-data from content such as concepts, entities, keywords, categories, relations and semantic roles. Returns both overall sentiment and emotion for a document, and targeted sentiment and emotion towards keywords in the text for deeper analysis.	Commercial	Extraction d'informations ; Reconnaissance d'entités nommées ; Extraction de relations ; Etiquetage de rôles sémantiques ; Analyse de sentiments ; Détection de la langue ;	Non	https://www.ibm.com/watson/services/natural-language-understanding/	États-Unis ;
ILSP NLP	**Natural Language Processing services developed by the NLP group of the Institute for Language and Speech Processing**	**Libre**	**PoS tagging ; Lemmatisation ; Parsing ; Chunking ; Sentence splitting ; Reconnaissance d'entités nommées ; Tokenisation ;**	**Oui**	**http://nlp.ilsp.gr/soaplab2-axis/**	**Grèce ;**
ILSP NLP Web Services	Natural Language Processing services developed by the NLP group of the Institute for Language and Speech Processing	Libre	Parsing ; Chuncking ; Lemmatisation ; Reconnaissance d'entités nommées ;	Oui	http://nlp.ilsp.gr/soaplab2-axis/	Grèce ;
INCEpTION	**Semantic annotation**	**Libre**	**Annotation ;**	**Oui**	**https://inception-project.github.io**	**Allemagne ;**
Indico	Text and image analysis to create transformative tools.	Commercial	Apprentissage automatique ;	Non	https://indico.io/	États-Unis ;
Intellexer	**Based on the use of Natural Language Processing and Machine Learning technologies, tools for text analytics solutions that can be used as standalone applications as well as integrated in the existing systems.**	**Commercial**	**Recherche d'informations ; PoS tagging ; Segmentation ; Parsing ; Extraction de relations ; Analyse de sentiments ; Reconnaissance d'entités nommées ; Résumé automatique ; Q&A ; Classification ; Clustering ; Détection de la langue ; Vérification de l'orthographe ;**	**Non**	**https://www.intellexer.com/products.html**	**États-Unis ;**
Iramuteq	Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires.	Libre	Classification ;	Non	http://iramuteq.org/	France ;
Java Automatic Term Extraction	**Java Automatic Term Extraction toolkit - a library of state-of-the-art term extraction algorithms and framework for developing term extraction algorithms**	**Libre**	**Extraction terminologiques ;**	**Non**	**https://code.google.com/archive/p/jatetoolkit/**	**Royaume-Uni ;**
Jazzy	Java Spell Check API	Libre	Vérification de l'orthographe ;	Oui	https://sourceforge.net/projects/jazzy/	?
JCoRE	**The JULIE Lab Component Repository (JCoRe) is an open software repository for full-scale natural language processing based on the UIMA middleware framework.**	**Libre**	**Sentence segmentation ; Tokenisation ; PoS tagging ; Reconnaissance d'entités nommées ;**	**Non**	**http://julielab.github.io/**	**Allemagne ;**
Jubatus	Online distributed machine learning on the data streams of Big Data	Libre	Classification ; Clustering ; Analyse de régression ;	Non	http://jubat.us/en/overview.html	Japon ;

KAF annotator	**Stand-alone application for annotating KAF files with any set of tags to any level. This annotator is used to create gold-standard data for evaluating the Kybot output.**	**Libre**	**Annotation ;**	**Non**	**http://kyoto-project.eu/xmlgroup.iit.cnr.it/kyoto/index2091.html?option=com_content&view=article&id=**	**Europe ; Japon ;**
KEA	KEA is an algorithm for extracting keyphrases from text documents. It can be either used for free indexing or for indexing with a controlled vocabulary.	Libre	Extraction de mots clés ;	Oui	http://community.nzdl.org/kea/	Nouvelle-Zélande ;
Keatext	**Keatext is an AI – driven text analytics technology that makes it easy for you to analyze large volumes of unstructured customer feedback**	**Commercial**	**Analyse de sentiments ; Analyse d'opinions ;**	**Non**	**https://www.keatext.ai/**	**Canada ;**
KEEL	KEEL (Knowledge Extraction based on Evolutionary Learning) is an open source (GPLv3) Java software tool that can be used for a large number of different knowledge data discovery tasks.	Libre	Découverte de connaissances	Non	http://www.keel.es/	Espagne ;
Keyterm	**rank extracted terms and concepts according to a relevance score estimating how well they capture the most important and discriminant subject matter of a document given a background collection of technical and scientific documents**	**Libre**	**Extraction de phrases clés**	**Non**	**http://science-miner.com/term-and-concept-relevance/**	**France ;**
KH coder	KH Coder is a free software for quantitative content analysis or text mining. It is also utilized for computational linguistics. You can analyze Japanese, English, French, German, Italian, Portuguese and Spanish text with KH Coder. Also, Chinese (simplified), Korean and Russian language data can be analyzed with the latest Alpha release (Version 3).	Libre	Clustering ; Concordancier ; Réseau de co-occurrences ;	Non	http://khc.sourceforge.net/en/	Japon ;
KIWI	**Keyword extractor**	**Libre**	**Extraction de mots clés ;**	**Non**	**http://www.quaero.org/module_technologique/kiwi-keyword-extractor/**	**France ;**
KNIME	Open source data analytics, reporting and integration platform.	Libre	Apprentissage automatique ;	Non	https://www.knime.com/	Suisse ;
KnowledgeREADER	**Integrated customer intelligence by combining visual text discovery and sentiment analysis with the power of predictive analytics**	**Commercial**	**Analyse de sentiments ;**	**Non**	**http://www.angoss.com/**	**Canada ;**

LanguageComputer (Cicero, Ferret)	Understanding the information stored in any large collections of text.	Commercial	Reconnaissance d'entités nommées ; Résolution de coréférence ; Annotation ; Extraction de relations ; Extraction d'évènements ; Q&A ;	Non	http://www.languagecomputer.com/	États-Unis ;
LAPPS Grid	**An open, interoperable web service platform for natural language processing (NLP) research and development**	**Libre**	**Architecture logicielle ; Tokenisation ; Reconnaissance d'entités nommées ; PoS tagging ; Sentence splitting ; Parsing ; Chunking ;**	**Non**	**http://www.lappsgrid.org/**	**États-Unis ;**
Lavastorm analytics engine	Visual data discovery solution	Libre	Prétraitement ;	Non	http://www.lavastorm.com/	États-Unis ;
Le Trameur	**Le Trameur est un programme d'analyse comportant de nombreuses fonctionnalités pour l'analyse automatique, statistique et documentaire de textes en vue de leur profilage sémantique, thématique et de leur interprétation. Ce logiciel est à l'origine un outil de textométrie : il intègre les fonctionnalités classiques de ce type d'outils dans ce domaine. Il dispose aussi des fonctionnalités particulières qui permettent d'annoter dynamiquement des corpus ou d'explorer des ressources richement annotées (treebanks monolingues/multilingues ou des alignements).**	**Libre**	**Textométrie ; Annotation**	**Non**	**http://www.tal.univ-paris3.fr/trameur/#p4**	**France ;**
Lexalytics (Semantria)	Natural language processing	Commercial	PoS tagging ; Extraction de relations ; Classification ; Tokenisation ; Extraction de relations ; Analyse d'opinions ; Analyse de sentiments ; Résolution d'anaphore ; Racinisation ; Reconnaissance d'entités nommées ;	Non	https://www.lexalytics.com/	États-Unis ;
Lexi-co	**Analyses textometriques**	**Libre**	**Textométrie ;**	**Non**	**http://www.lexi-co.com/index.html**	**France ;**
Leximancer	Leximancer automatically analyses your text documents to identify the high level concepts in your text documents, delivering the key ideas and actionable insights you need with powerful interactive visualisations and data exports.	Commercial	Extraction d'informations ;	Non	https://info.leximancer.com/	États-Unis ;
Liblinear	**Library for large linear classification**	**Libre**	**Classification ;**	**Non**	**http://www.csie.ntu.edu.tw/~cjlin/liblinear/**	**Taïwan ;**
LibShortText	A Library for Short-text Classification and Analysis	Libre	Classificationde textes ; Tokénisation ; Racinisation ;	Non	https://www.csie.ntu.edu.tw/~cjlin/papers/libshorttext.pdf	Taïwan ;

