

Introduction au Text Mining

Principes et applications

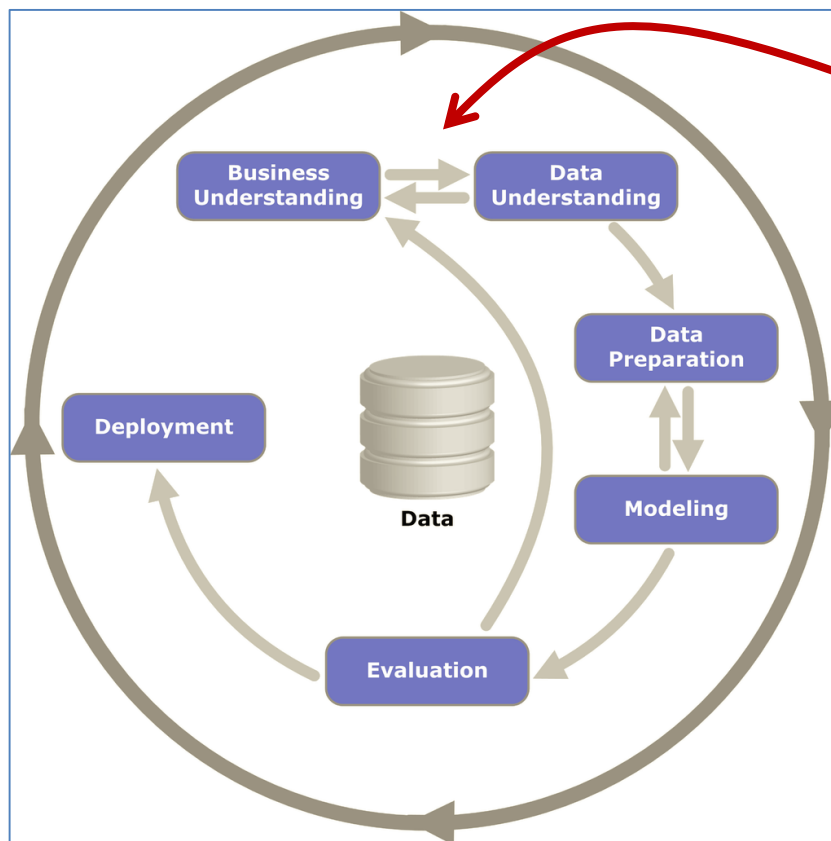
Ricco Rakotomalala

1. Du data mining au text mining
2. Principales applications
3. La représentation « bag-of-words » (BOW)
4. Au-delà de BOW
5. Bibliographique

Etendre la démarche data mining au traitement des données textuelles

DU DATA MINING AU TEXT MINING

Le data mining est un processus d'extraction de structures (connaissances) inconnues, valides et potentiellement exploitables dans les bases (entrepôts) de données (Fayyad, 1996), à travers la mise en œuvre des techniques statistiques et de machine learning.



La statistique exploratoire, l'économétrie, n'ont pas attendu le data mining pour s'intéresser à la modélisation sur des données issues d'enquêtes, de recueils divers,..., au sein des organisations ou en dehors (données de notre environnement socio-économique).

Les données textuelles constituent également une source d'information qui permettrait d'extraire de la connaissance (détecter des régularités [*patterns*], recherche des similarités, identifier les relations de causalité, etc.).

```
<document>  
< sujet>acq</ sujet>  
< texte>
```

Resdel Industries Inc said
it has agreed to acquire San/Bar Corp in a share-for-share
exchange, after San/Bar distributes all shgares of its
Break-Free Corp subsidiary to San/Bar shareholders on a
share-for-share basis.

The company said also before the merger, San/Bar would
Barry K. Hallamore and Lloyd G. Hallamore, San/Bar's director
of corporate development, 1,312,500 dlrs and 1,087,500 dlrs
respectviely under agreements entered into in October 1983.

```
</ texte>  
</ document>  
< document>  
< sujet>acq</ sujet>
```

```
< texte>
```

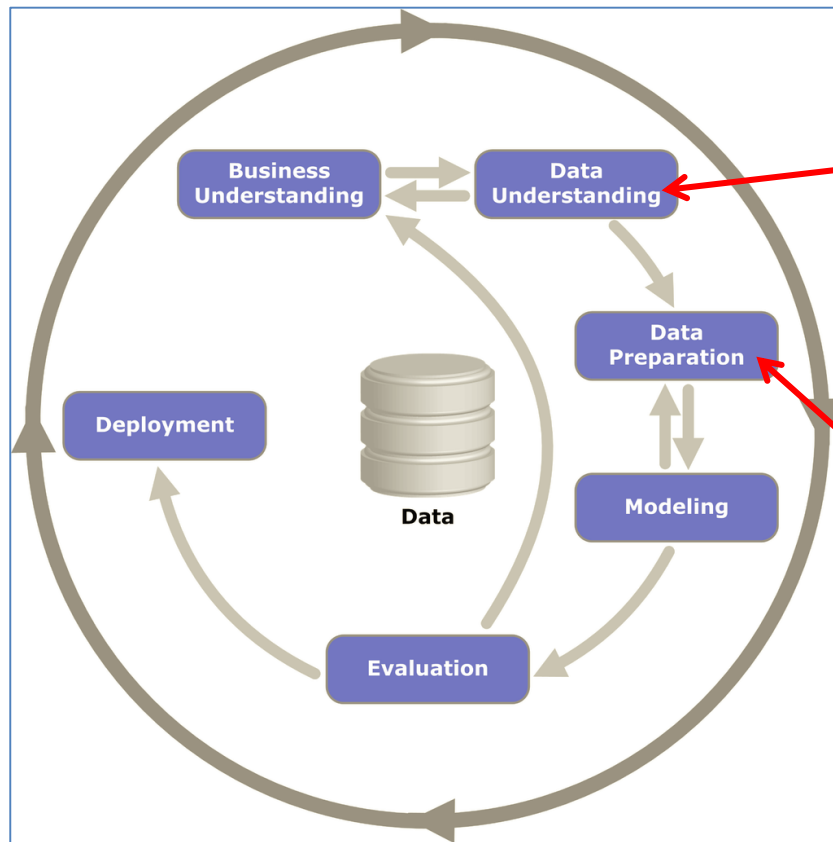
Warburg, Pincus Capital Co L.P., an
investment partnership, said it told representatives of Symbion
Inc it would not increase the 3.50-dlr-per-share cash price it
has offered for the company.
Last Month Warburg Pincus launched a tender offer to buy up
to 2.5 mln Symbion common shares.

```
</ texte>  
</ document>
```

Exemple : un ensemble de nouvelles (Reuters)

On dispose d'une collection de
« **documents** », composés de
« textes » associés à des
« sujets ». Est-il possible – à l'aide
de techniques de data mining –
d'établir une liaison fonctionnelle
entre « textes » et « sujets » :
indexation automatique,
catégorisation de textes.

