

Programmation de modèles linguistiques (II)

L6SOPROG L3 SDL

Alice Millour

STIH EA 4509, Sorbonne Université Sorbonne Université

La séance d'aujourd'hui

- Contrôle de connaissances
- Retour sur Jaccard et TF/IDF
- introduction à la représentation sémantique implicite
- Fin du TD commencé avec V. Lully en autonomie (**À rendre**)

Contrôle de connaissances

- Quelle est la différence entre `print()` et `return`? donnez deux exemples de fonctions
- À quoi servent les commentaires dans le code? Quand faut-il les écrire?
- À quoi sert d'avoir des données de test?

Écrire un programme (sur papier) qui prend en entrée une liste d'animaux [*animal1*, *animal2*, ...] et une liste de personnes [*personne1*, *personne2*, ...] et qui crée un dictionnaire

animaux_preferes = {'*personne1*' : '*animal1*', '*personne2*' : '*animal2*'}.

Donnez un exemple de données de test et de résultat attendu afin de tester votre programme.

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 ?”

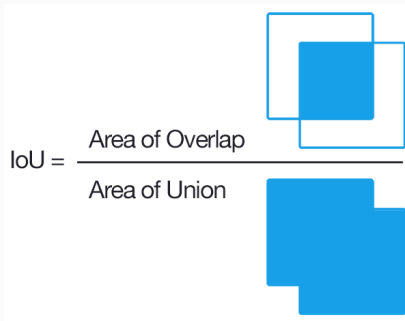
“À quel point décrivent-ils des objets semblables ?”

- Mesure de Jaccard
- Similarité cosinus

Quelle différence entre ces deux mesures ?

Similarité de Jaccard

Comment ça marche ?



By Adrian Rosebrock - <http://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/>,CCBY-SA4.0,<https://commons.wikimedia.org/w/index.php?curid=57718560>

Similarité de Jaccard - exemples

Exemple : calcul de la similarité entre ces trois textes :

texte_1 = "Lundi j'ai cours de bases de données, mardi de programmation, jeudi de sémantique : je n'ai plus le temps pour faire du sport."

texte_2 = "Qu'est-ce que la programmation? La programmation est une discipline qui demande temps concentration et vigilance."

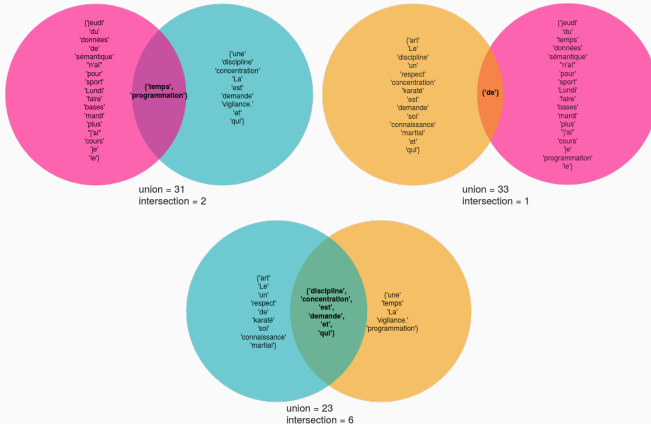
texte_3 = "Le karaté est un art martial qui demande respect discipline concentration et connaissance de soi."

Similarité de Jaccard - exemples

texte_1 = "Lundi j'ai cours de bases de données, mardi de programmation, jeudi de sémantique : je n'ai plus le temps pour faire du sport."

texte_2 = "Qu'est-ce que la programmation ? La programmation est une discipline qui demande temps concentration et vigilance."

texte_3 = "Le karaté est un art martial qui demande respect discipline concentration et connaissance de soi."



Similarité de Jaccard - exemples

texte_1 = "Lundi j'ai cours de bases de données, mardi de programmation, jeudi de sémantique : je n'ai plus le temps pour faire du sport."

texte_2 = "Qu'est-ce que la programmation ? La programmation est une discipline qui demande temps concentration et vigilance."

texte_3 = "Le karaté est un art martial qui demande respect discipline concentration et connaissance de soi."

$$\text{similarite_Jaccard}(\text{texte_1}, \text{texte_2}) = \frac{2}{28} = 0.07$$

$$\text{similarite_Jaccard}(\text{texte_1}, \text{texte_3}) = \frac{1}{33} = 0.03$$

$$\text{similarite_Jaccard}(\text{texte_2}, \text{texte_3}) = \frac{6}{20} = 0.30$$

Similarité de Jaccard - exemples

texte_1 = "Lundi j'ai cours de bases de données, mardi de programmation, jeudi de sémantique : je n'ai plus le temps pour faire du sport."

texte_2 = "Qu'est-ce que la programmation ? La programmation est une discipline qui demande temps concentration et vigilance."

texte_3 = "Le karaté est un art martial qui demande respect discipline concentration et connaissance de soi."

$$\text{similarite_Jaccard}(\text{texte_1}, \text{texte_2}) = \frac{2}{28} = 0.07$$

$$\text{similarite_Jaccard}(\text{texte_1}, \text{texte_3}) = \frac{1}{33} = 0.03$$

$$\text{similarite_Jaccard}(\text{texte_2}, \text{texte_3}) = \frac{6}{20} = 0.30$$

conclusion ?