LIBSVM	**Library for support vector machine**	**Libre**	**Classification ;**	**Oui**	**http://www.csie.ntu.edu.tw/~cjlin/libsvm/**	**Taïwan ;**
LingPipe	LingPipe is tool kit for processing text using computational linguistics.	Libre/Commercial	Reconnaissance d'entités nommées ; Détection de la langue ; Classification ; Sentence splitting ; Tokenisation ; PoS tagging ;	Oui	http://alias-i.com/lingpipe/index.html	États-Unis ;
Linguistic Inquiry Word Count	**LIWC2015 is the gold standard in computerized text analysis. Learn how the words we use in everyday language reveal our thoughts, feelings, personality, and motivations.**	**Commercial**	**Annotation ; Classification ;**	**Non**	**http://liwc.wpengine.com/**	**États-Unis ;**
LIONoso	Integrated tool for Machine Learning and Intelligent Optimization	Commercial	Apprentissage automatique ;	Non	http://lionoso.com/	Italie ;
LPU	**LPU (which stands for Learning from Positive and Unlabeled data) is a text learning or classification system that learns from a set of positive documents and a set of unlabeled documents (without labeled negative documents). This type of learning is different from classic text learning/classification, in which both positive and negative training documents are required.**	**Libre**	**Classification ;**	**Non**	**https://www.cs.uic.edu/~liub/LPU/LPU-download.html**	**États-Unis ;**
Luminoso	Understand, measure and act on large amounts of unstructured text.	Commercial	Apprentissage automatique ; Classification ; Analyse de tendances ;	Non	https://luminoso.com/	États-Unis ;
Magaputer	**Data and text mining solutions**	**Commercial**	**Classification ; Clustering ; Analyse de régression ; Reconnaissance d'entités nommées ; Extraction de relations ; Détection de la langue ; PoS tagging ; Extraction de mots clés ; Parsing ; Analyse de sentiments ; Résolution d'anaphore ;**	**Non**	**http://megaputer.com/site/index.php**	**États-Unis ;**
MALLET	MALLET is a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text.	Libre	Clustering ; Classification ;	Oui	http://mallet.cs.umass.edu/	États-Unis ;
MaltParser	**MaltParser is a system for data-driven dependency parsing, which can be used to induce a parsing model from treebank data and to parse new data using an induced model**	**Libre**	**Parsing ;**	**Oui**	**http://www.maltparser.org/**	**Suède ;**
Massive Online Analysis (MOA)	Open source framework for data stream mining	Libre	Clustering ; Classification ; Analyse de régression ;	Non	https://moa.cms.waikato.ac.nz/	Nouvelle-Zélande ;
MATE	**Multilevel Annotation, Tools Engineering**	**Libre**	**Annotation ;**	**Oui**	**http://xml.coverpages.org/mate.html**	**Royaume-Uni ; Allemagne ; Danemark ;**

MatheoSoftware	Patents search and analysis, technological trends, data analysis	Commercial	Recherche d'informations ; Analyse de tendances ;	Non	https://www.matheo-software.com/	France ;
MATLAB	**Analyse de données, développement d'algorithmes et création de modèles mathématiques, deep learning**	**Commercial**	**Clustering ; Classification ; Analyse de régression ; Apprentissage profond ;**	**Non**	**https://fr.mathworks.com/?s_tid=gn_logo**	**États-Unis ;**
Meaning Cloud	Extract the meaning of all kind of unstructured content: social conversation, articles, documents...	Libre/Commercial	Analyse de sentiments ; Analyse de tendance ; Clustering ; Classification de textes ; Résumé automatique ; Détection de la langue ; Lemmatisation ; PoS tagging ; Etiquetage morphosyntaxique ;	Non	https://www.meaningcloud.com/	États-Unis ;
MeCab		**Libre**	**Parsing ;**	**Oui**	**https://taku910.github.io/mecab/libmecab.html**	**Japon ;**
MER	Minimal name entity recognizer	Libre	Reconnaissance d'entités nommées ;	Non	https://github.com/lasigeBioTM/MER	Portugal ;
Merlin	**Deep learning**	**Libre**	**Apprentissage profond ;**	**Non**	**http://www.cstr.ed.ac.uk/projects/merlin/**	**Royaume-Uni ;**
MetaMAp	Tool for recognizing UMLS concepts in texts	Libre	Annotation ;	Non	https://metamap.nlm.nih.gov/	États-Unis ;
MicroFocus	**Big Data and analytics software**	**Commercial**	**Recherche d'informations ; Découverte de connaissances ; Apprentissage automatique ;**	**Non**	**https://software.microfocus.com/en-us/software/big-data-analytics-software**	**Royaume-Uni ;**
Microsoft Azure	Détectez le sentiment, les phrases clés, les sujets et la langue du texte	Commercial	Analyse de sentiments ; Extraction de mots clés ; Détection de la langue ; Annotation sémantique ;	Non	https://azure.microsoft.com/fr-fr/services/cognitive-services/text-analytics/	États-Unis ;
Microsoft Cognitive ToolKit	**The Microsoft Cognitive Toolkit (CNTK) is an open-source toolkit for commercial-grade distributed deep learning. It describes neural networks as a series of computational steps via a directed graph. CNTK allows the user to easily realize and combine popular model types such as feed-forward DNNs, convolutional neural networks (CNNs) and recurrent neural networks (RNNs/LSTMs). CNTK implements stochastic gradient descent (SGD, error backpropagation) learning with automatic differentiation and parallelization across multiple GPUs and servers.**	**Libre**	**Apprentissage profond ;**	**Non**	**https://docs.microsoft.com/en-us/cognitive-toolkit/**	**États-Unis ;**

Microsoft Distributed Machine Learning Toolkit	Distributed machine learning has become more important than ever in this big data era. Especially in recent years, practices have demonstrated the trend that more training data and bigger models tend to generate better accuracies in various applications. However, it remains a challenge for common machine learning researchers and practitioners to learn big models from huge amount of data, because the task usually requires a large number of computation resources. In order to tackle this challenge, we release the Microsoft Distributed Machine Learning Toolkit (DMTK), which contains both algorithmic and system innovations.	Libre ?	Topic modeling ; Apprentissage automatique ; Apprentissage profond	Non	http://www.dmtk.io/index.html	États-Unis ;
Microsoft SQL Server Analysis Services	**Ensemble d'outils pour la fouille de données**	**?**	**Classification ; Clustering ; Anayse de régression ;**	**Non**	**https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/data-mining-tools**	**États-Unis ;**
MiningMart	Prétraitement des données	Libre	Prétraitement ;	Non	http://mmart.cs.uni-dortmund.de/research/index.html	Allemagne ;
MiniPar		**Libre**	**Parsing ;**	**Oui**	**https://webdocs.cs.ualberta.ca/~lindek/minipar.htm**	**Canada ;**
ML-Flex	an open-source software package designed to enable flexible and efficient processing of disparate data sets for machine-learning (classification) analyses	Libre	Classification ;	Non	http://mlflex.sourceforge.net/	États-Unis ;
MLPACK	**Scalable machine learning library, written in C++**	**Libre**	**Classification ; Clustering ; Analyse de régression ;**	**Non**	**http://mlpack.org/**	**États-Unis ;**
mlpy	mlpy provides a wide range of state-of-the-art machine learning methods for supervised and unsupervised.	Libre	Classification ; Clustering ;	Non	http://mlpy.sourceforge.net/	Italie ;
Modalisa	**Création de questionnaires en ligne ou papier, diffusion et recueil de données, transformation de variables, codification de textes, analyses univariées et multivariées, indicateurs spécifiques, régressions, rapports dynamiques exportables sous PowerPoint, export et import des données sous format Excel et Texte...**	**Commercial**	**Classification**	**Non**	**https://modalisa.com/logiciel/modalisa.php**	**France ;**

