FOUILLE DE TEXTE

U.E Fouille de données supervisée

sept 2021

Text Mining

- Ensemble des techniques et méthodes destinées au traitement automatique de données textuelles **en langage naturel**
  - Disponible sous format numérique
  - En grande quantité
  - → En dégager et/ou structurer le contenu
    - → Ex : Détection de thème, Détection d'opinions, Extraction d'entités nommées, Résumé automatique…

Data Mining / Text Mining

- Data mining
  - Données généralement issues de BDR d'entreprises, de laboratoire de recherche
    - → Structurées
  - Définition explicite des attributs
- Text mining
  - Acquisition de connaissances à partir de *corpus* textuels
  - Corpus : recueil de documents d'une même discipline
  - Web : mine géante de corpus!

Exemples concrets de tâches en TAL

- Catégorisation
  - SNCF :
    - Classification automatique de documents techniques
    - Clustering pour recherche d'information
  - LS2N : Dossier de candidatures à la fac
- Détection d'opinions
  - Orange : Satisfaction client à partir de plateforme service client
  - MMA : Analyse besoins clients sur retours
  - IPPON : Prospection commerciale
- Extraction d'entités nommées
  - Ina : Indexation des fonds audiovisuels

Spécificités du corpus

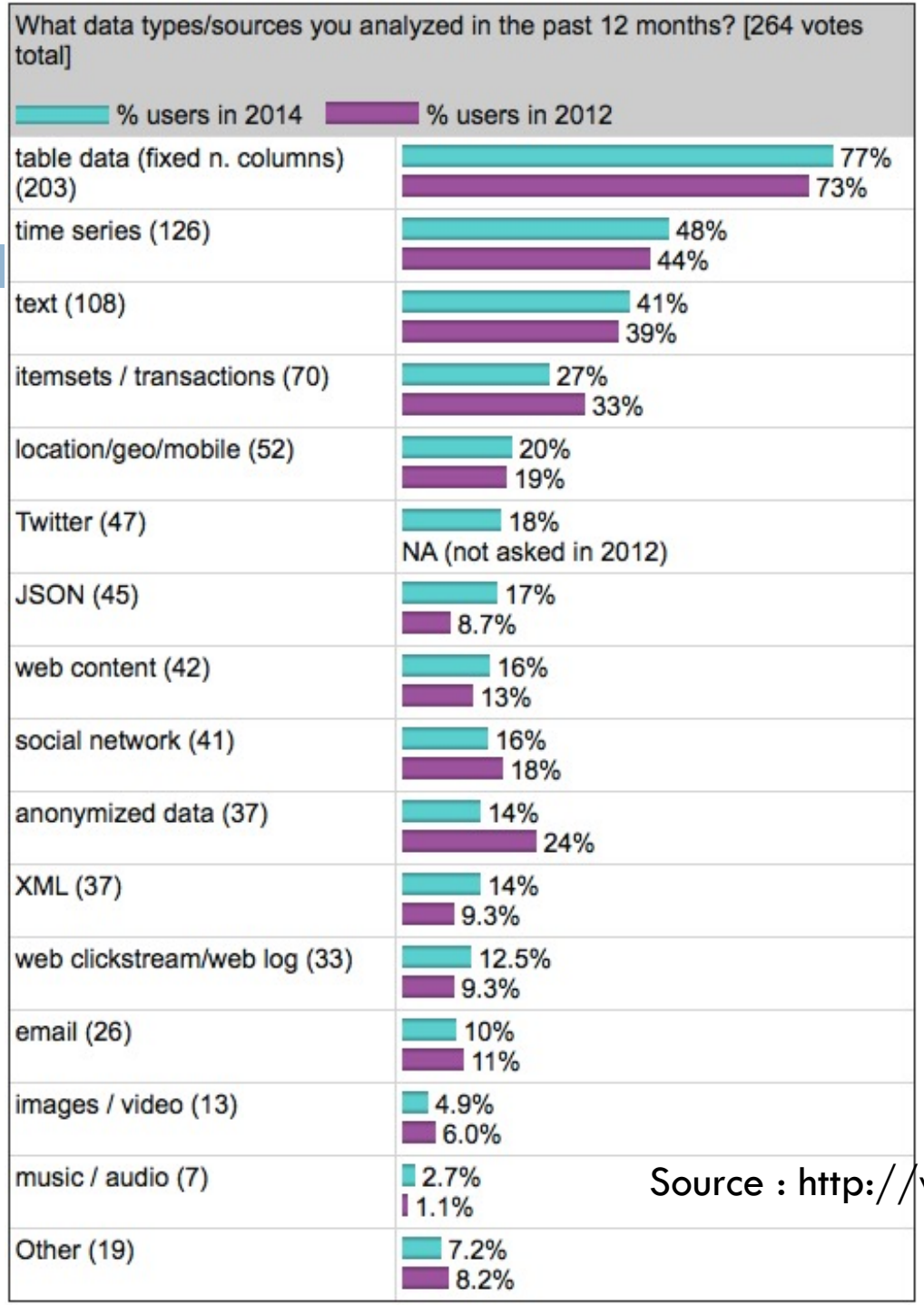
- Satisfaction aux conditions :
  - Corpus d'assez grande taille
  - Compréhensibilité et cohérence pas trop basses
  - Faire le moins souvent possible appel au sous-entendu, à l'ironie, à l'antiphrase

