Natural Language Processing (NLP)

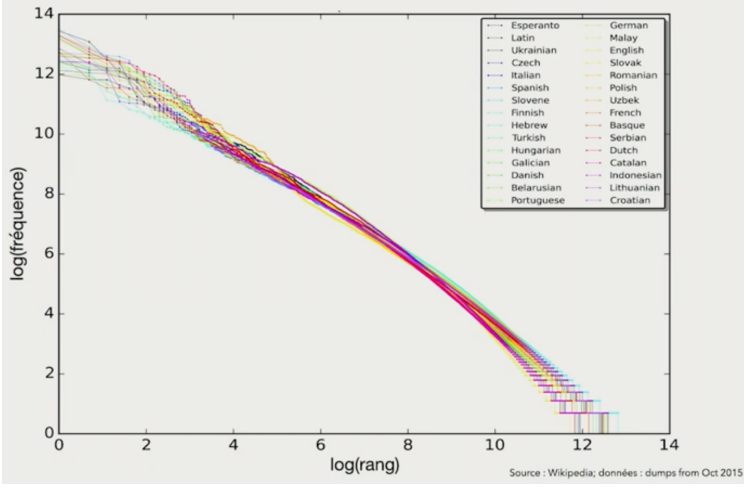
5 — Représenter les tokens

Georges-André Silber

2023/2024

École des mines de Paris

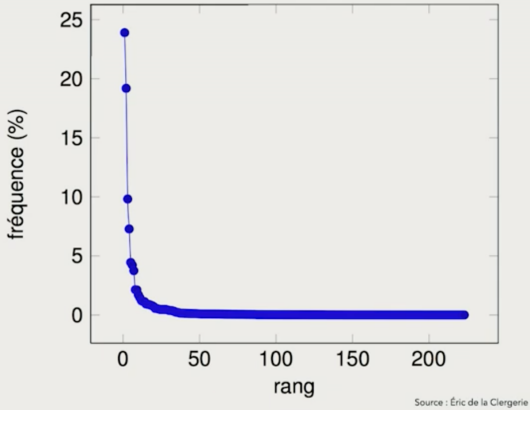
Loi de Zipf sur les mots d'un corpus



Rang / fréquence pour les 10 premiers millions de mots de 30 Wikipedia.

Source : cours de B. Sagot 2023

Loi de Zipf sur les phrases d'un corpus



Fréquence de constructions syntaxiques dans un corpus de 10 000 phrases analysées automatiquement.

Source : cours de B. Sagot 2023

Représenter les propriétés des tokens par des vecteurs

- Chaque token, dans le contexte où il apparaît, porte des propriétés (morpholoqiques, syntaxiques, sémantiques)
- On représente généralement les tokens par des vecteurs :
  - représentation utilisée depuis longtemps par les réseaux de neurones
  - permet également de tenter d'encoder la sémantique dans un espace vectoriel

Représentation *One hot*

1	a
2	à
3	abaca
4	abacas
5	abacule
6	abacules
7	abaissa
8	abaissable
9	abaissables
10	abaissai
11	abaissaient
12	abaissais
...	
29 000	zythum
30 000	zythums

	1	2	3	4	5	6	7	8	...	29 000	30 000
abaissa	0	0	0	0	0	0	1	0	...	0	0

- variables catégorielles
- taille du vecteur = nombre de tokens dans le modèle
- chaque token est représenté par un vecteur de 0 où une seule composante est à 1
- pas de notion de proximité

Source : cours de B. Sagot 2023

Hypothèse distributionnelle

- Le contexte donne des informations sur un token.
- Exemple [tiré de Nida 75, Lin 78, Sagot 23] :

  Il y a une bouteille de *tesgüino* sur la table.
  Tout le monde aime le *tesgüino*.
  Le *tesgüino* rend ivre.
  On produit le *tesgüino* à partir de maïs.
- Hypothèse : deux tokens sont similaires s'ils apparaissent dans un même contexte
- Firth (1957) : *you shall know a word by the company it keeps*.