Modular toolkit for Data Processing	Modular toolkit for Data Processing (MDP) is a library of widely used data processing algorithms that can be combined according to a pipeline analogy to build more complex data processing software.	Libre	Classification ; Clustering ;	Non	https://pypi.python.org/pypi/MDP/2.4	États-Unis ; Allemagne ;
Monkeylearn	**Text Analysis with machine learning**	**Libre/commercial**	**Reconnaissance d'entités nommées ; Analyse de sentiments ; Extraction de topics ; Apprentissage automatique ; Prétraitement ;**	**Non**	**https://monkeylearn.com/**	**États-Unis ;**
MorphAdoerner	MorphAdorner is a Java command-line program which acts as a pipeline manager for processes performing morphological adornment of words in a text	Libre	Tokenisation ; PoS tagging ; Reconnaissance d'entités nommées ; Sentence splitting ; Lemmatisation ; Annotation ;	Non	http://morphadorner.northwestern.edu/morphadorner/	États-Unis ;
MSTParser	**MSTParser is a non-projective dependency parser that searches for maximum spanning trees over directed graphs. Models of dependency structure are based on large-margin discriminative training methods. Projective parsing is also supported.**	**Libre**	**Analyse en dépendances**	**Non**	**https://www.seas.upenn.edu/~strctlrn/MSTParser/MSTParser.html**	**États-Unis ;**
MutationFinder	MutationFinder is a biomedical natural language processing (NLP) system for extracting mentions of point mutations from free text. MutationFinder achieves high performance (99% precision, 81% recall on blind test data) as an information extraction system	Libre	Extraction d'informations ;	Oui	https://sourceforge.net/projects/mutationfinder/	États-Unis ;
mutext	**Analytics and decision science solutions.**	**Commercial**	**Parsing ; Classification de textes ; Clustering ;**	**Non**	**https://www.mu-sigma.com/**	**Inde ;**
MXNet	A flexible and efficient library for deep learning	Libre	Apprentissage profond ;	Non	https://mxnet.incubator.apache.org	États-Unis ;
NaCTeM Software Tools	**The National Centre for Text Mining bases its service systems on a number of text mining software tools.**	**Libre**	**Reconnaissance d'entités nommées ; PoS tagging ; Parsing ; Sentencfe splitting ; Paragraph splitting ; Extraction d'évènements ; Annotation ;**	**Oui**	**http://nactem.ac.uk/software.php**	**Royaume-Uni ;**

Narrative Science Quill	Quill transforms data into automated, human-sounding Intelligent Narratives that empower your people with insights to improve every aspect of your business.	Commercial	Génération de langage naturel ;	Non	https://narrativescience.com/	États-Unis ;
NaturalText	**For Life sciences: NaturalText's Machine Learning Algorithms can process Scientific Papers, Bio Sequences to find patterns and help scientists, researchers to advance their research. For financial sector : NaturalText's Machine Learning Algorithms can combine various data formats, cross verify for incorrect information, help companies to know more from the data**	**Commercial**	**Extraction d'informations ; Extraction de relations ; Apprentissage automatique ; Découverte de connaissances ;**	**Non**	**http://naturaltext.com/**	**Inde ;**
NCBI Text Mining Tools	Ensemble de applications web ou de bureau pour la fouille de textes dans le domaine biomédical	Libre	Annotation ; Recherche d'informations ; Reconnaissance d'entités nommées ; Normalisation ; Désambiguisation ;	Non	https://www.ncbi.nlm.nih.gov/research/bionlp/Tools/	États-Unis ;
NERD	**Named entity recognition and desambiguation**	**Libre**	**Reconnaissance d'entités nommées ;**	**Non**	**http://nerd.eurecom.fr/**	**France ;**
NERSuite	Named Entity Recognition toolkit	Libre	Reconnaissance d'entités nommées ;	Non	http://nersuite.nlplab.org/	Japon ;
Netowl	**Text and Entity Analytics Products**	**Commercial**	**Reconnaissance d'entités nommées ; Analyse de sentiments ;**	**Non**	**https://www.netowl.com/**	**États-Unis ;**
Neural designer	Neural Designer is a software tool for advanced analytics. It includes tools for descriptive, diagnostic, predictive and prescriptive analytics. It allows you to get actionable insights resulting in smarter decisions and better business outcomes. Neural networks are the most powerful method to discover intricate relationships, recognize complex patterns or predict current trends in your data.	Libre/commercial	Apprentissage automatique ;	Non	https://www.neuraldesigner.com/	Espagne ;
NeuroNER	**A Named-Entity Recognition program based on neural networks**	**Libre**	**Reconnaissance d'entités nommées ;**	**Non**	**http://neuroner.com/**	**États-Unis ;**
NLPCube	Natural Language Procecssing Toolkit with support for tokenization, sentence splitting, lemmatization, tagging and parsing for more than 60 languages	Libre	Tokenisation ; Segmentation ; Lemmatisation ; Parsing ; POS tagging ; Analyse en dépendances : Reconnaissance d'entités nommées ;	Non	https://pypi.org/project/nlpcube/	États-Unis ;
NLTK	**Platform for building Python programs to work with human language data**	**Libre**	**Parsing ; Chunking ; Concordiancier ; Classification ; Clustering ; Extraction de relations sémantiques ; Anlyse de sentiments ; Racinisation ; Tokenisation ; Traduction automatique ;**	**Non**	**http://www.nltk.org/**	**États-Unis ;**

NooJ	Linguistic development environment software as well as a corpus processor.	Libre	Annotation ;	Non	http://www.nooj4nlp.net/	France ;
Noopsis	**Noopsis automatise la collecte d'informations stratégiques par la fouille de documents textuels.**	**Commercial**	**Recherche d'informations ; Analyse sémantique ;**	**Non**	**http://www.noopsis.fr/index.fr.html**	**France ;**
ntent	Semantic search technology to determine user intent and the contextual meaning of words	Commercial	Recherche d'informations ; Reconnaissance d'entités nommées ; Détection de la langue ; Classification ; Lemmatisation ; Toklenisation ; Extraction d'informations ; Apprentissage automatique ;	Non	http://www.ntent.com/	États-Unis ;
Nvivo	**NVivo est un logiciel qui supporte des méthodes de recherches qualitatives et combinées. Il est conçu pour vous permettre d'organiser, analyser et trouver du contenu perspicace parmi des données non structurées ou qualitatives telles que des interviews, des réponses libres obtenues dans le cadre d'un sondage, des articles, des médias sociaux et des pages Web.**	**Commercial**	**Classification ; Visualisation**	**Non**	**http://www.qsrinternational.com/nvivo-french**	**Australie ;**
Odintext	Text analytics software	Commercial	Analyse de sentiments ; Apprentissage automatique ;	Non	http://odintext.com/	États-Unis ;
OGER	**Biomedical entity recognizer**	**Libre**	**Reconnaissance d'entités nommées ;**	**Oui**	**http://www.ontogene.org/resources/oger**	**Suisse ;**
OntoText	Ontotext provides a complete set of Semantic Technology enabling better content management, knowledge discovery and semantic search.	Commercial	Annotation sémantique ; Extraction de relations ; Désambiguisation ; Apprentissage automatique ; Classification ;	Non	https://ontotext.com/	Bulgarie ;
Open Calais	**Way to tag the people, places, companies, facts, and events in your content to increase its value, accessibility and interoperability.**	**Commercial**	**Annotation ; Reconnaissance d'entités nommées ; Extraction de relations ; Extraction d'évènements ; Clustering ;**	**Oui**	**http://www.opencalais.com/about-open-calais/**	**États-Unis ;**
Open semantic search	Free Software for your own Search Engine, Explorer for Discovery of large document collections, Media Monitoring, Text Analytics, Document Analysis & Text Mining platform based on Apache Solr or Elasticsearch open-source enterprise-search and Open Standards for Linked Data, Semantic Web & Linked Open Data integration	Libre	Recherche d'informations ; Annotation ;	Non	https://www.opensemanticsearch.org/	Allemagne ;