    « ne dites pas merci surtout! »

    « oh qu'il est beau…»
- Deux problématiques …
  - Compréhension / Structuration
  - Représentation

Principales sources de textes utilisées

- Revues de presses, dépêches AFP
- Transcriptions d'entretiens téléphoniques
- Curriculum vitae
- Enquêtes d'opinion
- Blogs, Tweets, Réseaux sociaux
- Lettres de réclamation
- Messageries électroniques
- …



Source : http://www.kdnuggets.com/

Exemples d'applications de text mining

- Classification de courriers
  - Le courrier est-il désiré ?
    → Détection automatique de spam
- Système de question/réponse
  - En fonction d'une question précise, quels sont les documents (ou passages) susceptibles d'apporter la réponse
  - → ≠ Moteur de recherche classique
- Résumé automatique
  - Sélectionner les phrases représentatives d'un texte, voire reformuler

Au delà du texte

- Exploration de données audio, ex : Fonds radiophoniques de l'Ina
  - Que contient un corpus retrouvé de plusieurs centaines d'heures ?
  - À partir de transcriptions automatiques, pouvoir structurer le corpus (thèmes, mots clés, locuteurs, ...)

Au delà du texte

- TRECVID : fouille de vidéos
  - Fouille de données audiovisuelles, à partir de connaissances extraites automatiquement du flux vidéo
    - Analyse d'images
    - Reconnaissance de formes
    - Textes incrustés
    - Transcription automatique de la parole

Les particularités…

- Différents formats

  → où se trouve l'information?

- Différents encodage
  - Problème récurrent!
- Traitement du langage naturel = ambiguïté
  - Problème de caractères, de casse, de polysémie, d'homonymes…

Différents formats de texte

- Texte « brut »
- Transcriptions d'oral
- Pages WEB (HTML), images et textes
- Textes structurés (XML)

Texte « brut »

Extrait de « Le monde »

- { \rtf1\ansi \deff0\plain Document soumis aux dispositions du droit d'auteur. Tous droits r\'E9serv\'E9s.
  \par ----------
  \par \b\fs34 Le Monde\b0\fs24
- \par
  \par
  \par 31 d\'E9cembre 1996, page 1\par
  \par
  \par
  \par HORIZONS - ANALYSES ET DEBATS\par \b\fs34 L'Allemagne se sent plut\'F4t bien\b0\fs24
  \par
  \par \b DELATTRE LUCAS\b0
  \par
  \par C'\'C9TAIT, il y a peu, \'E0 Bonn. Vendredi, 15 h 30. Helmut Kohl, seul, quitte son bureau et traverse tranquillement le parc de la chancellerie. Sa semaine de travail est termin\'E9e. Le chancelier allemand se rend dans sa villa priv\'E9e, au fond du jardin, ce que l'on appelle ici le "bungalow". L\'E0, quelques lectures d'agr\'E9ment l'attendent un roman historique ou une biographie, sans doute.
  \par
  \par Surprenante image.
- }

Format ctm

- Forme de transcription automatique d'un document audio ou vidéo

08730_1007.l 1 47.87 0.12 oui

08730_1007.l 1 54.63 0.22 euh

08730_1007.l 1 54.85 0.49 réserver

08730_1007.l 1 55.61 0.15 [CARILLON]

08730_1007.l 1 55.83 0.07 [PAROLE]

08730_1007.l 1 55.9 0.1 [PAROLE]

08730_1007.l 1 56.0 0.11 [CARILLON]