Schématiquement : Le text mining est un processus d'extraction de structures (connaissances) inconnues, valides et potentiellement exploitables dans les documents textuels, à travers la mise en œuvre de techniques statistiques ou de machine learning. Mais d'autres applications spécifique aux textes sont possibles : résumé automatique, extraction d'information, etc.



- Document = individu statistique
- Collection de documents = Corpus = Base d'apprentissage

La préparation des données va jouer un rôle fondamental parce que les techniques statistiques usuelles ne sont pas armées pour traiter de la donnée non structurée.

Document structuré vs. Document non structuré en data mining

Les données usuellement exploités en data mining se présentent sous forme de tableaux attributs-valeurs. Ligne = individu statistique, colonne = attribut (descripteur). Les algorithmes sont taillées pour les traiter.

sep_length	sep_width	pet_length	pet_width	type
5.1	3.8	1.6	0.2	setosa
5.8	2.7	5.1	1.9	virginica
4.9	2.4	3.3	1.0	versicolor
7.7	2.6	6.9	2.3	virginica
6.6	3.0	4.4	1.4	versicolor
6.2	2.8	4.8	1.8	virginica
5.0	2.3	3.3	1.0	versicolor
4.4	3.0	1.3	0.2	setosa
5.7	2.8	4.5	1.3	versicolor

Les données en text mining se présentent sous forme de textes bruts. Les algorithmes de data mining ne savent pas les appréhender nativement.

<texte>

Resdel Industries Inc said it has agreed to acquire San/Bar Corp in a share-for-share exchange, after San/Bar distributes all shgares of its Break-Free Corp subsidiary to San/Bar shareholders on a share-for-share basis.

The company said also before the merger, San/Bar would Barry K. Hallamore and Lloyd G. Hallamore, San/Bar's director of corporate development, 1,312,500 dlrs and 1,087,500 dlrs respectviely under agreements entered into in October 1983.

</texte>



Une piste de travail possible consiste à transformer la collection de textes en tableau de valeurs, en espérant ne pas trop perdre d'informations. En ligne, nous aurions les documents, mais que mettre en colonne (*features, termes*) ? Et quelles valeurs mettre à l'intérieur du tableau (*weight, pondération*) ?

Cadre pour la catégorisation de textes multilingues

Radwan JALAM, Jérémy CLECH, Ricco RAKOTOMALALA
{jalam, jclech, rakotomaj}@eric.univ-lyon2.fr

Laboratoire ERIC
Université Lumière Lyon 2
5, av. Pierre Mendès-France
69676 Bron, France
fax: (33) 4 78 77 23 75

Abstract

In this paper, we propose an original framework for multilingual text categorization. The objective is to classify a set of texts, written in some language, using a predictive model learned from a set of texts written in a given language, called learning language. Contrary to the unilingual classical phase of text categorization, the classification phase contains two new steps : firstly identify the language of the text, and then automatically translate it into the learning language. As shown in this paper, first applications of multilingual text categorization on real data, that is over English, French and German newspapers, indicate that the approach is viable.

Résumé

Dans cet article, nous proposons un cadre pour la catégorisation de textes multilingues. L'objectif est de pouvoir, à partir d'un modèle de prédiction construit par apprentissage sur un corpus de textes rédigés dans une langue donnée, inférer sur une série de textes qui sont rédigés dans une langue quelconque. Cette phase d'inférence, par rapport à la généralisation classique, comprend deux étapes supplémentaires : la détection de la langue du texte, puis sa traduction automatique vers la langue de référence. Nos premiers résultats sur une application réelle : la catégorisation d'articles de journaux allemands, anglais et français, montrent la viabilité de l'approche.

Keywords: Multilingual text categorization, Language identification, Machine learning, n-grams representation, CLEF collection, Translation effects on text categorization.

1 Introduction

La catégorisation de texte consiste à chercher une liaison fonctionnelle entre *un ensemble de textes* et *un ensemble de catégories* (étiquettes, classes). Cette liaison fonctionnelle, que l'on appelle également *modèle de prédiction*, est estimée par un apprentissage automatique (traduction de *machine learning method*). Pour ce faire, il est nécessaire de disposer d'un ensemble de textes préalablement étiquetés, dit *ensemble d'apprentissage*, à partir duquel nous estimons les paramètres du modèle de prédiction le plus performant possible, c'est-à-dire le modèle qui produit le moins d'*erreurs* en prédiction (voir la figure 1).

Formellement, la catégorisation de texte consiste à associer une valeur booléenne à chaque paire $(d_j, c_i) \in \mathcal{D} \times \mathcal{C}$, où \mathcal{D} est l'ensemble des textes et \mathcal{C} est l'ensemble des catégories. La valeur V (Vrai) est alors associée au couple (d_j, c_i) si le texte d_j appartient à la classe c_i tandis que la valeur F (Faux) lui sera associée dans le cas contraire. Le but de la catégorisation de texte est de

Titre

On parle de document non structuré au sens du data mining, cela ne veut pas dire qu'un document n'obéit pas à une certaine organisation. *Ex. article scientifique.*

Résumés

Un titre est censé être très important, il doit refléter le contenu de l'article.

Le résumé est un condensé de l'article, il en reprend les grandes lignes et les conclusions.

Etc.

Mots-clés

Organisation en sections, paragraphes.



L'analyse statistique du texte doit en tenir compte.

L'analyse textuelle est un problème ancien

L'exemple le plus simple (et ancien) est le traitement des questions ouvertes dans les enquêtes.

Text mining une nécessité de plus en plus patente avec la profusion des documents au format numérique

Profusion des documents, augmentation des capacités de stockage, apparition de nombreux corpus spécialisés (ex. [MEDLINE](#)).