OpenCCG	**OpenCCG, the OpenNLP CCG Library, is an open source natural language processing library written in Java, which provides parsing and realization services based on Mark Steedman's Combinatory Categorial Grammar (CCG) formalism**	**Libre**	**Parsing ; PoS tagging ;**	**Oui**	**http://openccg.sourceforge.net/**	**États-Unis ;**
OpenMinTED	OpenMinted sets out to create an open, service-oriented e-Infrastructure for Text and Data Mining (TDM) of scientific and scholarly content. Researchers can collaboratively create, discover, share and re-use Knowledge from a wide range of text-based scientific related sources in a seamless way	Libre			http://openminted.eu/	Europe ;
OpenNER	**OpeNER's main goal is to provide a set of ready to use tools to perform some natural language processing tasks, free and easy to adapt for Academia, Research and Small and Medium Enterprise to integrate them in their workflow. More precisely, OpeNER aims to be able to detect and disambiguate entity mentions and perform sentiment analysis and opinion detection on the texts, to be able for example, to extract the sentiment and the opinion of customers about certain resource (e.g. hotels and accommodations) in Web reviews.**	**Libre**	**Reconnaissance d'entités nommées ; Analyse de sentiment ; Analyse d'opinion**	**Non**	**https://www.opener-project.eu/index.html**	**Italie ; Espagne ; Pays Bas ;**
OpenNN	OpenNN is an open source class library written in C++ programming language which implements neural networks, a main area of machine learning research.	Libre	Apprentissage profond ;	Non	http://www.opennn.net/	Espagne ;
OpenRefine	**Nettoyage, mise en forme et transformation de données**	**Libre**	**Prétraitement ;**	**Non**	**http://openrefine.org/**	**États-Unis ;**
OpenText	Digitize processes and discover the value in information using analytics and Artificial Intelligence.	Commercial	Recherche d'informations ; Classification ;	Non	http://www.opentext.com/	Canada ;
Oracle Data Mining	**Oracle Data Mining (ODM), a component of the Oracle Advanced Analytics Database Option, provides powerful data mining algorithms that enable data analytsts to discover insights, make predictions and leverage their Oracle data and investment.**	**Commercial**	**Classification ; Clustering ; Analyse de régression ; Détection d'anomalies ; Règle d'associations ;**	**Non**	**http://www.oracle.com/technetwork/database/options/advanced-analytics/odm/overview/index.html**	**États-Unis ;**
Orange Data Mining Toolbox	Open source machine learning and data visualization for novice and expert.	Libre	Clustering ; Classification ; Analyse de régression ;	Non	https://orange.biolab.si/	Slovénie ;

Overview	**Overview is a document mining application originally built for investigative journalists. It's also used for legal work, training machine learning models, and research of all types. It's a visualization and analysis tool designed for sets of documents, from dozens to millions of pages of materia**	**Libre**	**Clustering ; Recherche d'informations ; Annotation ; Reconnaissance d'entités nommées ;**	**Non**	**https://www.overviewdocs.com/**	**?**
Pagelyser	Pagelyzer is a tool which compares two web pages versions and decides if they are similar or not.	Libre	Analyse de pages web ;	Non	http://pagelyzer.openpreservation.org/	France ;
PANDAS	**Library providing high performance, easey-to-use data structure and data analysis tools for the Python programming language**	**Libre**	**Analyse de régression ;**	**Non**	**http://pandas.pydata.org/**	**États-Unis ;**
Pattern	Pattern is a web mining module for the Python programming language.	Libre	PoS tagging ; Analyse de sentiments ; Apprentissage automatique ; n-gram ; Clustering ;	Non	https://www.clips.uantwerpen.be/pages/pattern	Pays-Bas ;
Penelope	**Penelope is a cloud-based, open and modular platform that consists of tools and techniques for mapping landscapes of opinions expressed in online (social) media. The platform is used for analysing the opinions that dominate the debate on certain crucial social issues, such as immigration, climate change and national identity.**		**Tokenisation ; Lemmatisation ; Chunking ; PoS tagging ; Reconnaissance d'entités nommées ; Analyse en dépendances ; apprentissage automatique ; plongement lexical ; Sentencisation ;**	**Non**	**https://penelope.vub.be**	**Europe ;**
Philologic	PhiloLogic™ is the primary full-text search, retrieval and analysis tool developed by the ARTFL Project and the Digital Library Development Center (DLDC) at the University of Chicago. This is a Free Software implementation of PhiloLogic for large TEI-Lite document collections.	Libre	Recherche d'information ;	Non	https://sites.google.com/site/philologic3/home	États-Unis ;
Pingar	**Pingar DiscoveryOne Content Enrichment automatically tags and categorizes content. Typically, it is used to improve findability of information in an Electronic Content Management System (ECMS) by enabling faceted search.**	**Commercial**	**Classification de textes ; Recherche d'informations ; Découverte de connaissances ;**	**Non**	**http://pingar.com/**	**États-Unis ;**
PlaidML	Framework for making deep learning work everywhere	Libre	Apprentissage profond ;	Non	https://github.com/plaidml/plaidml	?
PoolParty Semantic Suite	**PoolParty is a world-class semantic technology suite that offers sharply focused solutions to knowledge organization and content business.**	**Commercial**	**Reconnaissance d'entités nommées ; Annotation ; Recherche d'informations ; Extraction de relations ; Extraction terminologique ; Classification de textes ;**	**Non**	**https://www.poolparty.biz/**	**Autriche ;**

PrediCX	Automated Text Analytics for Voice of Customer (VoC) data, chatbots, service desks, complaint handling, call center automation and early warning of issues. Quick and Easy to Deploy, High-Impact, AI and Machine Learning for Text Analysis.	Commercial	Apprentissage automatique ; Analyse de sentiments ;	Non	https://warwickanalytics.com/predicx/	Royaume-Uni ;
Prosuite	**ProSuite is an integrated collection of Provalis Research text analytics tools that allow one to explore, analyze and relate both structured and unstructured data. Provalis Research Text Analytics Software allows one to perform advanced computer assisted qualitative coding on documents and images using QDA Miner, to apply the powerful content analysis and text mining features of WordStat on textual data, and to perform advanced statistical analysis on numerical and categorical data using SimStat**	**Commercial**	**Clustering ; Concordancier ; Annotation ; Analyse de tendances ; Extraction de mots clés ; Classification de textes ;**	**Non**	**https://provalisresearch.com/products/prosuite/**	**Canada ;**
Proxem Studio	Transforme les données textuelles en prise de decision	Commercial	Classification de textes ; Recherche d'informations ; Découverte de connaissances ; Annotation ; Analyse de sentiments ;	Non	https://www.proxem.com/	France ;
PubRunner	**A framework for keeping biomedical text mining result up-to-date**	**Libre**	**Architecture logicielle ;**	**Oui**	**https://github.com/jakelever/pubrunner**	**États-Unis ;**
PubTator	PubTator is a Web-based tool for accelerating manual literature curation.	Libre	Annotation ;	Oui	https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/PubTator/index.cgi?user=User284144660	États-Unis ;
PyTorch	**PyTorch is a deep learning framework for fast, flexible experimentation.**	**Libre**	**Apprentissage profond ;**	**Non**	**https://pytorch.org/**	**États-Unis ; France ;**
Qlucore Omics Explorer	Identify patterns and structure when exploring biological data	Commercial	Classification ; Clustering ;	Non	https://www.qlucore.com/	Suède ;
QWAM	**Solutions logicielles métier répondant aux besoins de gestion des flux d'information (documentaire, textuelle, multimedia), de moteur de recherche, de veille et d'analyse et d'enrichissement sémantique**	**Commercial**	**Extraction d'informations ;**	**Non**	**http://www.qwamci.com/**	**France ;**
R	Free software environment for statistical computing and graphics.	Libre	Tokenisation ; Racinisation ; PoS tagging ; Parsing ; Reconnaissance d'entités nommées ; Analyse de sentiments ; Classification ; Clustering ; Apprentissage profond ;	Non	https://www.r-project.org/	Nouvelle-Zélande ; Canada ;