08730_1007.l 1 58.97 0.13 la

08730_1007.l 1 59.09 0.28 place

08730_1007.l 1 59.38 0.22 plein

08730_1007.l 1 59.59 0.4 tarif

08730_1007.l 1 60.0 0.42 et

Format XML


```
<?xml version="1.0"?>
<menu_petit_dejeuner>
        <nourriture>
                <nom>Cafe croissants</nom>
                <prix>*5.95</prix>
                <description> Caf^^e9 cr^^e8me avec deux croissants, beurre et confiture
</description>
                <calories>650</calories>
        </nourriture>
        <nourriture>
                <nom>Pain fromage</nom>
                <prix>*5.95</prix>
                <description> Choix de fromage, pain de seigle, beurre </description>
                <calories>750</calories>
        </nourriture>
        <nourriture>
                <nom>Petit d^^e9jeuner anglais</nom>
                <prix>*10.95</prix>
                <description> Oeufs avec bacon, pain et confiture, tranche de pud-ding
maison </description>
                <calories>750</calories>
        </nourriture>
</menu_petit_dejeuner>
```

Structuration des données textes

- Données non structurées
  - Fichier textes (txt, rtf, doc, …)
  - Pages web (article wikipedia, blog, site institutionnel,…)
- Données structurées
  - Format TEI
    - Text Encoding Initiative : www.tei-c.org
    - Représentation de textes sous formes balisées (SGML, XML)
  - Format RDF
    - Resource Description Framework : www.w3.org/RDF
    - Web sémantique

Gestion de l'encodage

- Encodage :
  - Association d'un jeu de caractères naturels à un jeu de caractère codifié (Ex : code morse, code braille, code ASCII)
- Prise en compte obligatoire
  - Utile pour la lecture
  - Indispensable pour les traitements
- Encodage utilisé
  - Précisé dans le document (fichier XML)
  - Commande shell « *file* »
- Encodages les plus fréquents
  - UTF-8 : comprend tous les caractères utilisés en français
  - Iso-Latin-1: ne comprend pas Ÿ, ni œ et Œ

Gestion de l'encodage

- Même encodage d'écriture et de lecture:

  Une police de caractères est un ensemble de glyphes ou représentations visuelles de caractères d'une même famille.

- Encodage iso-latin-1 et lecture UTF-8

  Une police de caract###s est un ensemble de glyphes ou repr###ntations visuelles de caract###s d'une m### famille.

- Encodage UTF-8 et lecture iso-latin-1

  Une police de caractÃ¨res est un ensemble de glyphes ou reprÃ©sentations visuelles de caractÃ¨res d'une mÃªme famille.

Problèmes d'ambiguïté

- Dans quelle langue est écrite le texte?

  « *Nixon put dire on tape* »

  « *Garage sale* »

→ Problème de la phrase polyglotte

Problèmes d'ambiguïté

- Polysémie, homographes…

« les poules du couvent couvent »

→Analyse des catégories grammaticales (pas si simple!)

- Typographie pauvre

« CET ENTREPOT FERME A CAUSE DES EMEUTES »

→ Quelle lecture? Quel événement déclenche l'autre?

Particularités des données textuelles

- Où se situe la connaissance ?
  - Au niveau des mots, des phrases ?
  - Porté par le document tout entier, une partie ?
  - Dans les balises?
- Comment représenter le langage naturel ?
  - Utiliser la forme de surface (le mot) ou une forme de plus haut niveau ?
  - Comment traiter la séquentialité?

Des prétraitements nécessaires

- Nettoyage
  - Étape primordiale de « préparation » du texte
- Normalisation
  - Étape d'homogénéisation du texte
- Représentation du texte
  - Choix de l'unité de base
  - Choix des descripteurs