Text mining incontournable avec l'explosion du web

Le web 2.0 a démultiplié dans des proportions gigantesques la disponibilité de documents rédigés, stockés numériquement (ex. newsgroups, réseaux sociaux [twitter], forums, etc.). Les opportunités d'extraction de connaissances utiles sont innombrables.

Requête sur les mots clés « text mining » sur stack overflow

Une question clé est « comment a été quantifiée la pertinence ? » (cf. recherche d'information).

The screenshot shows the Stack Overflow search results for the query 'text mining'. The search bar at the top left contains the text 'text mining' and a blue 'search' button. Below the search bar, it indicates '1,382 results'. A green arrow points from the text box in the yellow callout to the 'relevance' sorting option, which is currently selected. The results list shows three questions:

- Q: Data Mining and Text Mining**
What is the difference between Data **Mining** and **Text Mining**? Both refers to the extraction of unstructured data to structured ones. Is both forms work in the same fashion? please provide a clarity on that. ...
asked Mar 1 '15 by Vineeth Bhaskaran
Tags: nlp, bigdata, nltk, data-mining, text-mining
3 answers
- Q: Text mining on mobile**
I have some **text** files or html file. I want to retrieve some words from the **text**. My worry is that I have to do **text-mining** on mobile. Which algorithm can I use for **text-mining** on a mobile. A sample example would be appreciated. ...
asked Jul 24 '12 by user1545059
Tags: android
1 vote, 0 answers
- Q: Chinese Text Mining**
I used Chinese word segment to do **Text Mining**. And I changed data type to dataframe had comma and double quotation mark. So the wordcloud is strange. Like this: My syntax as below: inspect ...
asked Mar 1 by Chris Chung
Tags: r, text-mining, word-cloud
1 answer

On the right side, there is a yellow banner for 'Want a python job?' with job listings for 'Python Developer' and 'Data Engineer'. Below that is a section for 'Hot Network Questions' with a question about handling nitric acid spill.

Nous nous situons dans le cadre du TALN (traitement automatique du langage naturel ; en anglais NLP : natural language processing). Nous le considérons sous l'angle des approches statistiques.

Ce que nous traiterons

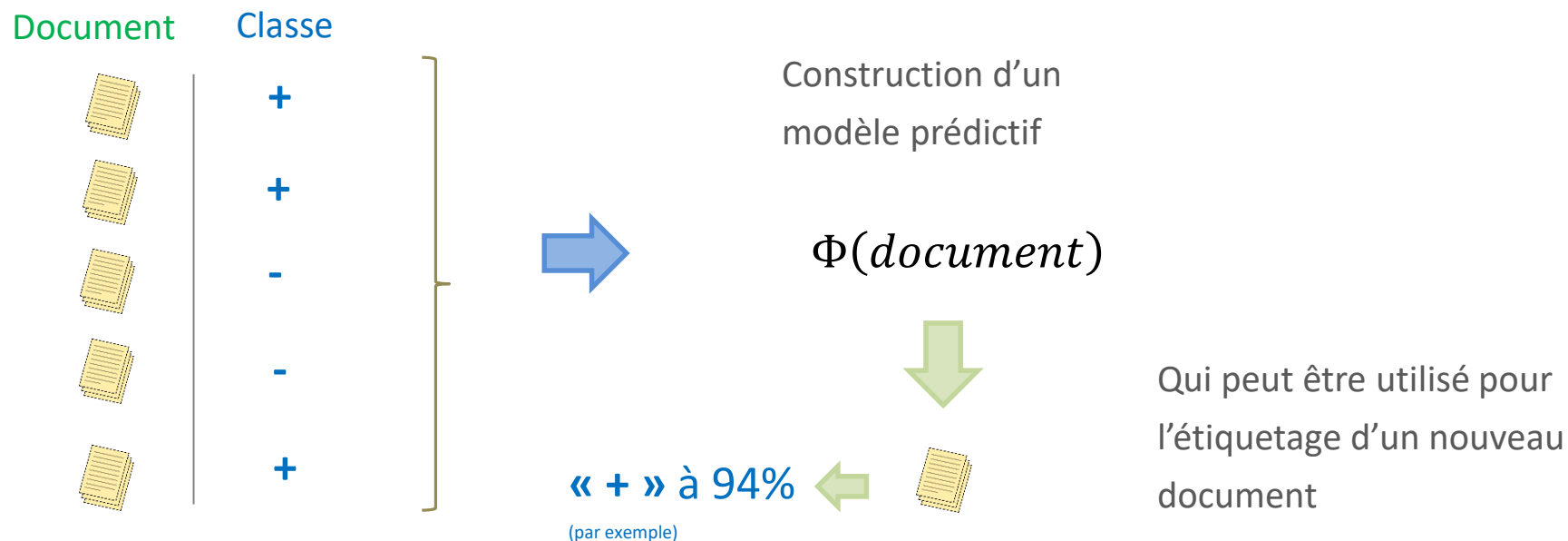
- Traitement des corpus à l'aide des algorithmes de statistique et de data mining
- Les descripteurs sont automatiquement extraits des corpus.
- Sans connaissances approfondies sur le domaine étudié.

Ce que nous ne traiterons pas

- Les approches nécessitant des connaissances fines sur le domaine.
- Ou nécessitant des connaissances linguistiques avancées.