R.TeMIS	**Environnement graphique de travail sous R permettant de créer, manipuler et analyser des corpus de textes.**	**Libre**	**Racinisation ; Chunking ; Clustering ;**	**Non**	**http://rtemis.hypotheses.org/**	**France ;**
RapidMiner	RapidMiner is code free data science platform that unifies data prep, machine learning, and model deployment.	Libre/commercial	Tokenisation ; Racinisation ; Reconnaissance d'entités nommées ; Analyse de sentiments ; Classification ; Clustering ;	Non	https://rapidminer.com/	Allemagne ;
Rasp	**The RASP system includes state-of-the-art modules for finding sentence boundaries, finding individual words, analyzing words to identify the word root and any suffixes, assigning part-of-speech labels to words in running text, and analyzing the grammatical relations between words and larger units within sentences.**	**Libre/commercial**	**Tokenisation ; Lemmatisation ; PoS tagging ; Parsing ;**	**Oui**	**https://www.ilexir.co.uk/rasp/index.html**	**Royaume-Uni ;**
Rattle	GUI for data mining using R	Libre	Classification ; Clustering ;	Non	https://rattle.togaware.com/	États-Unis ;
RepKnight	**Software platform provides real-time cyber intelligence to keep people, companies and assets safe from internal and external threats**	**Commercial**	**Recherche d'informations ;**	**Non**	**https://www.repknight.com/**	**Royaume-Uni ;**
Resoomer	Résumé automatique	Libre/Commercial	Résumé automatique	Non	https://resoomer.com	France ;
Rocket Folio	**Automated content search and publishing for desktop, digital media, web, or intranet**	**Commercial**	**Recherche d'informations ;**	**Non**	**http://www.rocketsoftware.com/products/rocket-folionxt**	**États-Unis ;**
Rosetta	ROSETTA is a toolkit for analyzing tabular data within the framework of rough set theory. ROSETTA is designed to support the overall data mining and knowledge discovery process: From initial browsing and preprocessing of the data, via computation of minimal attribute sets and generation of if-then rules or descriptive patterns, to validation and analysis of the induced rules or patterns.	Libre	Découverte de connaissances ;	Non	http://bioinf.icm.uu.se/rosetta/	Suède ;
Rosette Text Analytics	**Multilingual Text Analytics Solution**	**Commercial**	**Classification de textes ; Clustering ; Analyse de sentiments ; Reconnaissance d'entités nommées ; Entity linking ; Extraction de relations ; Détection de la langue ; Traduction automatique ; Tokenisation ; PoS tagging ; Lemmatisation ;**	**Non**	**https://www.rosette.com/**	**États-Unis ;**

S4	S4 (Structured and Semantic Search Service) is a generic end-to-end infrastructure to rapidly build and deploy a search service providing state-of-the-art structured and semantic searches in collections of technical and scientific documents.	Libre	Recherche d'informations ;	Non	http://science-miner.com/structured-and-semantic-search-infrastructure/	France ;
SANSA	**SANSA is a big data processing engine for scalable processing of large-scale RDF data. SANSA uses Spark and Flink which offer fault-tolerant, highly available and scalable approaches to process massive sized datasets efficiently. SANSA provides the facilities for Semantic data representation, Querying, Inference, and Analytics**	**Libre**	**Classification ; Clustering ; Règlesd'association ; Détection d'anomalie ;**	**Non**	**http://sansa-stack.net/**	**Allemagne ;**
SAP predictive analytics	Predictive modeling suite	Commercial		Non	https://www.sap.com/products/analytics/predictive-analytics.html	Allemagne ;
SAS Enterprise Mining	**Create accurate predictive and descriptive models for large volumes of data.**	**Commercial**	**Prétraitement ; Apprentissage automatique ;**	**Non**	**https://www.sas.com/en_id/software/analytics/enterprise-miner.html**	**États-Unis ;**
SAS Text Miner	Text mining software from SAS automatically finds information buried in unstructured text data.	Commercial	Découverte de connaissances ; Reconnaissance d'entités nommées ; Clustering ; Extraction de relations ; Apprentissage automatique ;	Non	https://www.sas.com/en_us/software/text-miner.html	États-Unis ;
Scikit-learn	**Simple and efficient tools for data mining and data analysis. Accessible to everybody, and reusable in various contexts. Built on NumPy, SciPy, and matplotlib**	**Libre**	**Classification ; Clustering ; Analyse de régression ;**	**Non**	**http://scikit-learn.org/stable/index.html**	**France ;**
Scrapy	An open source and collaborative framework for extracting the data you need from websites. In a fast, simple, yet extensible way.	Libre	Indexation ;	Non	https://scrapy.org	?
SDL MultiTerm Extract	**Computer assisted translation**	**Commercial**	**Traduction automatique ;**	**Non**	**https://www.sdltrados.com/**	**Allemagne ;**
Semdee	Comprendre et exploiter de gros volumes de données textuelles	Commercial	Apprentissage automatique ;	Non	http://www.semdee.com/	France ;
SemRep	**SemRep is a UMLS-based program that extracts three-part propositions, called semantic predications, from sentences in biomedical text.**	**Libre**	**Extraction de relations**	**Non**	**https://semrep.nlm.nih.gov/**	**États-Unis ;**
SENNA	ENNA is a software distributed under a non-commercial license, which outputs a host of Natural Language Processing (NLP) predictions: part-of-speech (POS) tags, chunking (CHK), name entity recognition (NER), semantic role labeling (SRL) and syntactic parsing (PSG).	Libre	PoS tagging ; Chunking ; Reconnaissance d'entités nommées ; Etiquetage de rôle sémantique ; Parsing ;	Non	http://ronan.collobert.com/senna/	États-Unis ;