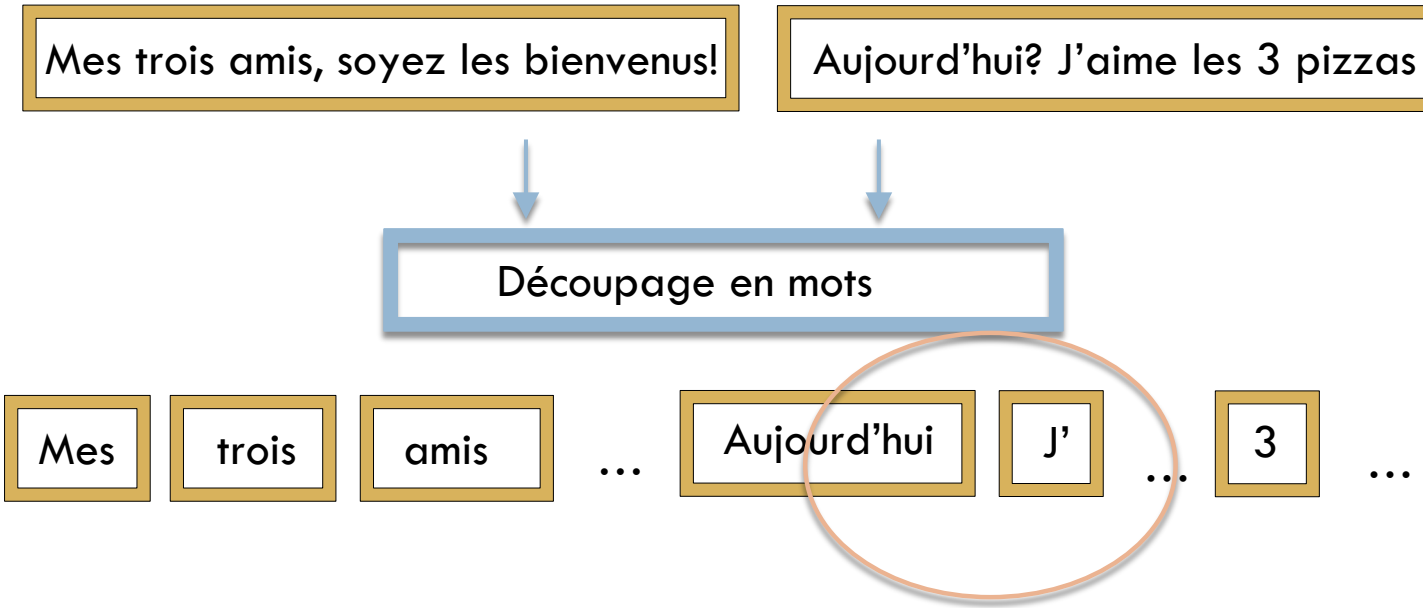
Nettoyage

- Étape 1 : Isoler et structurer l'information
  - Extraire les parties utiles
  - Aucune mise en forme
  - Traiter l'encodage
- Exemple : une page html
  - Isoler les zones intéressantes (titre, auteur, date, contenu)
  - Ecarter le reste
  - Gérer les caractères

Ou pire… un pdf2txt !

Découpage en mots (tokenisation)

- Segmenter une séquence de caractères en des unités appelées *mots* (les éléments d'information)



  - Basé sur les espaces et la ponctuation, mais…

Qu'est-ce qu'un mot?

- Chaine de caractères entre 2 espaces?
  - *mots.* et *mots*
  - *Le* et *le*
  - *dix mille, 10000, 10,000, 10 000, …*
  - *5 minutes, 5m, 5 min, 5 ms, …*
  - Pomme de terre
  - Le président de la république française

Qu'est-ce qu'un mot?

- Comment savoir où couper?
  - Problème des apostrophes
    - *J'aime*, *aujourd'hui* : un ou deux mots ?
  - Problème des tirets
    - *ira-t-on*, *compte-rendu* : combien de mots ?
  - Problème des espaces
    - *San Francisco* : un ou deux mots ?
  - Problème des nombres
    - *1,3 ou 1.3*, *127.0.0.1*, *15/09/2013*
- Selon les langues, encore plus de problème!
  - Agglutination, alphabet non latin, sens de lecture…

Normalisation

- Tokenisation
- Étapes classiques (basiques)
  - Traitement de la casse
  - Traitement de la ponctuation
  - Traitement des chiffres en mots
  - Création de classes d'équivalence

Traitement de la casse

- Casse:
  - Distinction des lettres majuscules des lettres minuscules
- Prise en compte
  - Conserver la distinction majuscule/minuscule
    - Ex : Le premier mot de chaque phrase commence par une majuscule

    →« Mot » et « mot » seront 2 mots différents
  - Normaliser
    - Initiale en majuscule, le reste en minuscule
    - Tout mettre en minuscule

    →Perte d'information?

Traitement de la ponctuation

- Ponctuation:
  - Ensemble des caractères spéciaux indiquant la structure de l'énoncé
- Comment reconnaître un signe de ponctuation?
  - Point : fin de phrase, séparateur dans un acronyme, séparateur entre unités et dixièmes,…?
  - Deux points : ponctuation ou séparateur heure/minute?
- Prise en compte
  - Oui :
    - Isoler la ponctuation des mots (ajout d'espaces)
    - Remplacer les caractères spéciaux par des mots
  - Non :
    - Suppression
    - Perte d'information?

