Quelques éléments de lexicométrie, analyse statistique de textes

Frédéric Béchet

Aix Marseille Université

Master M1

Lexicométrie, statistique linguistique

- Le langage est vu comme un processus pouvant être modélisé *statistiquement*
- Le *contenu* des textes n'est pas analysé d'un point de vue linguistique, mais uniquement sous l'angle de l'*analyse de données statistiques*
- Principes
  - Recherche de structures collectives dans le langage
  - Caractériser un texte en analysant les propriétés statistiques de ses éléments
  - Déterminer le *lexique* de chaque texte, trouver les lois statistiques qui régissent les fréquences des éléments du lexique

Aix*Marseille université

Lexicométrie, statistique linguistique

- Historique
  - Ouvrage de G.K. Zipf en 1949 : *Human Behavior and the Principle of Least Effort. Cambridge, Massachusetts : Addison-Wesley.*
    - la fréquence d'utilisation d'un mot est inversement proportionnelle à son rang, les mots d'un texte étant classés par fréquence décroissante
    - travaux connus depuis 1935 sous le nom de la fameuse loi de Zipf dont les applications dépasse l'application au seul langage
  - Développement des capacités de traitement des ordinateurs
    - Mise à disposition de *corpus* de plus en plus gros de données textuelles

Aix*Marseille université

Loi de Zipf

Pourquoi la fréquence d'un mot n'est pas un bon critère pour caractériser son importance ?
→ Loi de Zipf

Loi de Zipf

- Zipf et la **Principe du moindre effort**
  - George Kingsley Zipf (1902-1950)
  - Observations empiriques sur la fréquence des mots dans un texte
  - Formalisation sous la forme d'une loi par Jean-Baptiste Estoup
  - **Principe du moindre effort**
    - les mots les plus courts sont les plus fréquents
    - la fréquence d'un mot dépend de son rang tel que : $r \times f = constante$