Sentic API	**Sentic API provides the semantics and sentics (i.e., the denotative and connotative information) associated with the concepts of SenticNet 4, a semantic network of commonsense knowledge that contains 50,000 nodes (words and multiword expressions) and thousands of connections (relationships between nodes).**	**Libre**	**Analyse de sentiments ;**	**Non**	**http://sentic.net/api/**	**États-Unis ;**
SentiStrength	SentiStrength estimates the strength of positive and negative sentiment in short texts, even for informal language	Libre	Analyse de sentiments ;	Non	http://sentistrength.wlv.ac.uk/	Royaume-Uni ;
Shogun	**The Shogun Machine learning toolbox offers a wide range of efficient and unified Machine Learning methods.**	**Libre**		**Non**	**http://shogun-toolbox.org/**	**?**
Simple Extractor	Software application oriented to extracting terminology from texts. Some of its main features are its simple use and its intuitive interfaces. This tool allows the setting of different extraction criteria and exporting files.	Commercial	Extraction terminologique ;	Non	http://www.dail-software.com/help/9_en/index.html	Espagne ;
Sisense	**Business intelligence tool for simplifying complex data preparation and analysis.**	**Commercial**	**Prétraitement ; Apprentissage automatique ;**	**Non**	**https://www.sisense.com/get/pricing/**	**États-Unis ; Israël ;**
Sketch Engine	Sketch Engine is the ultimate tool to explore how language works. Its algorithms analyze authentic texts of billions of words (text corpora) to identify instantly what is typical in language and what is rare, unusual or emerging usage. It is also designed for text analysis or text mining applications	Commercial	Collocation ; Concordancier ; Extraction terminologique ;	Non	https://www.sketchengine.eu	République Tchèque ;
SMART Text Miner	**The SMART Text Miner is a sophisticated software tool that can extract hidden knowledge from legacy texts.**	**?**	**Extraction terminologique ;**	**Non**	**http://www.smartny.com/miner.htm**	**États-Unis ;**
Smartlogic	Semantic platform that allows organizations to realize the business value of their information. By leveraging a common vocabulary and sophisticated semantic techniques Semaphore:	Commercial	Découverte de connaissances ; Reconnaissances d'entités nommées ; Extraction de relations ; Classification ; Recherche d'informations ; Désambiguisation ;	Non	https://www.smartlogic.com/	États-Unis ;
SoftLaw	**Analyse de documents juridiques**	**Commercial**	**Extraction d'informations ;**	**Non**	**https://www.softlaw.digital/**	**France ;**
SpaCy	Software library for NLP.	Libre	Apprentissage profond ; Tokenisation ; PoS tagging ; Segmentation de phrases ; Parsing ; Reconnaissance d'entités nommées ; classification de textes ;	Non	https://spacy.io/	Australie ;
Stanbol	**Set of reusable components for semantic content management.**	**Libre**	**Détection de la langue ; Tokenisation ; PoS tagging ; Chunking ; Lemmatisation ; Reconnaissance d'entités nommées ;**	**Non**	**https://stanbol.apache.org**	**États-Unis ;**

Stanford CoreNLP	Stanford CoreNLP provides a set of human language technology tools. It can give the base forms of words, their parts of speech, whether they are names of companies, people, etc., normalize dates, times, and numeric quantities, mark up the structure of sentences in terms of phrases and syntactic dependencies, indicate which noun phrases refer to the same entities, indicate sentiment, extract particular or open-class relations between entity mentions, get the quotes people said, etc.	Libre	POS tagging ; Reconnaissance d'entités nommées ; Résolution de coréférence ; parsing ; analyse en dépendances ; Analyse de sentiments ; Extraction d'informations	Non	https://stanfordnlp.github.io/CoreNLP/	États-Unis ;
Stanford NLP	**An integrated suite of natural language processing tools for English and (mainland) Chinese, including tokenization, part-of-speech tagging, named entity recognition, parsing, and coreference.**	**Libre**	**PoS Tagging ; Résolution de coréférence ; Tokenisation ; Reconnaissance d'entités nommées; Parsing ; Extraction de relations ; Classification ;**	**Oui**	**https://nlp.stanford.edu/**	**États-Unis ;**
Stanford Topic Modeling Toolbox	The Stanford Topic Modeling Toolbox (TMT) brings topic modeling tools to social scientists and others who wish to perform analysis on datasets that have a substantial textual component	Libre	Clustering ; Topic modeling ;	Non	https://nlp.stanford.edu/software/tmt/tmt-0.4/	États-Unis ;
Statistica Text Miner	**Statistica Text Miner is an optional extension of Statistica Data Miner, ideal for translating unstructured text data into meaningful, valuable clusters of decision-making "gold."**	**Commercial**	**Lemmatisation ; Clustering ;**	**Non**	**https://www.statsoft.fr/logiciels/textminer.php**	**États-Unis ;**
Stratifyd	Stratifyd's data analytics platform allows users to integrate, analyze, and visualize data in a single platform, empowering analysts through a holistic view of both structured and unstructured data.	Commercial	Clustering ; Analyse de sentiments ; Extraction d'nformations ;	Non	https://www.stratifyd.com/	États-Unis ;
streamDM	**streamDM is a new open source software for mining big data streams using Spark Streaming**	**Libre**	**Classification ; Clustering ; Analyse de régression ;**	**Non**	**http://huawei-noah.github.io/streamDM/**	**Chine ;**
Synaptica	Software solution for knowledge organization and discovery.	Commercial	Annotation ; Classification ;	Non	http://www.synaptica.com/	États-Unis ;
Sysomos	**Sysomos is a unified, insights-driven social platform that gives marketers the easiest way to Search, Discover, Listen, Publish, Engage, and Analyze at scale across earned, owned, and paid media.**	**Commercial**	**Recherche d'informations ; Découverte de connaissances ;**	**Non**	**https://sysomos.com/**	**Canada ;**

Systran	Understanding, analyze and act in over 50 languages	Commercial	Analyse de sentiments ; Traduction automatique ; Reconnaissance d'entités nommées ; Détection de la langue ; Segmentation ; Tokenisation ; PoS tagging ; Analyse morphologique ;	Non	http://www.systran.io/	Corée du Sud ;
TACIT	**Text Analysis,Crawling and Interpretation Tool**	**Libre**	**Classification de textes ; Clustering ;**	**Non**	**http://tacit.usc.edu/**	**États-Unis ;**
Tagtog	Biomedical annotation tool	Libre	Annotation ;	Non	https://www.tagtog.net/	Pologne ;
Talismane	**NLP framework: sentence detector, tokeniser, pos-tagger and dependency parser**	**Libre**	**Tokenisation ; POS tagging ; Analyse en dépendances ; Détection de phrases ;**	**Non**	**https://github.com/joliciel-informatique**	**France ;**
TAMS	TAMS stands for Text Analysis Markup System. It is a convention for identifying themes in texts (web pages, interviews, field notes). It was designed for use in ethnographic and discourse research.	Libre	identification de thèmes	Non	http://tamsys.sourceforge.net	États-Unis ;
TANAGRA	**TANAGRA is a free DATA MINING software for academic and research purposes. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area.**	**Libre**	**Apprentissage automatique ;**	**Non**	**https://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html**	**France ;**
TapoRware	TAPoRware is a set of text analysis tools that enables users to perform text analysis on HTML, XML and plain text files, using documents from the users' machine or on the web.	Libre	Concordancier ; Tokenisation ; Extraction de liens ; Résumé automatique ;	Non	http://taporware.ualberta.ca/~taporware/about.shtml	Canada ;
TBXTools	**TBXTools allows easy and rapid Terminology Extraction and Management. This tool implements both statistical and linguistic methods, along with several utilities to create and manage terminological databases. It is written in Python and uses NLTK (Natural Language Toolkit)**	**Libre**	**Extraction terminologique ;**	**Non**	**https://sourceforge.net/projects/tbxtools/**	**Espagne ;**
TensorFlow	Machine learning library	Libre	Apprentissage automatique ;	Non	https://www.tensorflow.org/	États-Unis ;
Termsuite	**Outil d'extraction terminologique et d'alignment multilingue de termes.**	**Libre**	**Extraction terminologique ;**	**Oui**	**https://termsuite.github.io/fr/**	**France ;**
Text2data	Advanced text analytics	Commercial	Analyse de sentiments ; Résumé automatique ; Classification de textes ; Reconnaissance d'entités nommées ; Clustering ; Extraction de mots clés ;	Non	http://text2data.org/	?
Textable	**Free open source software to analyze and process texts visually**	**Libre**	**Concordancier ; Collocation ; Lemmatisation ; POS tagging ; Clustering ; Classification ; Visualisation ; Pétraitement ;**	**Non**	**http://textable.io/**	**Suisse ;**