Traitement des chiffres

- Chiffres:
  - Tout caractère de 0 à 9
- Comment transformer en lettres?
  - « Il est 10:55 »
  - « je vais prendre 3,5 kg de pommes »
  - « Sa note de FDD est de 15,5 »
- Normaliser
  - dix sept et dix-sept
  - 15,5 et quinze → ?

Classes d'équivalence

- Regroupement de plusieurs mots au sein d'un groupe pour les représenter tous
- Exemples :
  - U.S.A , USA, US, U.S, United States …

Représentation en mots

Dans les textes, quelles sont les données?

- Les données → les mots
  - Entrée d'un dictionnaire
    - courant
    - vocabulaire spécialisé
  - Mots à caractères spéciaux
    - balises XML
    - liens hypertexte
    - hashtags

Représentation en mots

- Mots simples (*token*)
  - chaîne de caractères entre deux espaces
- Abréviations
  - Unités de mesures « kg », « ms »
  - Courantes « kilo », « sec. », « ciné »
- Mots composés
  - « aujourd'hui », « cent quarante-huit »
- Formes composées
  - « est-ce que », « c'est à dire »

Au delà de la forme de surface

- Lemme
  - Lemmatisation : réduction des formes fléchies à la forme canonique du mot
    - Verbe → infinitif
    - Mot → masculin singulier
    - Ex : souhaitons → souhaiter
  - Exemple :
    - suis, es, est, sommes, sera, étions → être
    - Les jolies souris vertes → le joli souris vert
  - Nécessite un étiqueteur morpho-syntaxique qui dépend de la langue

Au delà de la forme de surface

- Stemme
  - Racinisation (ou stemming) : découpage du mot pour obtenir sa racine (radical
    - Suppression du suffixe et du préfixe
  - Exemples :
    - souhaitons → souhait
    - *automate, automatisme, automatique* ⟶ *automat*
  - Contrairement au lemme, le stemme peut être un mot qui n'existe pas (« cherch » pour chercher, chercheur…)
  - Nécessite un étiqueteur morpho-syntaxique qui dépend de la langue

Au delà de la forme de surface

- Etiquette morphosyntaxique (POS)
  - Analyse de la fonction du mot dans la phrase
    - Ex : V2P pour « souhaitons », PONC « . »
    - !! Le jeu d'étiquettes dépend de l'outils utilisé
- Etiquette sémantique
  - Analyse du sens du mot dans la phrase
    - Ex : jour pour « lundi » ou ville pour « marseille »
- Etiquette de polarité
  - Analyse de la polarité du mot
    - Ex : négatif pour « méchant »
    - Associé à un poids?

Et encore d'autres phénomènes…

- Collocations
- Entités nommées
- Coréférences
- Stop-liste
- …

Prise en compte des cooccurrences

- Cooccurrence : Ensemble de mots en apparition fréquente dans une même fenêtre
  - Ex: Médecin et malade
- Collocation : Cooccurrence particulière
  - groupe de mots qui produit un sens nouveau par rapport au sens initial des mots qui le composent.
  - Par exemple:
    - pomme de terre
    - au fur et à mesure
    - passer son tour
- Normaliser en mots et groupes de mots?

Entités nommées

- Mots ou groupe de mots catégorisables en classes telles que : nom de personne, d'organisation, de lieu, fonction, dates, …
- Exemple
  - *François Fillon* "ne décrochera pas" de la vie politique après son échec dans la bataille à la présidence de l'*UMP*, ont déclaré, *mardi 20 novembre*, plusieurs de ses proches, dont *Laurent Wauquiez* et *Eric Ciotti*, à l'issue d'une réunion avec l'*ancien premier ministre à l'Assemblée*.

Les coréférences

- Phénomène qui permet de désigner le même objet par plusieurs expressions différentes
  - « **L'agriculteur** a dit qu'**il** aimait les pommes. »
  - « **Apple** sort un nouveau smartphone. **La firme à la pomme** se lache sur le prix! »
- Plusieurs types : anaphores, cataphores…
- Résolution de coréférences
  - Domaine de recherche à part entière
  - Pose de nombreux problèmes…

Sélection des mots informatifs?