Embeddings non contextuels

- *Word embedding* : plongement lexical
- Représentation vectorielle des tokens
- Étant donné un mot on lui assigne une représentation vectorielle unique sur la base de toutes ses apparitions dans un grand corpus
- Approches par comptage ou statistiques
- Approches prédictives par modèle neuronal
- Voir Embedding projector

Approche par comptage

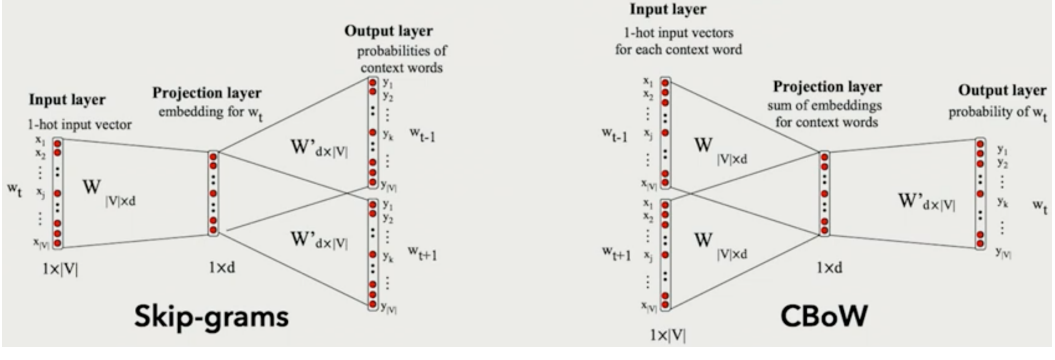
	autruche	ordinateur	donnée	salade	résultat	sucre	...	...
abricot	0	0	0	1	0	1		
ananas	0	0	0	1	0	1		
numérique	0	2	1	0	1	0		
information	0	1	6	0	4	0		

	As you like it	Twelfth night	Julius Caesar	Henry V	...	...
battle	1	0	7	13		
good	114	80	62	89		
fool	36	58	1	4		
wit	20	15	2	3		

- Matrices de co-occurrence mot/mot et mot/document.
- Vecteurs ”creux” et de grande dimension
- Mots rares et fréquents pèsent autant (1 dimension)
- Solutions : TF-IDF, Latent Semantic Analysis, Latent Dirichlet Allocation

Source : cours de B. Sagot 2023, adapté du cours de Jurafsky et Martin

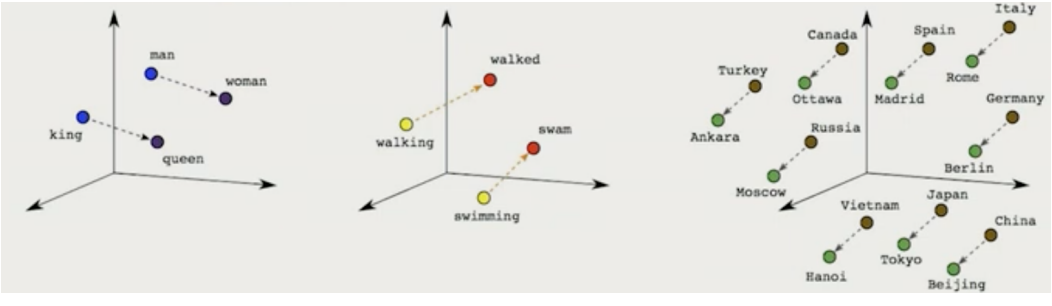
Approche prédictive neuronale



- Word2vec, fastText, GloVe
- Vecteurs One-hot
- La couche cachée est lue comme un word embedding

Source : cours de B. Sagot 2023

Structuration de l'espace



- Succès de word2vec : structuration de l'espace
- Calculs par analogie : Paris - France + Italy ≡ Rome

Source : cours de B. Sagot 2023, d'après Irina Sigler

Limite des embeddings non contextuels

Portrait de Frédéric Thomas (avocat, littérateur, journaliste). Musée Carnavalet

Pierre est un excellent avocat

Source : wiktionary

Pierre mange un excellent avocat

- ”avocat” est représenté par le même vecteur quel que soit son contexte
- Les modèles de langues et leurs *embeddings contextuels* permettent de lever cette limitation

Source : cours de B. Sagot 2023

Retour sur TF-IDF : moteurs de recherche

- Problème : pour une recherche q, dans quel ordre renvoyer les documents ?
- TF-IDF

$$\text{tfidf}(t, d, D) = \text{tf}(t, d) \cdot \log \frac{N}{1 - |\{d \in D : t \in d\}|}$$

- Okapi BM25 (sacs de mots)

$$\text{score}(D, Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}$$

$$\text{IDF}(q_i) = \ln\left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1\right)$$

- Démonstration d'indexation du JORF avec Solr
- Relevant Search