Autre méthode : utiliser TF/IDF

Hypothèses linguistiques sous-jacente :

- Tous les mots n'ont pas la même importance dans un texte
- Les mots présents dans tous les textes ne permettent pas de classer les textes par groupe thématique

ex : classer des personnalités par biographies

- quels sont les mots qui vont avoir un TF fort ?
- quels sont les mots qui vont avoir un IDF faible ?

Cours

Modélisation sémantique

Hypothèse linguistique

C'est ce qu'on **observe** en tant qu'**humain**.

Observation : *en Italien, les mots finissent plus souvent par "a" ou par "o" qu'en français*

Hypothèse : *Les proportions de terminaison des mots permet de distinguer l'italien du français*

implémentation

On *modélise* l'observation pour qu'elle puisse être traitée par un **programme**.

Dans cet exemple, on *compte* :

1. Dans un corpus de l'italien, 196 mots (9%) finissent par a, 98 mots (4,5%) finissent par o.
2. Dans un corpus du français, 24 mots (1%) finissent par a, 3 mots (0,2%) finissent par o.

implémentation

On **modélise** l'observation pour qu'elle puisse être traitée par un **programme**.

Dans cet exemple, on **compte** :

1. Dans un corpus de l'italien, 196 mots (9%) finissent par a, 98 mots (4,5%) finissent par o.
2. Dans un corpus du français, 24 mots (1%) finissent par a, 3 mots (0,2%) finissent par o.

Attention ! pour que la modélisation soit correcte, il y a des conditions !

implémentation

On **modélise** l'observation pour qu'elle puisse être traitée par un **programme**.

Dans cet exemple, on **compte** :

1. Dans un corpus de l'italien, 196 mots (9%) finissent par a, 98 mots (4,5%) finissent par o.
2. Dans un corpus du français, 24 mots (1%) finissent par a, 3 mots (0,2%) finissent par o.

Conclusion de la modélisation : *en Italien, les mots finissent 9 fois plus "a" et 22 fois plus par "o" qu'en français*

L'hypothèse est **validée** par le protocole scientifique !

D'autres exemples ?

Modélisation “pauvre” : On se contente de

découper, filtrer, compter

des mots

Ce n'est pas suffisant !

Imaginez un moteur de recherche qui ne fonctionne que par mots-clé...

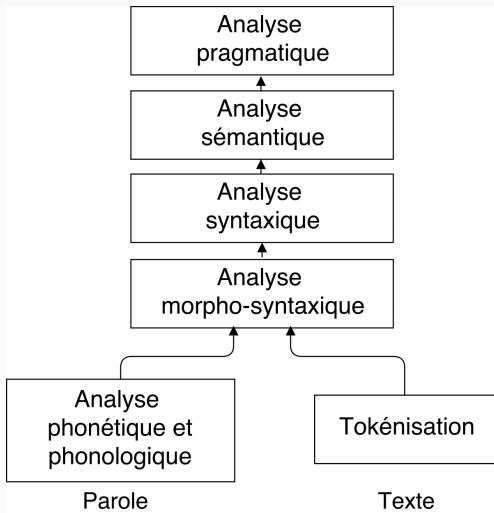
On se contente de découper, filtrer, compter des mots

Ce n'est pas suffisant !

- ambiguïté (homonymie, polysémie)
- on ne modélise pas les liens entre les mots (*boulangerie* vs *pâtisserie*) ?

Le sens... difficile à saisir

Les niveaux d'analyse linguistique :



Comment modéliser le sens ?

Hypothèse linguistique

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

Hypothèse linguistique

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

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.

Quels sont les liens sémantiques qui existent entre les mots ?

Un peu de sémantique lexicale

- hypéronymie et hyponymie : *tulipe* est un hyponyme de *fleur*
- partie-tout : *main* est une partie de *bras* qui est une partie de *corps*
- synonymie : *docteur* / *médecin*
- antonymie : *mort* / *vivant* // *étudiant* / *professeur*
- polysémie : *souris* = *souris* (ordinateur) et *souris* (animal)
- **homonymie** = une même forme a plusieurs sens : une *porte* / je *porte*

Que pensez-vous des relations de synonymie et d'antonymie ?

Objectif de la représentation sémantique :

- texte qui parle d'un **mulot** \simeq texte qui parle de l'**animal souris** \neq **souris d'ordinateur**
- texte qui parle de **médecin** \simeq texte qui parle de **docteur** \simeq texte qui parle de **patient**

`http://live.babelnet.org/`

Fin du TD commencé avec M. Lully