- Idée : suppression des mots peu informatifs
  - Oui : indexation ou catégorisation de documents
  - Non : typage de texte
- Mots fréquents ou mots outils
  - Mot avec rôle syntaxique plus important que rôle sémantique
    - Ex : Article conjonctions, pronoms …

La stop-liste

- Liste de mots qui *a priori* n'apporte pas d'information
  - Préposition : de, sur…
  - Déterminant : le, une…
  - Pronom : je, nous…
  - Quelques adverbes et adjectifs : déjà, plusieurs…
  - Quelques noms et verbes : faire…
- Usage d'une stop-liste standard
  - Liste SMART, 571 mots anglais
  - Liste en français disponible : http://www.up.univ-mrs.fr/~veronis/data/antidico.txt
- Généralement, supprimer ces mots améliore les résultats

Représentation du texte

Dans les textes, quelles sont les données?

- Les données → les caractères
- Représentation en caractères?
  - Utilisées dans les applications:
    - Identification de la langue
      - Analyse de la fréquence de successions de caractères

Format de représentation

- Une fois « le dictionnaire » défini :
  - après nettoyage, normalisation
  - comprenant mots, classe d'équivalence, collocations, lemmes, étiquettes sémantiques…
- Sous quel « format » vont être présentées ces unités d'information aux méthodes d'apprentissage automatique?

Représentation du texte

- En pratique, comment représenter le corpus?
  - Corpus = Ensemble de documents
  - Une donnée, un individu = un document
  - Descripteurs?
    - mot ou étiquette de plus haut niveau
    - valeur globale au document
- Choisir une représentation vectorielle

RI : Représentation des textes en vecteur de mots

- Représentation des documents dans l'espace
- Espace vectoriel de dimension le vocabulaire choisi (+ descripteurs globaux ?)
  - Taille du vecteur :
    - Représentation en sac de mots
    - Vocabulaire = Ensemble des descripteurs
  - Valeur de la composante :
    - Présence/absence
    - Nombre d'occurrence
    - Pondération

Représentation en sac de mots

- L'ordre des mots n'est plus considéré
  - Perte de la séquentialité…
- Même représentation pour :
  - Le chat mange la souris
  - La souris mange le chat

Représentation en n-grammes

- Succession de n mots consécutifs
  - Ex Phrase *: « tous les mots que je dis dansent la farandole »
    - Unigramme : tous / les / mots / que / je / dis / dansent / la / farandole
    - Bigramme : tous les / les mots / mots que / que je / je dis / dis dansent/ dansent la / la farandole
    - Trigramme : tous les mots / les mots que / mots que je / que je dis / je dis dansent / dis dansent la / dansent la farandole
    - …

* *L'accent*, Miguel Zamacoïs

Résumé taille du vecteur

- Exemple :
  - vocabulaire=2 mots,
  - valeur=nb d'occurrences
    - Doc1 : je je suis , v1={2,1}
    - Doc2 : je je je je suis suis , v1={4,2}
    - Doc3 : je suis suis, v1={1,2}
- Ajout potentiel
  - de différents niveaux de représentation
  - de descripteurs globaux

Bien pondérer son vecteur

- Objectif
  - Donner un poids au mot qui permet de mesurer leur importance dans le document
  - → Donner un poids faible aux mots trop communs
  - → Donner un poids élevé aux mots plus spécifiques
- Calcul au sein d'un corpus de documents textuels
  - Déterminer les mots communs ou spécifiques
  - → La mesure tf.idf

Pondération par tf.idf

- Objectif : pondérer les mots suivant leur pertinence
- Calcul de deux mesures
  - tf : Fréquence du mot:
    - Fréquence d'apparition du mot i dans le document j
  - idf : Inverse de la fréquence dans le document :
    - Importance du mot dans le document inversement proportionnelle au nombre de documents dans lequel il apparaît
- Pondération du mot i dans le document j = $tf_{ij}.idf_i$

Fréquence des termes : tf

- Fréquence du terme
  - Ratio du nombre de fois où le mot i apparaît dans le document j sur le nombre total de mots du document j :

$$tf_{ij} = \frac{n_{ij}}{\|d_j\|}$$

- Information insuffisante :
  - Si le mot apparaît beaucoup de fois dans beaucoup de documents, est-ce plus intéressant que s'il apparaissait un peu moins mais que dans quelques documents?

