La vectorisation des textes : TF-IDF

La séance précédente, nous avons vu l'approche de vectorisation de texte BOW, une approche simple et peu efficace. TF-IDF (Term Frequency-Inverse Document Frequency) est une autre approche plus forte que BOW.

En effet, il existe de nombreux mots courants comme « le », « est », « je » qui apparaissent fréquemment dans les phrases, mais ne contribuent pas de manière significative à apporter des informations. Si nous ne regardions que le terme fréquence, ces mots apparaîtraient plus importants que les autres mots. Pour cette raison, TF-IDF est introduit pour améliorer la représentation vectorielle des textes.

I. Principe de l'approche TF-IDF:

1. La notation standard de TF-IDF :

La formule de calcul de TF-IDF est définie comme suite :

TF-IDF(t, d) = TF(t, d) x IDF(t)

Où Term Frequency (TF) est la fréquence du mot ***t*** dans le document ***d***. En d'autres termes, c'est le rapport entre le nombre de mots dans le document et le nombre total de mots :

TF(t, d) = Le nombre d'occurrence de t dans d / nombre des mots dans d

Comme nous l'avons dit précédemment, la fréquence des mots n'est pas suffisante pour fournir des mesures efficaces. Nous devons également la combiner avec une autre valeur, appelé Inverse Document Frequency. C'est une transformation logarithmique d'une fraction, calculée comme le nombre total de documents dans le corpus divisé par le nombre de documents contenant le mot.

IDF(t) = log (nombre de documents / nombre de documents contenants le mot t)

2. La notation TF-IDF de Sklearn :

IDF(t) = log (1 + n / 1 + df(t)) + 1

Avec df(t) est nombre de documents contenants le mot **t**

Vous pouvez facilement observer que la version sklearn d'IDF ajoute 1 au numérateur et au dénominateur pour éviter les divisions nulles. De plus, une constante égale à 1 est ajoutée au terme du logarithme.

Par exemple, si on a **n=3** documents et **df(t)=3**, ce qui implique que le mot apparaît dans tous les documents, l'IDF(t) est égal à **log((1+3)/(1+3))+1 = 1** suivant la définition Scikit-learn, tandis que **IDF(t) = log(3/3) = 0** dans le cas standard.

La formule générale de calcule ne change pas : ***TF-IDF(t, d) = TF(t, d) x IDF(t)***

II. Vectorisation TF-IDF avec Sklearn :

```
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer()
response = vectorizer.fit_transform([liste de vos documents])
df_tfidf_sklearn = pd.DataFrame(response.toarray(),columns=vectorizer.get_feature_names())
df_tfidf_sklearn
```

III. Travaux pratiques :

1. Vectoriser les textes d’avis de la séance précédente avec la méthode TF-IDF en utilisant la bibliothèque Sklearn.
2. Refaire le même travail cette fois-ci avec Python pure.