Textalytics	Textalytics is a meaning extraction service that produces meaningful data from social media content, contracts, news, and other documents	Commercial	Annotation ; PoS tagging ; Parsing ; Lemmatisation ; Analyse de sentiments ; Clustering ;	Oui	https://textalytics.io/	?
Textblob	**TextBlob is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.**	**Libre**	**Extraction terminologique ; PoS tagging ; Analyse de sentiments ; Classification ; Traduction automatique ; Détection de la langue ; Tokenisation ; Parsing ; Correction orthographique ;**	**Non**	**https://textblob.readthedocs.io/en/dev/**	**États-Unis ;**
TextCat	http://odur.let.rug.nl/~vannoord/TextCat/	Libre	Classification de textes ;	Oui	http://odur.let.rug.nl/~vannoord/TextCat/	Pays-Bas ;
TextObserver	**Outil de d'observation et d'exploitation des données textuelles multidimensionnelles.**	**Libre**	**Textométrie ;**	**Non**	**http://textopol.u-pec.fr/textobserver/**	**France ;**
TextRazor	The TextRazor API helps you extract and understand the Who, What, Why and How from your news stories with unprecedented accuracy and speed	Libre	Reconnaissance d'entités nommées ; Classification ; Annotation ; Désambiguisation ; Extraction de relations ; Extraction de mots clés ;	Oui	https://www.textrazor.com/	Royaume-Uni ;
Theano	**Deep learning**	**Libre**	**Apprentissage profond ;**	**Non**	**http://www.deeplearning.net/software/theano/**	**Canada ;**
Thematic	AI text analytics and visualizations	Commercial	Clustering ;	Non	https://getthematic.com/	?
Theysay	**Emotional AI and advanced data analytics to stream, interpret, and bring together opinions, moods, and feelings across the Web.**	**Commercial**	**Analyse de sentiments ; Analyse d'opinions ;**	**Non**	**http://www.theysay.io/**	**Royaume-Uni ;**
Think Analytics	The ThinkAnalytics Search and Recommendations Engine provides a powerful, scalable, real-time and comprehensive multi-content/multi-platform Recommendations Engine supporting across content delivery of recommendations and search to multiple platforms such as the set top box, mobile, web, smart TV, games consoles, and others.	Commercial	Apprentissage automatique ;	Non	https://thinkanalytics.com/	Royaume-Uni ;
TnT	**Statistical Part-of-Speech Tagging**	**Libre**	**POS tagging**	**Non**	**http://www.coli.uni-saarland.de/~thorsten/tnt/**	**Allemagne ;**
Torch	Torch is a scientific computing framework with wide support for machine learning algorithms that puts GPUs first. It is easy to use and efficient, thanks to an easy and fast scripting language, LuaJIT, and an underlying C/CUDA implementation.	Libre	Apprentissage automatique ;	Non	http://torch.ch/	?

TreeTagger	**The TreeTagger is a tool for annotating text with part-of-speech and lemma information**	**Libre**	**PoS tagging ; Lemmatization ; Chunking ;**	**Oui**	**http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/**	**Allemagne ;**
Tropes	Analyse sémantique de textes	Libre	Textométrie ;	Non	http://www.tropes.fr/	France ;
Tweet NLP	**We provide a tokenizer, a part-of-speech tagger, hierarchical word clusters, and a dependency parser for tweets, along with annotated corpora and web-based annotation tools.**	**Libre**	**Tokenisation ; PoS tagging ; Parsing ; Clustering ; Annotation ;**	**Non**	**http://www.cs.cmu.edu/~ark/TweetNLP/**	**États-Unis ;**
TwinWord	Text analysis APIS to understand and associate words	Commercial	Analyse de sentiments ; Clustering ; Classfication de textes ; Lemmatisation ; Extraction de mots clés ;	Non	https://www.twinword.com/	États-Unis ;
TXM	**Analyses textométriques**	**Libre**	**Textométrie**	**Non**	**http://textometrie.ens-lyon.fr/**	**France ;**
U-compare	U-Compare is an integrated text mining/natural language processing system based on the UIMA Framework	Libre	Annotation ; Reconnaissance d'entités nommées ; Tokenisation ; PoS Tagging ; Lemmatisation ; Parsing ; Extraction d'évènements ;	Non	http://u-compare.org/	Japon ;
UAIC NLP	**The Natural Language Processing (NLP) Group at UAIC-FII has been involved in many national and European projects dealing with: morphology, information retrieval, dialogue systems, anaphora resolution, WordNet, discourse parsing and summarization, question-answering, textual entailment, etc.**	**Libre**	**PoS tagging ; Chunking ; Reconnaissance d'entités nommées ; Parsing; Sentence splitting ; Résolution d'anaphore ;**	**Oui**	**http://nlptools.info.uaic.ro/Resources.jsp**	**Roumanie**
UDPipe	Trainable pipeline for tokenizing, tagging, lemmatizing and parsing Universal Treebanks and other CoNLL-U files	Libre	Lemmatisation ; POS tagging ; parsing ; Analyse en dépendances ; Annotation ; Tokenisation ;	Non	https://ufal.mff.cuni.cz/udpipe	République Tchèque ;
UltiPro Perception	**Understand what employees are saying and how they truly feel about the workplace, with surveys and sentiment analysis.**	**Commercial**	**Apprentissage automatique ; Analyse d'opinions ;**	**Non**	**https://www.ultimatesoftware.com/UltiPro-Solution-Features-Employee-Surveys**	**États-Unis ;**
Unitex	Open Source Corpus Processing Suite.	Libre	Reconnaissance d'entités nommées ; Désambiguïsation ;	Non	http://unitexgramlab.org/	France ;
Vertica	**Data mining and analysis**	**Commercial**	**Apprentissage automatique ;**	**Non**	**https://www.vertica.com/**	**États-Unis ;**
VisualText	Integrated development environment for building information extraction systems, natural language processing systems, and text analyzers.	Libre/Commercial	Extraction d'informations ; Résumé automatique ; Clustering ; Reconnaissance d'entités nommées ; Recherche d'informations ; Annotation ;	Non	http://www.textanalysis.com/Products/products.html	États-Unis ;
VizTrails	**Open-source scientific workflow and provenance management system that supports data exploration and visualization.**	**Libre**	**Analyse de données ; Visualisation ;**	**Non**	**https://www.vistrails.org/index.php/Main_Page**	**États-Unis ;**

VosViewer	VOSviewer is a software tool for constructing and visualizing bibliometric networks. These networks may for instance include journals, researchers, or individual publications, and they can be constructed based on citation, bibliographic coupling, co-citation, or co-authorship relations. VOSviewer also offers text mining functionality that can be used to construct and visualize co-occurrence networks of important terms extracted from a body of scientific literature.	Libre	Extraction terminologique ; Co-occurrence ;	Non	http://www.vosviewer.com/	Pays-Bas ;
Vowpal WAbbit	**The Vowpal Wabbit (VW) project is a fast out-of-core learning system**	**Libre**	**Classification ; Analyse de régression ;**	**Non**	**https://github.com/JohnLangford/vowpal_wabbit/wiki**	**États-Unis ;**
Voyant	Environnement en ligne de lecture et d'analyse de textes numériques.	Libre	Clustering ; Concordancier ;	Non	http://voyant-tools.org/	Canada ;
Voziq	**Unify every customer experience data source, and apply combined power of text analytics and predictive algorithms for strategic customer intelligence.**	**Commercial**	**Analyse de sentiments ; Apprentissage automatique ;**	**Non**	**http://voziq.com/**	**États-Unis ;**
WarpLDA	Cache efficient implementation for Latent Dirichlet Allocation	Libre	Topic modeling ;	Non	https://github.com/thu-ml/warplda	Chine ;
WebAnno	**WebAnno is a general purpose web-based annotation tool for a wide range of linguistic annotations including various layers of morphological, syntactical, and semantic annotations.Additionaly, custom annotation layers can be defined, allowing WebAnno to be used also for non-linguistic annotation tasks.**	**Libre**	**Annotation sémantique ;**	**Non**	**https://webanno.github.io/webanno/**	**Allemagne ;**
Weblicht	WebLicht is an execution environment for automatic annotation of text corpora. Linguistic tools such as tokenizers, part of speech taggers, and parsers are encapsulated as web services, which can be combined by the user into custom processing chains. The resulting annotations can then be visualized in an appropriate way, such as in a table or tree format.	Libre ?	Tokenisation ; PoS tagging ; Parsing ;	Oui	https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page	Allemagne ;
Weka	**Collection of machine learning algorithms for data mining tasks.**	**Libre**	**Classification ; Clustering ; Analyse de régression ; Arbre de décision ; Règle d'association ;**	**Oui**	**http://www.cs.waikato.ac.nz/ml/weka/**	**Nouvelle-Zélande ;**

