Programmation de modèles linguistiques (II)

L6SOPROG L3 SDL

Alice Millour

STIH EA 4509, Sorbonne Université Sorbonne Université

La séance d'aujourd'hui

- Quelques rappels sur la sémantique
- Présentation des word embeddings
- Démo TF/IDF pour ceux et celles qui en ont besoin (TP 3 pour les autres)
- TP 3 pour tout le monde (à rendre)

Comment modéliser le sens ?

Hypothèse linguistique

"les mots qui apparaissent dans des contextes similaires tendent à avoir des sens proches" (Harris, 1954)

"A man is known by the company that he keeps" — Ésope

Hypothèse linguistique

"les mots qui apparaissent dans des contextes similaires tendent à avoir des sens proches" (Harris, 1954)

"A man is known by the company that he keeps" — Ésope

Modéliser le sens = modéliser **les liens** entre les mots

Donc : pour modéliser un mot, j'ai besoin des autres mots...

On tourne en rond ?

Donc : pour modéliser un mot, j'ai besoin des autres mots...

On tourne en rond ?

Non ! On va utiliser les notions de

- co-occurrence
- distribution

pour modéliser les sens *relatifs* des mots les uns par rapport aux autres.

TP : construction de représentation vectorielle de mots

Pour ce cours, les représentations vectorielles sont déjà construites.

- le vecteur d'un mot peut être de taille variable (exemple de la polarité)
- le vecteur d'un mot dépend du texte à partir duquel il a été calculé

À partir des statistiques de co-occurrence, de position (relative et absolue) dans la phrase, on répartit les mots dans "l'espace du sens"

TP : manipulation de représentation vectorielle de mots

- pour l'instant, on a comparé des représentation vectorielles de **textes**
- pour avoir une représentation plus *fine*, on va maintenant utiliser des représentations vectorielles de **mots**

ouvrir le fichier
`frWac_no_postag_no_phrase_500_skip_cut100.txt.extrait` et
observer les vecteurs (faire une recherche du mot "pour" par exemple)

Retour sur TF/IDF ou TP

Les mesures de similarité

Objectif : calculer la similarité entre deux textes

“À quel point ces deux textes se ressemblent-ils ?”

“À quel point parlent-ils de la même chose ?”

“À quel point décrivent-ils des objets semblables ?”

Les mesures de similarité

Objectif : calculer la similarité **entre deux textes**

TF/IDF permet d'obtenir une représentation **vectorielle** des textes ce qui les rend plus facile à comparer avec des métriques classiques (du type *similarité cosinus*)

Autre méthode : utiliser TF/IDF

exemple avec trois phrases :

“0 = Simple example with Cats and Mouse ”

“1 = Another simple example with dogs and cats "

”2 = Another simple example with mouse and cheese ”

Pour chaque phrase, on calcule TF et IDF pour **l’ensemble des mots du corpus**

$\Rightarrow$ on obtient des vecteurs de même taille pour toutes les phrases

	and	another	cats	cheese	dogs	example	mouse	simple	with
0	1	0	1	0	0	1	1	1	1
1	1	1	1	0	1	1	0	1	1
2	1	1	0	1	0	1	1	1	1

	and	another	cats	cheese	dogs	example	mouse	simple	with
0	0.0	0.000000	0.067578	0.000000	0.000000	0.0	0.067578	0.0	0.0
1	0.0	0.057924	0.057924	0.000000	0.156945	0.0	0.000000	0.0	0.0
2	0.0	0.057924	0.000000	0.156945	0.000000	0.0	0.057924	0.0	0.0

References i