Fréquence des documents inverse : idf

- Fréquence des documents
  - Ratio du nombre de document dans lequel le mot i apparaît (n_i) sur le nombre de documents (n)
- Inverse de la fréquence dans le document :
  - Importance du mot dans le document inversement proportionnelle au nombre de documents dans lequel il apparaît

$$idf_i = \log\left(\frac{n}{n_i}\right)$$

- Pondération du mot i dans le document j = $tf_{ij}.idf_i$

Résumé pondération du vecteur

- 3 possibilités :
  - Absence/Présence
  - Nombre d'occurrences
  - Pondération tf.idf
- Problèmes
  - Vecteurs très sparses
  - Vecteurs très grands : complexité de calcul
  - Pas de représentation du langage naturel (syntaxe ou sémantique…)

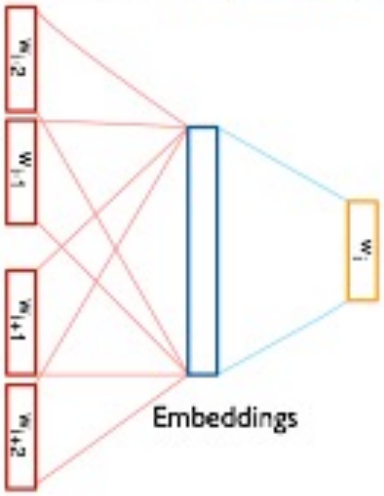
Les plongements de mots

- Idée :
  - Projeter un ensemble de mots dans un espace vectoriel pour y modéliser les relations syntaxiques ou sémantiques
- Objectif :
  - Obtenir un vecteur condensé
  - Chaque mot est représenté par un vecteur de nombres réels
- En anglais : word embeddings

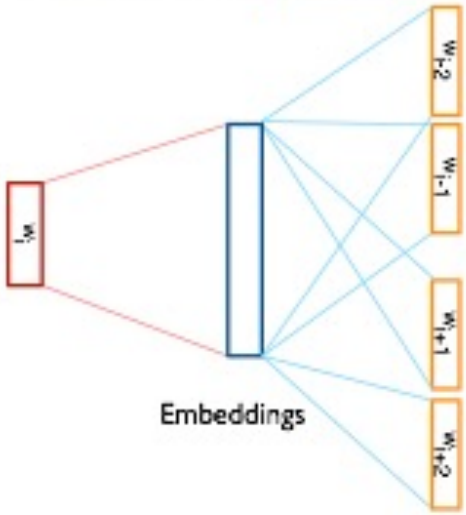
Word embeddings

Embeddings Linguistiques

CBOW [T. Mikolov *et al.* 2013]



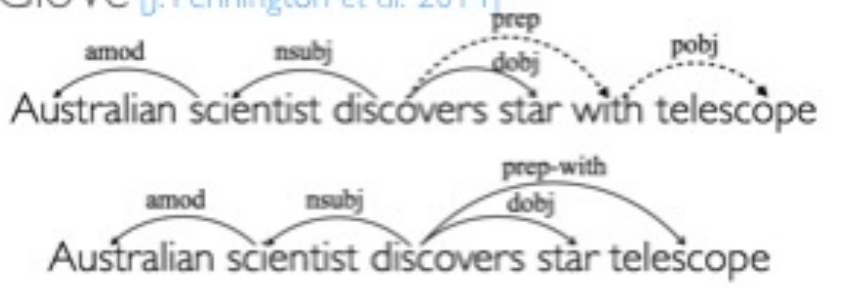
Skip-gram [T. Mikolov *et al.* 2013]



w2vf-deps [O. Levy *et al.* 2014]

- Calcul d'une matrice de co-occurence X
- Factorisation de X pour obtenir les word embeddings

GloVe [J. Pennington *et al.* 2014]



Parenthèse RI : Comparaison de documents

- Calcul de distance entre deux vecteurs
  - Documents similaires = vecteurs proches

    → directions quasi-identiques
  - Calcul de distances
    - Distance euclidienne, distance de Manhattan, indice de Jaccard, similarité cosine, …
- Exemple : vocabulaire=2 mots, valeur=nb d'occurrences
  - Doc1 : je je suis , v1={2,1}
  - Doc2 : je je je je suis suis , v1={4,2}
  - Doc3 : je suis suis, v1={1,2}