Word2Vec	Word embeddings	Libre	Apprentissage profond ; Plongement de mots	Non	https://github.com/dav/word2vec	États-Unis ;
WordFreak	**WordFreak is a java-based linguistic annotation tool designed to support human, and automatic annotation of linguistic data as well as employ active-learning for human correction of automatically annotated data.**	**Libre**	**Annotation ;**	**Non**	**http://wordfreak.sourceforge.net/**	**États-Unis ;**
YaTeA	Term extraction	Libre	Extraction terminologique ;	Oui	http://search.cpan.org/~thhamon/Lingua-YaTeA-0.5/	France ;

Annexe 4 : Recensement de laboratoires spécialisés en fouille de textes et TAL

Organisme/laboratoire	Université	Ville	Rattachement	Site web
ATILF	Université de Lorraine	Nancy	UMR CNRS	http://www.atilf.fr/
CLESTHIA	Université Sorbonne Nouvelle	Paris		http://www.univ-paris3.fr/clesthia-langage-systemes-discours-ea-7345-98241.kjsp
CLLE	Université Toulouse Jean Jaurès	Toulouse	UMR CNRS	https://clle.univ-tlse2.fr/
DGLFLF	Mission langues et numérique	Paris	Ministère de la culture	http://www.culture.gouv.fr/Thematiques/Langue-francaise-et-langues-de-France
ERIC-DMD	Université Lyon 1 et 2	Lyon		https://eric.ish-lyon.cnrs.fr/
ERTIM - INALCO	Université Sorbonne Paris Cité	Paris	CNRS IRD	http://www.er-tim.fr/
GERIICO	Université Lille 3	Lille		https://geriico-recherche.univ-lille3.fr/
GREYC-Codag, Hultech	Université de Caen	Caen	UMR CNRS	https://www.greyc.fr/
IMSIC (anciennement I3M)	Université Sophia Antipolis Université de Toulon Université Côte d'Azur	Nice Toulon		http://www.imsic.fr/
INRA-MaIAGE		Jouy-en -Josas	UPR INRA	
INRIA-Alpage	Université Paris 7	Paris	UMR INRIA	https://www.rocq.inria.fr/alpage-wiki/tiki-index.php?page=accueil
INRIA-Orpailleur	Université de Lorraine	Nancy		https://www.inria.fr/equipes/orpailleur
INRIA-Zénith	Université Montpellier 2	Montpellier		https://www.inria.fr/equipes/zenith
Institut des systèmes complexes		Paris	UPS CNRS	https://iscpif.fr/
IRISA-Texmex	Université de Rennes 1 Université Bretagne Sud	Rennes	UMR CNRS INRIA	https://www.irisa.fr/
IRIT-Melodi	Université de Toulouse	Toulouse	UMR CNRS	https://www.irit.fr/-Equipe-MELODI-
L3i	Université de la Rochelle	La Rochelle		https://l3i.univ-larochelle.fr/
Laboratoire de linguistique professionnelle	Université Paris 7 Diderot	Paris	UMR CNRS	http://www.llf.cnrs.fr/
Laboratoire Hubert Curien- Data Intelligence	Université de Saint-Etienne	Saint-Etienne	UMR CNRS	https://laboratoirehubertcurien.univ-st-etienne.fr/en/index.html
LaBRI-MABioVis	Université de Bordeaux	Bordeaux	UMR CNRS	https://mabiovis.labri.fr/
Lattice	Université Sorbonne Nouvelle Université Paris Cité Université Paris Sciences et Lettres	Montrouge	Ecole Normale Supérieure UMR CNRS	http://www.lattice.cnrs.fr/

Organisme/laboratoire	Université	Ville	Rattachement	Site web
LIA	Université d'Avignon et des Pays de Vaucluse	Avignon		https://lia.univ-avignon.fr/
LIDILEM	Université Grenoble 3	Grenoble		https://lidilem.univ-grenoble-alpes.fr/
LIFAT	Université de Tours	Tours	ERL CNRS	https://www.univ-tours.fr/sciences-et-technologies/laboratoire-d-informatique-fondamentale-et-appliquee-de-tours-lifat--116922.kjsp
LIFO	Université d'Orléans	Orléans	INSA Val de Loire	https://www.univ-orleans.fr/lifo
LIG **Ecole doctorale "Langages, espaces, temps, sociétés"**	Université Grenoble Alpes	Grenoble	UMR CNRS INRIA Grenoble Rhône Alpes Grenoble INP	https://www.liglab.fr/ https://lig-getalp.imag.fr/
LIGM	Université Paris-Est Marne-la-Vallée	Marne-la-Vallée	UMR CNRS	http://ligm.u-pem.fr/
LiLPa	Université de Strasbourg	Strasbourg		https://lilpa.unistra.fr/
LIMICS	Sorbonne Université Université Paris 13	Paris	UMRS INSERM	http://www.limics.fr/
LIMSI	Université Paris Sud Université Paris Saclay	Orsay	INS2I CNRS	https://www.limsi.fr/fr/
LIP6	Sorbonne Université	Paris	UMR CNRS	https://www.lip6.fr/
LIPN	Université Paris 13	Paris	UMR CNRS	https://lipn.univ-paris13.fr/
LIRIS	Université Claude Bernard Lyon 1 Université Lumière Lyon 2	Lyon	UMR CNRS INSA lyon Ecole centrale Lyon	https://liris.cnrs.fr/
LIRMM	Université de Montpellier	Montpellier	UMR CNRS	http://www.lirmm.fr/
LIS	Université Aix Marseille	Marseille	UMR CNRS	https://www.lis-lab.fr/
LIST CEA Tech	Université Paris Saclay	Gif-sur-Yvette	CEA	http://www-list.cea.fr/le-list-institut-de-cea-tech
LISTIC	Université Savoie Mont-Blanc	Annecy	Polytech Annecy Chambér	https://www.listic.univ-smb.fr/presentation/organisation/
LORIA	Université de Lorraine	Nancy	UMR CNRS INRIA	http://www.loria.fr/fr/
LS2N	Université de Nantes	Nantes		https://www.ls2n.fr/
LTCI		Paris	Télecom Paris	https://www.telecom-paris.fr/fr/recherche/laboratoires/laboratoire-traitement-et-communication-de-linformation-ltci
MoDyCo	Université Paris Nanterre Université Paris Lumière Campus Condorcet	Paris	UMR CNRS	https://www.modyco.fr/fr/
STL	Université Lille 3	Lille	UMR CNRS	https://stl.univ-lille.fr/
TETIS		Montpellier	UMR CNRS IRSTEA CIRAD AgroParisTech	https://www.umr-tetis.fr/index.php/fr/