DM2 Sémantique Computationnelle

Cong Jinyu

4 novembre 2024

1 Introduction

Les matrices de cooccurrence sont des outils couramment utilisés en traitement automatique du langage pour modéliser les relations sémantiques entre les mots. Elles permettent de représenter le contexte d'un mot dans un corpus en comptabilisant les occurrences conjointes de ce mot avec d'autres mots voisins dans une fenêtre définie. En mesurant la fréquence de cooccurrence, on peut ainsi déterminer la similarité sémantique entre différents mots. Une mesure courante pour quantifier cette similarité est la similarité cosinus, qui permet de comparer les vecteurs de contexte de chaque mot.

Dans le cadre de ce travail, j'ai choisi un article en français et j'ai procédé à l'analyse des cooccurrences des mot autour d'autres mots. Pour ce faire, j'ai défini une fenêtre de taille fixe autour de chaque mot et j'ai compté la fréquence d'apparition des mots dans ces fenêtres. Ensuite, j'ai organisé les données obtenues sous forme de tableau, afin de structurer les relations de cooccurrence entre les mots du corpus. Enfin, j'ai calculé la similarité cosinus entre quelques mots sélectionnés pour déterminer leur degré de similarité sémantique en fonction des contextes observés.

Ce travail permet de visualiser comment les mots, lorsqu'ils partagent des contextes similaires, peuvent être considérés comme sémantiquement proches grâce aux valeurs de similarité obtenues.

2 Explication du code python

Tout d'abord, j'ai défini une fonction context_count pour calculer le nombre d'occurrences de chaque mot apparaissant dans les fenêtres de chaque mot de l'article entier. Les résultats sont stockés dans un dictionnaire dont les clés sont les mots de l'article, et les valeurs sont des dictionnaires. Chaque dictionnaire interne contient le nombre d'occurrences des mots entourant le mot clé. Enfin, j'ai converti ce dictionnaire en un DataFrame avec pandas et l'ai utilisé comme sortie de cette fonction. La taille de la fenêtre est également changeable ; par exemple, j'ai choisi des fenêtres de taille 4 et 6. Les similarités cosinus entre les mots varient en fonction de la taille de la fenêtre sélectionnée. J'ai mis à la fin les tableaux générés pour ces deux tailles de fenêtre, afin de comparer les similarités cosinus obtenues dans chaque cas.

Ensuite, j'ai écrit une fonction appelée generate_candidate_words qui génère une liste contenant les mots et leurs vecteurs de contexte. Le vecteur de chaque mot correspond à la fréquence des mots dans son contexte, extraite du DataFrame précédent. Cette fonction permet de définir le nombre de mots que l'on souhaite analyser. Il suffit d'ajouter en entrée tous les mots pour lesquels on souhaite calculer la similarité cosinus.

Enfin, grâce aux mots extraits précédemment, il est possible de calculer la similarité cosinus entre leurs vecteurs en utilisant la fonction cosine_similarity de la bibliothèque sklearn. Le résultat est également organisé sous forme de DataFrame, permettant de visualiser facilement la similarité cosinus entre tous les mots sélectionnés.

3 Prétraitement de donnée

J'ai utilisé le module spacy pour éliminer de l'article tout ce qui n'est pas un mot, comme la ponctuation, et pour convertir tous les mots en leur lemme. Cela permet une meilleure comptabilisation de chaque occurrence de mot, afin de rendre les données plus homogènes et de rendre la similarité cosinus finale plus représentative. J'ai également retiré tous les mots vides, car ils sont nombreux et

influencent largement le vecteur de chaque mot, ce qui pourrait diminuer la pertinence de la similarité calculée.

4 Analyse des données

Comme s'affiche les deux tableaux de la dernière page, j'ai sélectionné cinq mots : *fiscal, financier, budget, jardin, démissionner*. Pour chacun de ces mots, j'ai compté le nombre de mots apparaissant dans leurs fenêtres respectives, ce qui m'a permis de créer cinq vecteurs pour chaque mot. Ensuite, j'ai calculé la similarité cosinus entre chaque paire de ces mots, et les résultats sont présentés dans les tableaux.

4.1 Analyse de similarité entre les mots

Si on regarde seulement le tableau de similarité entre les mots avec la fenêtre de taille 4, il n'est pas difficile de trouver que les similarité entre les paires de mots *fiscal, financier, budget* sont beaucoup haute que les autres paires de mots, comme par exemple *jardin* et *fiscal* (Sim(fiscal, jardin) = 0). Pour un jeu de donnée assez grand comme j'ai choisi (à peu près 9200 mots différents), la similarité dépasse 0.15 est assez significative. Ce résultat est tout à fait correspond à l"hypothèse distributionnelle, qui montre que pour les mots similaires, ou dans la même champ sémantique, leurs contexte (les mots se trouve dans leurs fenêtres) sont aussi similaires. Dans cette expérience, les mots financier, fiscal et budget ont un lien commun pour dire de l'argent, par contre les mots fiscal et jardin ont peu de similarité en commun dans un point de vue sémantique.

4.2 L'influence de la taille de fenêtre

Pour comparer le résultat de l'influence de la taille de fenêtre, j'ai choisi deux différente taille de fenêtre, taille 4 et 6. Comme montre le résultat dans la dernière page, les similarités avec fenêtre de taille 6 sont plus hautes que celles de taille 4, ce qui rend le résultat plus significative. Parce que les mots similaires, comme fiscal et financier, leur similarité a monté de 0.09, par contre, les mots peu similaires comme fiscal et jardin, leur similarité a monté aussi mais avec une valeur beaucoup moins élevé, juste 0.008. Donc nous pouvons conclure que, la taille de la fenêtre plus large rend le résultat plus significative, mais il faut la contrôler dans une intervalle approprié, c'est évidement que la taille 100 ne marche pas.

5 Conclusion

L'expérience montre que les mots sématiquement similaires partagent aussi une contexte similaire. La taille de fenêtre influence beaucoup le résultat obtenu par le calcul de similarité cosinus entre les mots.

	fiscal	financier	budget	jardin	démissionner
fiscal	1.000000	0.260011	0.151894	0.000000	0.032940
financier	0.260011	1.000000	0.124309	0.000000	0.082223
budget	0.151894	0.124309	1.000000	0.062994	0.062994
jardin	0.006588	0.000000	0.062994	1.000000	0.133333
démissionner	0.032940	0.082223	0.062994	0.133333	1.000000

TABLE 1 – Les similarités de cosinus entre les mots avec une fenêtre de taille 4

	fiscal	financier	budget	jardin	démissionner
fiscal	1.000000	0.347816	0.186287	0.008133	0.104638
financier	0.347816	1.000000	0.220175	0.021574	0.093953
budget	0.186287	0.220175	1.000000	0.054317	0.092161
jardin	0.008133	0.021574	0.054317	1.000000	0.082479
démissionner	0.104638	0.093953	0.092161	0.082479	1.000000

TABLE 2 – Les similarités de cosinus entre les mots avec une fenêtre de taille 6