APPLICATIONS DU TEXT MINING

Cadre de l'apprentissage supervisé : exploiter une collection de documents préalablement étiquetés



Remarques

- On parle de « document classification » en anglais
- Le processus permet également d'associer les documents à des termes prédéfinis, on parle alors d'indexation automatique (ex. Reuters)
- Le cadre le plus fréquent est binaire avec une classe d'intérêt (« + ») contre les autres (« - » qui peuvent être constitués de documents très hétérogènes)
- Une application phare est le filtrage automatique de documents (ex. e-mail, pages web, etc.) ; mais il y en a d'autres (analyse des sentiments, opinion mining, etc.)

```
Return-Path: <...@outlook.fr>
Received: from mailgw.univ-lyon2.fr ([159.84.182.44])
    by co5.univ-lyon2.fr with ESMTP ; Wed, 9 Nov 2016 10:31:49 +0100 (CET)
Received: from mx02.univ-lyon2.fr (mx02.univ-lyon2.fr [159.84.182.37])
    by mailgw.univ-lyon2.fr (Postfix) with ESMTP id E74EBEB311;
    Wed, 9 Nov 2016 10:31:29 +0100 (CET)
Received: from mx02.univ-lyon2.fr (localhost [127.0.0.1])
    by mx02.univ-lyon2.fr (Postfix) with ESMTP id 8B2758A0FA;
    Wed, 9 Nov 2016 10:31:27 +0100 (CET)
X-Virus-Scanned: amavisd-new at univ-lyon2.fr
X-Spam-Flag: NO
X-Spam-Score: -1.005
X-Spam-Level:
X-Spam-Status: No, score=-1.005 tagged_above=-5 required=3
    tests=[ADVANCE_FEE_3_NEW_MONEY=0.001, BAYES_00=-1.9,
    CRM114_CHECK=-0.12, FREEMAIL_FROM=0.001, HTML_MESSAGE=0.001,
    LOTS_OF_MONEY=0.001, MISSING_HEADERS=1.021,
    RCVD_IN_DNSWL_NONE=-0.0001, RCVD_IN_MSPIKE_H3=-0.01,
    RCVD_IN_MSPIKE_WL=-0.01, T_MONEY_PERCENT=0.01]
    autolearn=no autolearn_force=no

...
Hello ,

My Name is Mrs. Helen , a widow with my only daughter, she is 16 years of a
ge, I am writing to ask for your help to please accept my child in the name
of God and help me to raise her up, I grew up as an orphan and now a widow
, I have nobody to help take care of her because I was recently diagnosed c
ancer and I have only 2 months to live, I don't want her to suffer again be
cause she has suffered a lot ever since the death of her father and my illn
ess, we have the sum of 5.000.000 Five million dollars which we inherited f
rom her late father daniel Akah (my husband)I am willing to transfer money
to you while you help my daughter to come over to your country to continue

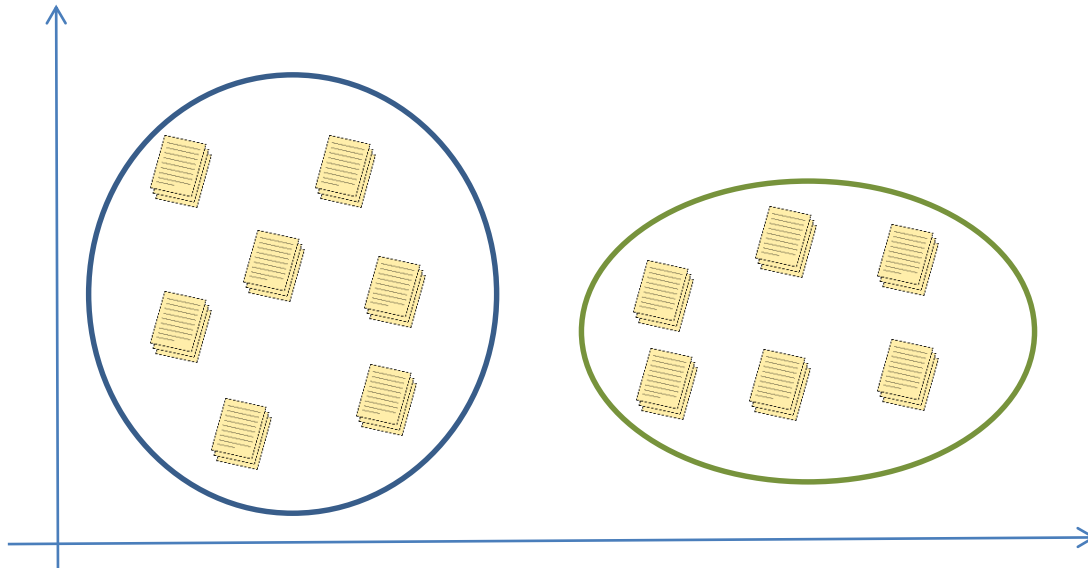
...
God bless you!
Best Regards.
```

Une base d'apprentissage constituée de documents SPAM et NON-SPAM ont permis de construire un modèle prédictif.

Ce modèle issu du processus data mining a été déployé sur le serveur. Il attribue un score aux e-mails entrants. Ceux qui présentent un score trop élevés sont bloqués.

Le serveur de l'université n'a pas réussi à détecter celui-ci !

Cadre de l'apprentissage non-supervisé : partitionner en groupes homogènes les documents



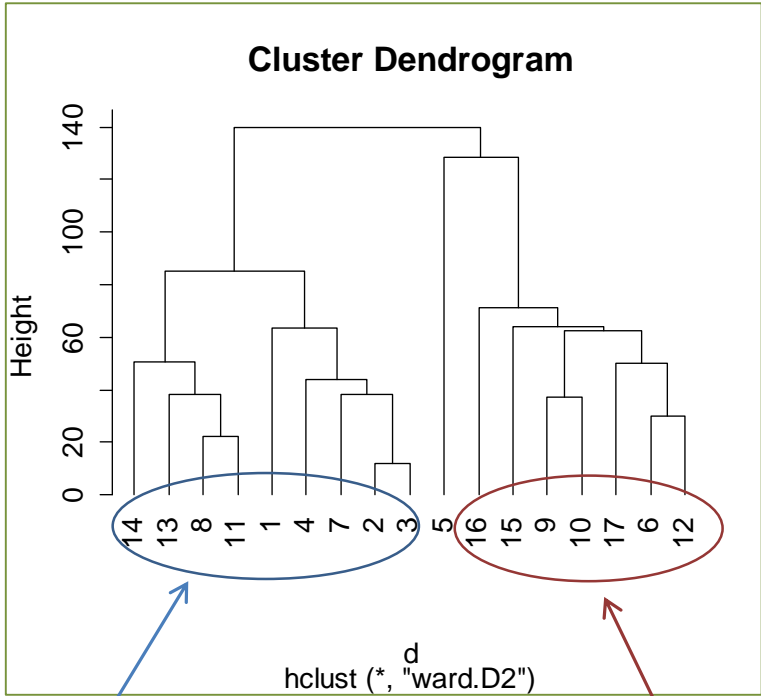
Les documents d'un même groupe sont le plus similaires possible.
Les documents de groupes différents sont le plus dissimilaires possible.



