TXM : présentation et commandes de base

Karën Fort

karen.fort@sorbonne-universite.fr / https://members.loria.fr/KFort/

LETTRES SORBONNE UNIVERSITÉ

Quelques sources d'inspiration

par ordre d'importance décroissant

- Atelier TXM du 25 et 26 septembre 2014
- Manuel de TXM : http://txm.sourceforge.net/doc/manual/manual1.xhtml
- Vidéo : http://txm.sourceforge.net/enregistrement_atelier_initiation_TXM_fr.html
- B. Pincemin (IHRIM) et S. Heiden (IHRIM)
- Site de TXM : http://textometrie.ens-lyon.fr/
- Modèle de TreeTagger : http://www.cis.uni-muenchen.de/%7Eschmid/tools/TreeTagger/

Une communauté active

- formations
- liste de diffusion active
- communauté d'utilisateurs et de développeurs
  - logiciel libre
  - supportant Unicode
  - multi-plateformes (Java)
  - modulaire (R et CQP)
- documentation sous différentes formes (pdf, vidéo, pages Web)

+ version portail Web : `http://portal.textometrie.org/demo/?locale=fr`

 La documentation et la vitalité de la communauté sont des critères fondamentaux dans le choix d'un logiciel

Quelle communauté ?

Sciences humaines et sociales :

- archives historiques
- dépouillement d'enquêtes avec questions ouvertes
- œuvres littéraires
- corpus scientifiques
- etc.

Textométrie ?

Spécificité française

Historiquement, évolution et élargissement avec les avancées techniques (annotations, structuration) :

- lexicométrie : statistiques lexicale (sur les mots)
- logométrie : statistiques sur les discours
- textométrie : statistiques sur les textes

$\rightarrow$ les calculs sont délégués à l'ordinateur, mais le chercheur reste maître de l'interprétation

Particularités de TXM

- interface très complète
- robustesse : permet de traiter jusqu'à 10 millions de mots
- puissance : permet d'intégrer toutes sortes de traitement *via* le logiciel R (de statistiques)
- rapidité : permet d'interroger des millions de mots très efficacement *via* CQP (Corpus Query Processor)

TXM et TAL

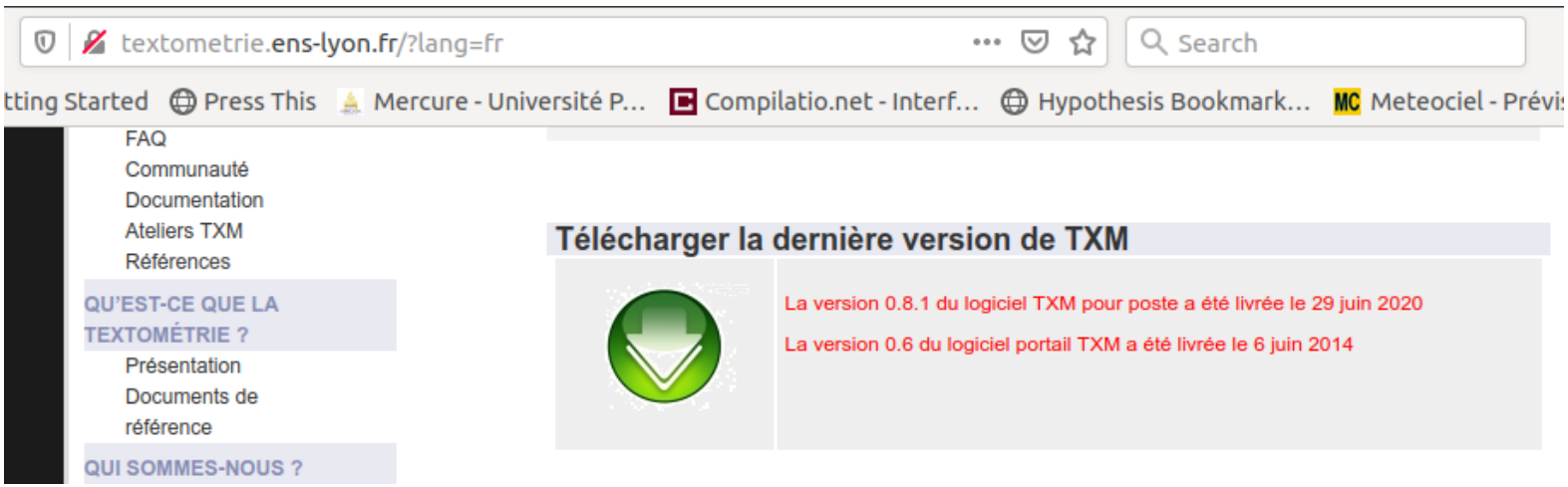
TXM n'est pas un outil de TAL [1] en tant que tel, mais