Parenthèse RI : Comparaison de documents

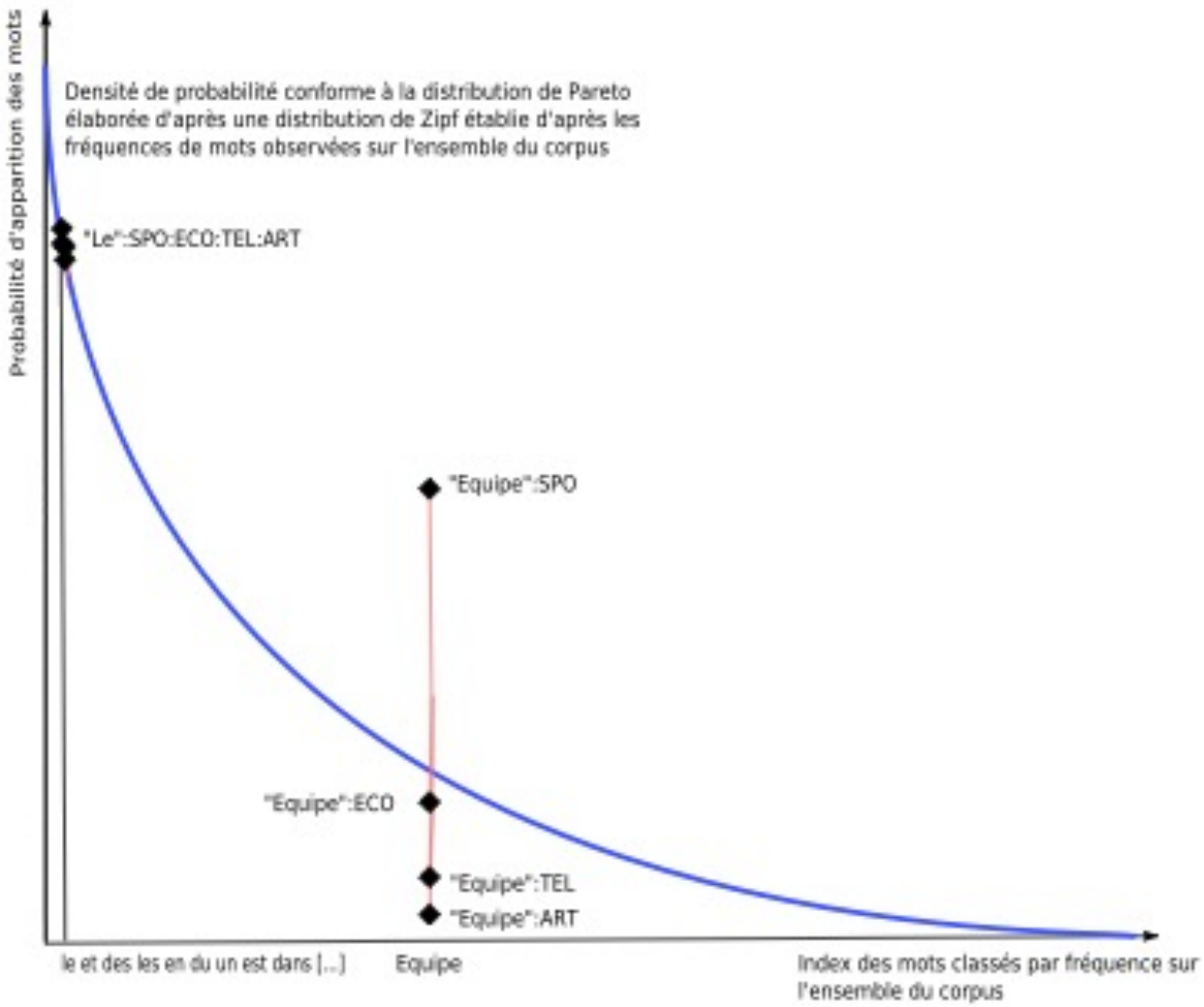
- Pour résumer
  - 2 documents contenant les mêmes mots, dans les mêmes proportions → similaires
  - 2 documents contenant les mêmes mots, dans des proportions différentes → dissemblables
  - 2 documents sans aucun mot commun

    → totalement dissemblables

La loi de Zipf

- Loi empirique découverte en 1936 par George K. Zipf (linguiste et philosophe)
- Classification des mots par fréquence décroissante
- Fréquence d'utilisation du mot inversement proportionnelle à son rang
  - Ex : le 10^{e} mot a une fréquence d'utilisation 10 fois moindre que le mot du 1er rang
  - Exemple : Le monde février 2007 : 1 286 915 mots
    - de (72395), la (40558), le (31647), les (26237), et (24676)…février (2851)… président(1484)… affairisme (1)

Loi de Zipf : exemple sur DEFT2008



Entre autres sources supplémentaires

- Cours de M. Grouin
  - http://perso.limsi.fr/grouin/inalco/1314/index.html
- Cours de M. Allauzen
  - http://perso.limsi.fr/Individu/amax/enseignement/ect/cours1-ECT.pdf
- Livre « Data Mining » de Tufféry (c.f. cours 1)