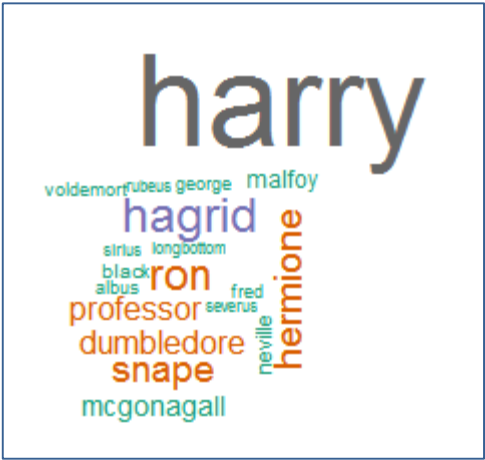
On bénéficie des avantages liés à la structuration : une organisation des documents pour faciliter la consultation et la recherche ; disposer d'un résumé du corpus (une forme de réduction de la dimensionnalité) ; etc.



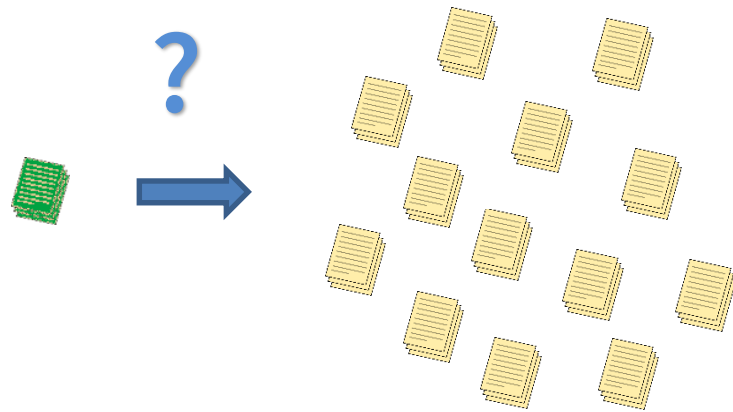
L'enjeu, toujours en clustering, est de comprendre la nature des groupes, en les associant à des thèmes par exemple.



Analyse par chapitre des occurrences des personnages dans le 1^{er} tome de Harry Potter



Harry est présent partout. Normal. Mais, selon les chapitres, Hagrid, Ron ou Hermione se distinguent.



L'objectif de la recherche d'information (*information retrieval* en anglais) est de mettre en place les stratégies permettant d'identifier, dans un corpus, les documents pertinents relatifs à un document requête. Il s'agit d'une **recherche par le contenu**, le texte est concerné, mais elle peut s'étendre à l'image, la vidéo, le son...



L'indexation joue un rôle fondamental dans la recherche, la **mesure de similarité** entre documents...



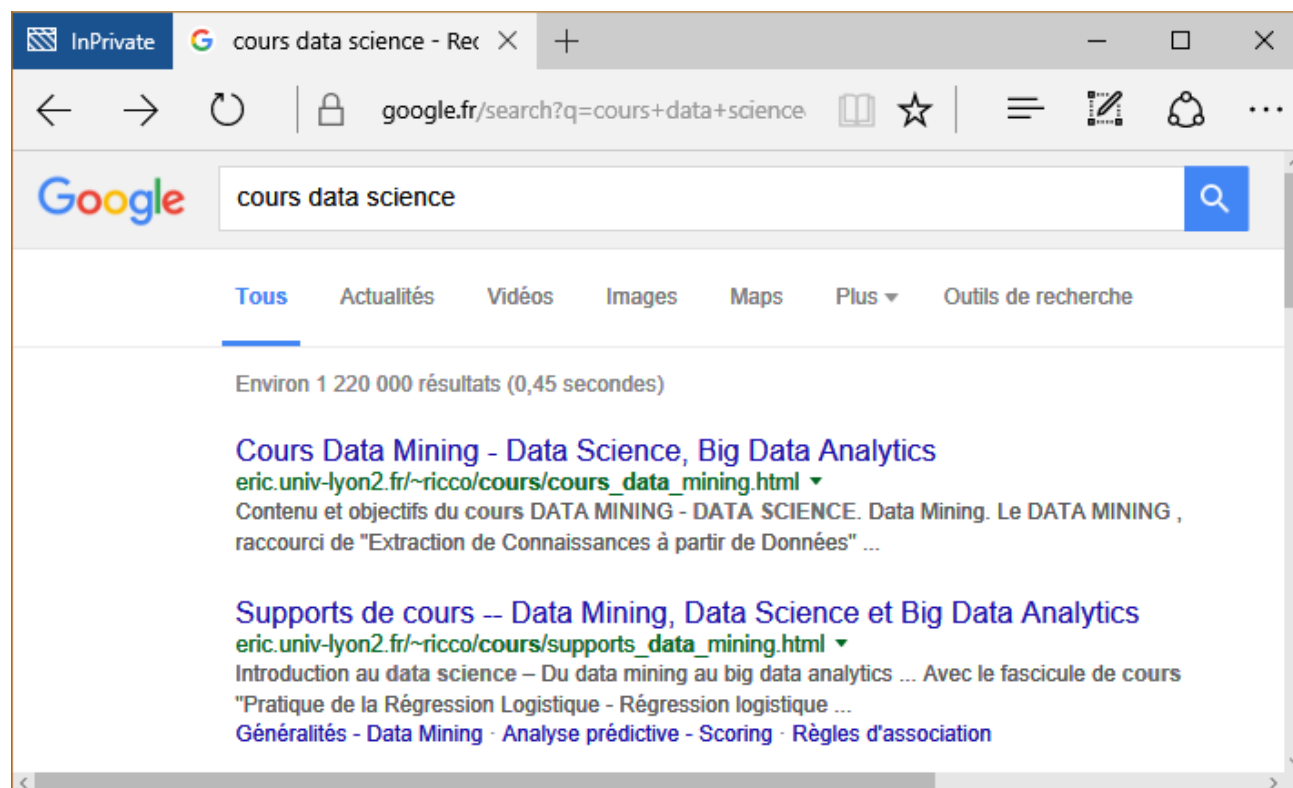
Les applications sont très nombreuses. Ex. détection de plagiat, recherche de dossiers médicaux, etc.