- il intègre des fonctionnalités de TAL, *via* TreeTagger (lesquelles ?)
- il permet d'explorer les corpus et de les analyser manuellement (préalable au TAL)

$\Rightarrow$ outil d'analyse très pratique (indispensable ?)

1. Traitement Automatique des Langues

Installer TXM



Installer TXM : prérequis

Prérequis d'installation

- Vous aurez besoin des droits d'installation sur votre machine pour pouvoir installer TXM ;
- Vous aurez besoin d'un accès à Internet ;
- TXM fonctionne sur des machines **64-bit** ;
- Windows
  - TXM est supporté pour Windows 7 et Windows 10
- Mac OS X
  - TXM est supporté pour Mac OS X 10.14
    - Réglages de sécurité : il faut modifier les paramètres de « sécurité » du système pour pouvoir installer correctement TXM. Aller dans « Préférences Système > Personnel > Sécurité » et autoriser les applications téléchargées depuis Internet. Si on ne fait pas ce réglage, le problème ne se manifeste pas au moment de l'installation, mais au premier lancement.

Installer TXM : TreeTagger

A. Installation de TreeTagger dans TXM 0.8.0

À partir de TXM 0.8.0, deux extensions dédiées à TreeTagger installent automatiquement le logiciel TreeTagger et les modèles français et anglais :

1. Lancer la commande “Fichier > Ajouter une extension”
2. Sélectionner les lignes “TreeTagger software” et “TreeTagger models” pour installer le logiciel TreeTagger et les modèles français et anglais
3. Valider les étapes suivantes
4. Après le redémarrage de TXM, TreeTagger est opérationnel
5. Pour installer d’autres modèles de langues, suivre l’étape 4 de l’installation manuelle ci-dessous
6. Fin

Corpus vœux et interface

Préalable

Télécharger le corpus Vœux :
http://sourceforge.net/projects/txm/files/corpora/voeux/voeux-bin.txm/download

Différents espaces, à explorer :

- onglets
- menuS (3 modes d'accès)
- console

Commandes de bas niveau

- charger (un corpus déjà importé auparavant)
- édition
- description

Corpus Vœux

Chargez le corpus Vœux
Que pouvez-vous dire sur le corpus Vœux grâce à `TXM` ?

Que manque-t-il ?

Commandes de bas niveau

- charger (un corpus déjà importé auparavant)
- édition
- description

Corpus Vœux

Chargez le corpus Vœux
Que pouvez-vous dire sur le corpus Vœux grâce à `TXM` ?

Que manque-t-il ?

→ le nombre entre parenthèses après `id` sous `text` donne le nombre de textes
→ mais il manque la licence et un descriptif !

Charger *vs* importer

`Charger` : corpus **déjà** importé dans `TXM` auparavant

`Importer` : corpus brut (txt, XML, voire en provenance du presse-papier)

Import *via* le presse-papier

- aller sur le site Web
- copier le contenu de la page (CTRL+C)
- dans `TXM`, sélectionner Fichier/Importer/Presse-papier
- tada !

Réglages

- vue interne
- ajout d'informations
- changement d'affichage

Qu'est-ce que le lexique pour TXM ?