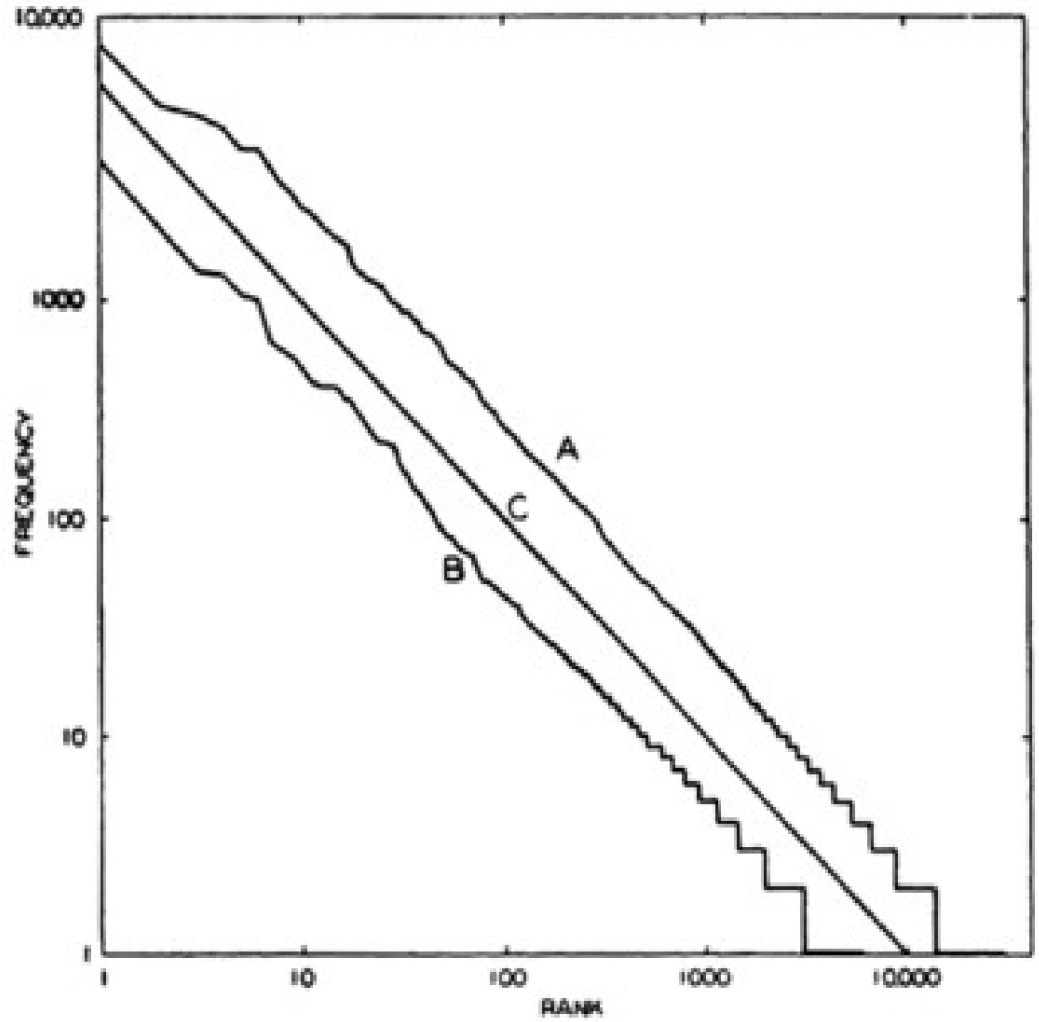
Aix*Marseille université

Loi de Zipf

Analyse de Ulysse, *de James Joyce*

rang r	fréquence f	$f \times r = c$
10	2 653	26 530
20	1 311	26 220
30	926	27 780
40	717	28 680
50	556	27 800
100	265	26 500
200	133	26 600
300	84	25 200
400	62	24 800
500	50	25 000
1 000	26	26 000
2 000	12	24 000
3 000	8	24 000
4 000	6	24 000
5 000	5	25 000
10 000	2	20 000
20 000	1	20 000
29 899	1	29 899

Loi de Zipf



Ce tableau, reproduit d'après Human Behavior and the Principle of least Effort, **montre la fréquence en fonction du rang des 298 000 mots de Ulysse, de James Joyce (courbe A) et de 43 900 mots de journaux quotidiens (courbe B). La ligne droite C illustre la loi de Zipf.**

(affichage des courbes en échelle *log*)

Utilisation de la lexicométrie pour l'analyse de documents électroniques

Comment représenter le langage de manière numérique ?

- Langage oral
  - langage = signal de parole → méthodes de traitement de signal
- Langage écrit
  - langage = texte → documents électroniques
  - texte = suites de caractères alphanumériques + symboles + formattage

Aix*Marseille université

Représentation d'un texte

- Choix de l'unité de base (**token**)
  - caractères, morphèmes, mots, syntagmes, phrases
- Modèle de représentation
  - **sac** de tokens
    - modèle booléen
    - modèle vectoriel
  - séquence de tokens générés par une grammaire
  - séquence de tokens générés par un processus statistique

Aix*Marseille université

Modélisation par sac de tokens

- Représentation "déstructurée" de textes
  - aucune notion de séquence
- Modèle booléen
  - présence (1) ou absence (0) d'un token dans un texte
  - représentation d'un texte sous une forme booléenne
    - $W_1 \& W_2 \& W_3 \ldots$
  - modèle simple (simpliste ! !) mais suffisant pour de nombreuses tâche de recherche d'information

Aix*Marseille université

Modélisation par sac de tokens

- Modèle vectoriel
  - A chaque mot w_i est associé un *poids*, relatif à l'importance de w_i dans le document
  - Un document est un vecteur dans un espace de grande dimension (taille du vocabulaire)
  - Chaque coordonnée correspond au degré d'importance d'un mot donné dans le texte