Il existe des bibliothèques logicielles performantes (Ex. Lucene, Elasticsearch).

Les moteurs de recherche web constituent l'exemple le plus emblématique de la recherche d'information, ... mais avec des spécificités : le document requête est très court, constitué de quelques mots clés ; des stratégies de spamming viennent polluer la recherche (Voir [Historique des moteurs de recherche](#))

La popularité de Google avec le concept [PageRank](#) repose en très grande partie sur sa capacité à parer les assauts des spammeurs, en tenant compte des liens, et de la crédibilité (critère d'autorité) de l'auteur.



L'extraction d'information consiste à recherche des champs prédéfinis dans un texte plus ou moins rédigé en langage naturel. On s'appuie plus sur l'analyse lexicale et morphosyntaxique pour identifier les zones d'intérêts.

Annonce 1

Description :
FIAT COUPE 2L 20V 155 CV
1999
136000 km
Contrôle technique ok
Prix: 2990 €
Equipement :
Cuir. Climatisation. Vitres électriques. Rétroviseurs électriques. Direction assistée. Fermeture centralisée. ABS. Airbag. Jantes aluminium

Ex. extraire automatiquement les informations additionnelles sur les annonces automobiles.

Annonce	2L	5 cylindres	Turbo	Bolidée	1ère main	Réparations à prévoir
1	oui	oui	non	non	?	non
2	oui	non	oui	oui	?	oui
3

Annonce 2

Fiat coupe T 16 (moteur lancia delta dorigine) an 95
tuning leger rabaisse av et ar blistein b8 ressort court ct ok en cours le vehicule roule tres peu (suite perte emploi) il es entretenu marche tres fort environ 200 ch il peut parcourir tous les trajet. je ne suis pas presse alors les marchand de tapis pas la peine de venir la peinture a ete refait par lancien proprietaire ainsi que lebrayage mais elle as des imperfections demande de renseignement par tel uniquement

Objectif : Récupérer automatiquement les informations qui ne sont pas standardisées sur leboncoin.fr (ex. pour les voitures : année modèle, kilométrage, carburant, boîte)

Résumé automatique ([Automatic summarization](#))



Résumé d'un document. Recherche des phrases les plus représentatives dans un document.

Résumé d'un corpus. Recherche du document le plus représentatif dans un corpus ou Recherche des phrases représentatives à partir de plusieurs documents.

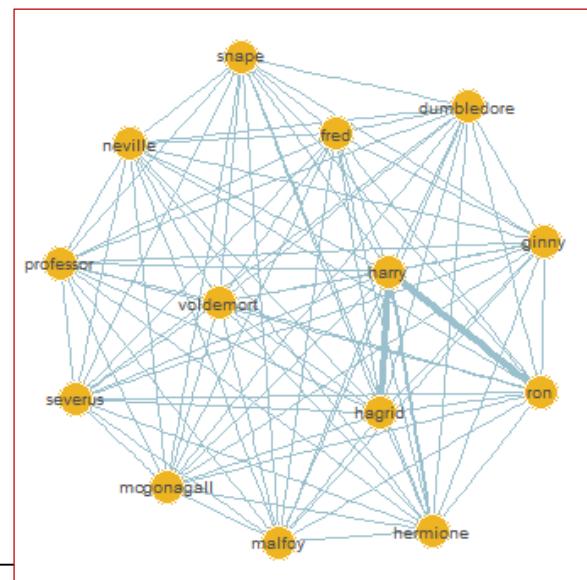
La notion de distance entre phrases, entre documents, joue un rôle important.

Identification des tendances

Les données arrivent en flux. S'appuyer sur la temporalité (ex. textes sur les réseaux sociaux). Analyser l'évolution de la popularité des thèmes, identifier les sujets émergents (*topic detection*).

Analyse des liens

Analyser les relations (cooccurrences) entre les termes, qui peuvent être des personnages (nous approfondirons cette idée dans l'analyse des réseaux sociaux)



Transformation d'une collection de textes en un tableau de données

Sans nécessiter de connaissances particulières sur le domaine étudié

REPRÉSENTATION « BAG-OF-WORDS »

Processus de « tokenization » : découpage d'un flux de caractères en phrases, symboles, mots. Nous nous intéressons plus particulièrement aux mots.

Plusieurs étapes :

1. Repérer les mots (*tokens*) présents dans les documents
2. Qui vont constituer le dictionnaire
3. Les mots deviennent des descripteurs (features, *termes*)
4. On associe alors l'absence ou la présence des mots à chaque document

L'identification du délimiteur de mots est important. Ce sera souvent l'espace, les ponctuations,... Certains caractères sont moins évidents (ex. « - » dans « écart-type »)

Potentiellement, le nombre de mots est très important. Il peut y avoir des redondances dans le dictionnaire, il faudra pouvoir les traiter (parfois évidentes : voiture vs. voitures ; parfois moins : voiture vs. automobile...).

Ca peut être aussi le nombre d'apparition des mots, ou autre type de valeurs. On parle de *pondération*.

4 documents (Coelho & Richert, 2015 ; page 55, seuls les 4 derniers documents ont été repris) :

1. Imaging databases can get huge
2. Most imaging databases save images permanently
3. Imaging databases store images.
4. Imaging databases store images. Imaging databases store images. Imaging databases store images



Document	imaging	databases	can	get	huge	most	save	images	permanently	store
1	1	1	1	1	1	0	0	0	0	0
2	1	1	0	0	0	1	1	1	1	0
3	1	1	0	0	0	0	0	1	0	1
4	1	1	0	0	0	0	0	1	0	1

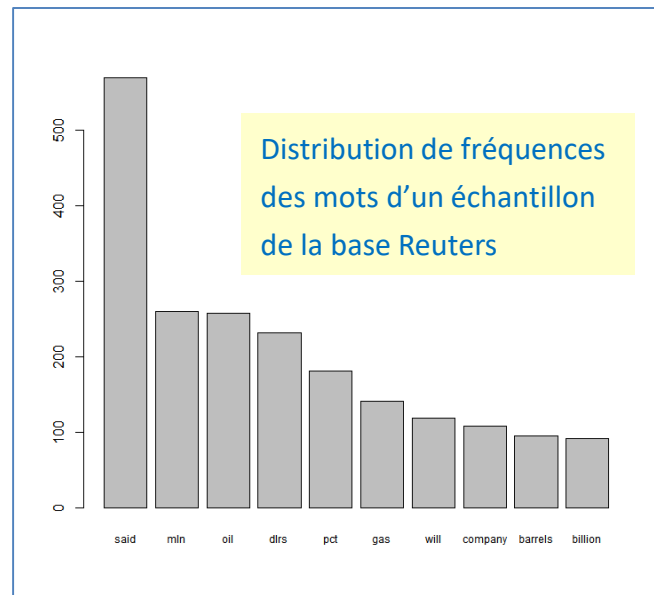
Nous percevons les problèmes liés à ce type de représentation :