- liste de formes (par défaut, mais paramétrable)
- fréquences d'apparition
- lemmatisation et étiquetage (par défaut) avec TreeTagger [Schmid, 1997], mais possibilité d'importer des corpus pré-annotés

- lien vers la concordance

 Le contexte est fondamental dans TXM (seule la remise en contexte permet l'analyse)

Qualité de l'étiquetage morpho-syntaxique

ou *POS tagging*

Exactitude (*accuracy* en anglais, à ne pas confondre avec la précision) :

- `TreeTagger` (1994) : 95,7 %
  [Allauzen and Bonneau-Maynard, 2008]
- `MElt` (2010) : près de 98 % [Denis and Sagot, 2010]

Quelle différence concrète ?

Qualité de l'étiquetage morpho-syntaxique

ou *POS tagging*

Exactitude (*accuracy* en anglais, à ne pas confondre avec la précision) :

- `TreeTagger` (1994) : 95,7 %
  [Allauzen and Bonneau-Maynard, 2008]
- `MElt` (2010) : près de 98 % [Denis and Sagot, 2010]

Quelle différence concrète ?

96 % d'exactitude, environ 10 mots par phrase
→ sur 10 phrases, un mot mal étiqueté dans 4 phrases

98 % d'exactitude → deux fois moins d'erreurs

Caractéristiques ?

Exporter et analyser

Exportez le lexique du corpus Vœux dans un tableur.
Que pouvez-vous constater concernant la répartition des fréquences de mots ?

Premiers pas en CQL

Corpus Query Language

- expressions régulières : *Europe*|*européen*.∗, [] (un mot), & et | (booléens)
- neutralisations (à ajouter **après** l'expression) :
  - %c pour neutraliser la casse ("europe"%c)
  - %d pour neutraliser les diacritiques (accents, cédille)
  - etc. (voir doc)
- assistant de requête
- tri du contexte droit **et** du contexte gauche

Trier, visualiser et chercher sont 3 actions différentes

Fréquences

Index permet de chercher la fréquence d'une expression

Rechercher

Trouver en une seule recherche les fréquences de « patrie », « patriote », « patriotisme », « compatriotes »

→ permet de tester une formule de recherche (avant de se lancer en concordance)

Les vœux dans le corpus Vœux

Rechercher les vœux

Trouver en une seule recherche le souhait de « bonne année » de chaque Président

Les vœux dans le corpus Vœux

Rechercher les vœux

Trouver en une seule recherche le souhait de « bonne année » de chaque Président

[frlemma="je"][] * [frlemma="souhaiter"][] * [frlemma="année"] within s
s = dans l'espace de la phrase

[frlemma="je"][] * [frlemma="souhaiter"][] * [frlemma="année"] within 25
= dans l'espace de 25 mots

Cooccurrences

Moyen de voir comment un mot ≪ résonne ≫ dans un corpus

À lire

La textométrie par les textomaîtres :
http://textometrie.ens-lyon.fr/spip.php?rubrique80

- lexicométrie, logométrie, textométrie
- manipulations de base :
  - lexique
  - concordance
  - cooccurrences
  - index
  - CQL

Mais aussi :

- loi de Zipf
- qualité des taggers

Illustrer la loi de Zipf

Exercice noté, à rendre avant le prochain cours, par mail

Construire, à l'aide de `TXM` et sur un tableur par exemple :

- deux (beaux) graphiques illustrant la loi de Zipf
- sur des textes de votre choix
- dans deux langues différentes
  - un bonus sera octroyé à qui traitera des langues peu courantes
  - si vous avez dû effectuer des traitements particuliers, signalez-les
  - commentez/analysez les résultats obtenus


Allauzen, A. and Bonneau-Maynard, H. (2008).
Training and evaluation of pos taggers on the french multitag corpus.
In Nicoletta Calzolari (Conference Chair), Khalid Choukri, B. M. J. M. J. O. S. P. D. T., editor, Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco. European Language Resources Association (ELRA).
http ://www.lrec-conf.org/proceedings/lrec2008/.


Denis, P. and Sagot, B. (2010).
Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morphosyntaxique état-de-l'art du français.
In
Traitement Automatique des Langues Naturelles : TALN 2010, Montréal, Canada.


Schmid, H. (1997).

New Methods in Language Processing, Studies in Computational Linguistics, chapter Probabilistic part-of-speech tagging using decision trees, pages 154–164.
UCL Press, London.