Aix*Marseille université

Pondération des mots d'un document

- Pondérations *intra-document*
  - Pour un lexique L et un document D
- booléenne
  - $\forall w_i \in L,\ \mathit{Bool}(w_i, D) = 1$ *if* $w_i \in D$ *otherwise* $\mathit{Bool}(w_i, D) = 0$
- nombre d'occurrences
  - $\forall w_i \in L,\ \mathit{Occ}(w_i, D) = C(w_i, D)$ avec $C(w_i, D)$ nombre d'occurences de w_i dans D
- fréquentielle normalisée (*Term Frequency - TF*)
  - $\forall w_i \in L,\ \mathit{TF}(w_i, D) = C(w_i, D) / \sum_j C(w_j, D)$

Aix*Marseille université

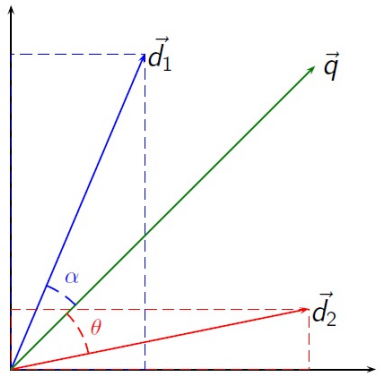
Pondération des mots d'un document

- Pondérations *intra+inter-document*
- Pour un lexique L, un document D, et un ensemble de documents $E = \{D_1, D_2, \ldots, D_M\}$
- Fréquence inverse de document (*Inverse Document Frequency - IDF*)
  - $\forall w_i \in L,\ IDF(w_i) = log(M / \sum_j\ bool(w_i, D_j))$
- *TF-IDF*
  - $TF\text{-}IDF(w_i, D) = TF(w_i, D) \times IDF(w_i)$
  - Critère assez ancien mais toujours pertinent
  - de nombreuses variantes (ex : Okapi BM25)

Aix*Marseille université

Utilisation de la représentation vectorielle

- Mesure de similarité entre documents ou entre un document et une requête
  - On mesure l'angle entre les vecteurs
  - Plus l'angle est petit entre deux documents D_1 et D_2, plus les distributions de termes sont proches, plus les documents sont semblables (au niveau des termes)



Utilisation de la représentation vectorielle

- Utilisation de la mesure *similarité cosinus* entre deux vecteurs
  - la similarité est mesurée par rapport à la mesure cosinus de l'angle
  - $\cos\theta$ est comprise dans l'intervalle $[-1, 1]$
    - la valeur -1 indiquera des vecteurs résolument opposés
    - 0 des vecteurs indépendants (orthogonaux)
    - 1 des vecteurs similaires (colinéaires de coefficient positif).
    - Les valeurs intermédiaires permettent d'évaluer le degré de similarité.

- La similarité cosinus entre 2 vecteurs s'obtient grâce au produit scalaire (*dot product*) et à leurs distance euclidienne

$$\cos(D_1, D_2) = \frac{D_1.D_2}{|D_1||D_2|}$$

avec $D_1.D_2 = \sum_i^V D_{1,i} \times D_{2,i}$ et $|D_1| = \sqrt{D_1 \times D_1}$

Aix*Marseille université

Applications

- Les mesures de similarité statistique de type *similarité cosinus* permettent de calculer une *ressemblance* entre deux documents
- Le domaine d'application principal est la recherche d'information ou recherche documentaire
  - Calculer une ressemblance entre une *requête* et un *document* $\rightarrow$ moteur de recherche sur internet
  - Classer un document : classification thématique
- Caractériser les propriétés statistiques d'un document
  - Etudes stylistiques
  - Détection de plagiat

Aix*Marseille université

Et le Traitement Automatique des Langues ?

- Est-ce que l'analyse statistique du texte fait partie du TAL ?
  - *question de point de vue ! !*
- Frontière nette entre les méthodes visant à faire apparaître des structures *linguistiques* dans le langage de manière explicite
  - Linguistique informatique
  - Analyse fondées sur des expertises et des modèles linguistiques pour *généraliser* l'étude d'un document
- Frontière flou avec les méthodes empiriques basées sur la découverte de structures *implicites* dans le langage pour réaliser une tâche
  - Méthodes basées uniquement sur la *performance* par rapport à une tâche
  - Paradgime du *end-to-end* (*bout-en-bout*)

Aix*Marseille université