Remarques

- Certains mots ne sont pas intrinsèquement porteurs de sens (ex. most) ;
- Certains mots véhiculent la même idée (ex. imaging vs. images) ou sont synonymes (ex. store vs. save)
- La pondération joue un rôle important (ex. n°3 et n°4 sont strictement identiques au sens de la présence/absence des mots, indûment ?)
- La question de la mesure de similarité/dissimilarité entre les documents sera centrale

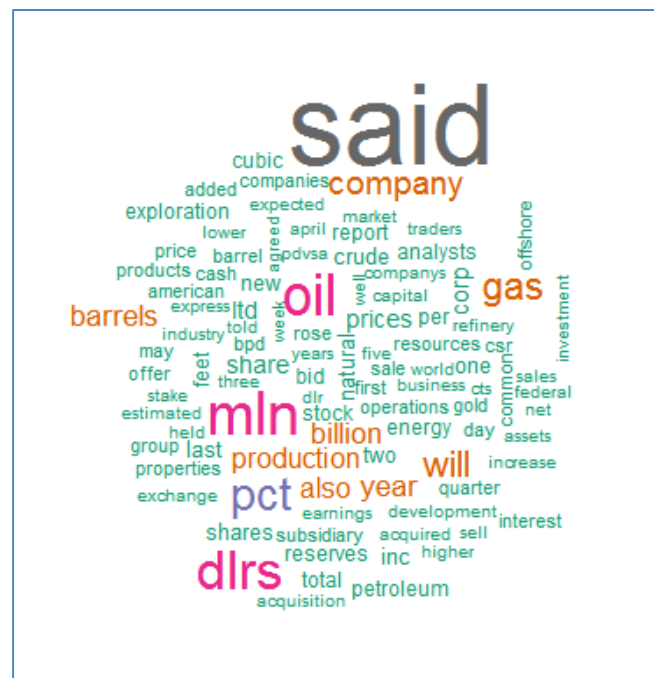
Loi de Zipf. La fréquence d'un mot est inversement lié (proportionnel) à son rang « r » dans l'ordre des fréquences, avec la formule

$$f(mot) = \frac{c}{r} \text{ (où } c \text{ est une constante)}$$




Word cloud. Une représentation populaire est le nuage « word cloud » où la fréquence d'un mot est signifiée par sa taille et sa couleur (pas la position).

Remarque : « said » est le terme le plus fréquent, on se rend compte que le nettoyage du texte est primordial.



Le tableau issu de la représentation BOW n'est pas tout à fait un tableau de données comme les autres....

- 
- La **dimensionnalité** est souvent très élevée. Le nombre de colonnes peut excéder le nombre de lignes. Problème pour les algorithmes de data mining souvent. **La réduction de la dimensionnalité est un enjeu crucial.**
 - Les termes sont présents dans peu de documents. Nombre des vecteurs lignes contiennent la valeur zéro. Il faut utiliser la représentation sous forme de « **sparse-vector** » pour économiser la mémoire. Encore faut-il que les implémentations des algorithmes sachent les exploiter.
 - Les valeurs du tableau sont toujours positives ou nulles. On se concentre sur les occurrences positives des termes (ex. avec les règles d'association)
 - Les descripteurs (colonnes) sont exprimés dans les mêmes unités. Il n'y a pas de normalisation ou de standardisation des variables à effectuer.
 - Il n'y a jamais de données manquantes (N/A). Un terme est forcément soit absent, soit présent dans un document.

Aller au-delà des mots présents (observés) dans un corpus

AU-DELÀ DE BAG-OF-WORDS

D'autres informations peuvent compléter la représentation bag-of-words dans la construction de la matrice des données.

Received: from PACIFIC-CARRIER-ANNEX.MIT.EDU by
po8.MIT.EDU (5.61/4.7) id AA21953; Sun, 5 Dec 99
17:31:46 EST

BENCHMARK SUPPLY
7540 BRIDGEGATE COURT
ATLANTA GA 30350

LASER PRINTER TONER CARTRIDGES
FAX AND COPIER TONER

CHECK OUT OUR NEW CARTRIDGE PRICES :

APPLE

LASER WRITER PRO 600 OR 16/600	\$69
LASER WRITER SELECT 300,310.360	\$69
LASER WRITER 300, 320	\$54
LASER WRITER LS,NT,2NTX,2F,2G & 2SC	\$54
LASER WRITER 12/640	\$79

HEWLETT PACKARD

LASERJET SERIES 2,3 & 3D (95A)	\$49
--------------------------------	------

Received: from PACIFIC-CARRIER-ANNEX.MIT.EDU by
po9.MIT.EDU (5.61/4.7) id AA10612; Sat, 10 Apr 99
14:19:17 EDT

From: <

Received: from ipa214.miami15.fl.pub-ip.psi.net by
MIT.EDU with SMTP

Reply-To:

Subject: The Hottest Site On The Net!! (270)

>>>NEVER Seen Before Amatuer Pics and Videos<<<

1000's of Hardcore Teen Pics!

1000's of "First Time" Girl Next Door Pics!

Teen Movies With Sound!

Hidden Bathroom and Bedroom Cams!

1000's of Video Feeds!

Videos With "Real Time" Chat!

Live Amateur Sex Shows!

Plus More...More...More...Something for Everyone!

It Doesn't Get Any Better Than This!!!

Ex. Spam = proportion du texte en majuscules, présence excessive de certaines ponctuations (ex. !), etc.

Répertorier, classer et mettre en relation le contenu sémantique et lexical d'une langue (ex. WORDNET pour la langue anglaise) ([Wikipédia](#)).

Synset

Les synset (synonym set) correspondent à des groupes de mots interchangeables. Ex. en anglais : car, automobile, machine, motorcar, etc. Dans la représentation bag-of-word, ils permettent de réduire considérablement la dimensionnalité.

Ontologies (Arbres de concepts)

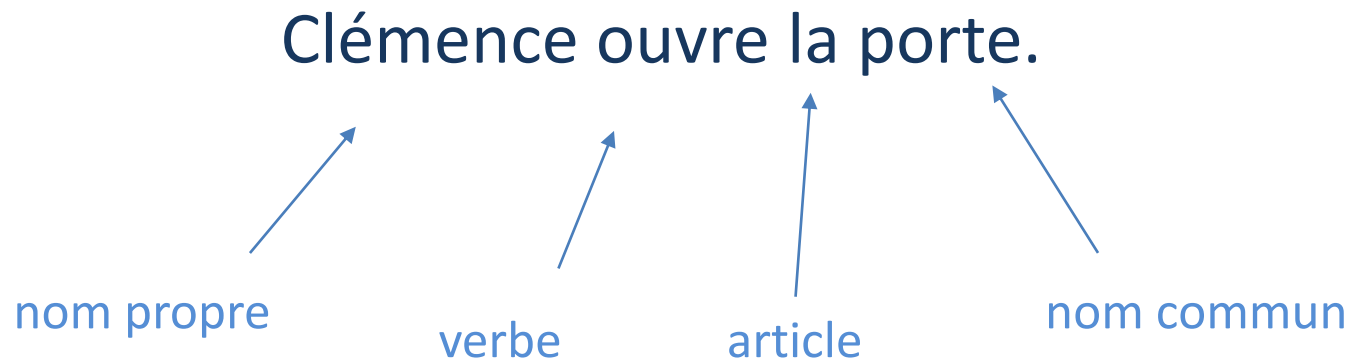
- car, auto, automobile, machine, motorcar motor vehicle, automotive vehicle
 - vehicle
 - conveyance, transport
 - instrumentality, instrumentation
 - artifact, artefact
 - object, physical object
 - entity, something



Ici aussi, il est possible de réduire considérablement la dimensionnalité.

Le schéma BOW fait partie du NLP, mais on peut aller encore plus loin en exploitant les connaissances que nous pouvons avoir sur le langage naturel.

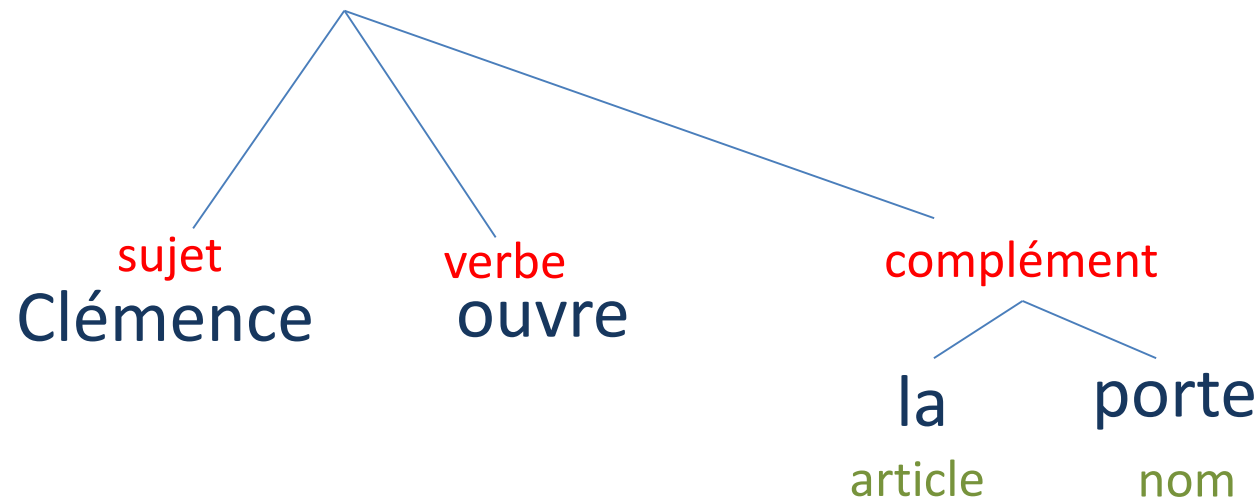
Ex. **Part of speech** (POS) se propose de distinguer les mots d'une phrase selon leurs catégories lexicales (ex. nom, verbe, adjectif, préposition, etc.).



On peut utiliser des bibliothèques prédéfinies. Mais on peut aussi tagger les mots par apprentissage [**part-of-speech tagging**]. En effet, il peut être spécifique à un domaine. Ex. « porte » peut venir du verbe « porter », comme désigner une « porte ». Dans la menuiserie, on a plus de chances d'être dans le second cas.

Ex. La bibliothèque NLTK (Python) peut construire le système de classement à partir d'un **corpus** où les termes ont été annotés. Dans le pire cas, on adopte la modalité la plus fréquente.

La décomposition de la phrase peut enrichir la compréhension du rôle des mots. En effet, elle (la phrase) obéit à une certaine organisation.



On comprend sans ambiguïté que « porte » est un nom dans ce contexte.



On comprend aussi que Clémence correspond à un nom propre. A la différence de son rôle dans « Je demande la clémence pour les cancre ».

L'objectif est d'associer le sens adéquat à un mot en fonction du contexte dans lequel nous nous situons.



Robert est poli.

Le meuble est poli.

Dans une représentation en BOW, le terme « poli » peut avoir des sens différents, on ne peut pas l'utiliser tel quel.



Une approche possible consiste à appliquer un algorithme supervisé de data mining sur un corpus où le sens des termes a été annoté manuellement par un expert.

Ouvrages

Weiss S., Indurkha N., Zhang T., Damerau F., « Text Mining – Predictive methods for analyzing unstructured information », Springer, 2005.

Feldman R., Sanger J., « The text mining handbook – Advanced approaches in analyzing unstructured data », Cambridge University Press, 2008.

Aggarwal C., Zhai C., « Mining Text Data », Springer, 2012.

Ignatow G., Mihalcea R., « Text mining – A guidebook for the social science », Sage Publications, 2016.

Perkins J., « Python text processing with NLTK 2.0 cookbook », Packt Publishing, 2010.