Gestion des données et accès à l'information

Représentation du texte

Olivier Schwander <olivier.schwander@sorbonne-universite.fr>

2023-2024

Exercice 1 - *Modèle des mots*

On va utiliser la bibliothèque `Gensim` qui fournit de nombreux outils relatifs au pré-traitements et aux représentations du texte. On y trouve notamment des architectures pré-entraînées pour les modèles de mots.

Les instructions d'installation sont disponibles ici https://radimrehurek.com/gensim/.

On télécharge immédiatement un modèle pré-entraîné:

```
import gensim.downloader
model = gensim.downloader.load('word2vec-google-news-300')
```

Il s'agit un modèle de mot, de type *word2vec*: à chaque mot du vocabulaire est associé un point dans un espace vectoriel:

```
model["queen"]
```

Question 1

Observez la représentation ainsi obtenue. Quelle est la dimension du vecteur représentant un mot ?

Question 2

On peut comparer deux représentations avec:

```
model.similarity("queen", "king")
```

Cette similirité utilise le cosinus entre les vecteurs.

Comparez les similarités entres quelques mots. Par exemple

- queen
- king
- woman
- man

Question 3

On peut avoir un aperçu du vocabulaire utilisé dans le modèle de la façon suivante:

```
model.key_to_index
```

Combien y a-t-il de mots dans le vocabulaire ?

Question 4

Avec `most_similar` on peut rechercher les mots les plus similaires à un autre dans tout le vocabulaire.

Commentez la sortie de

```
model.most_similar('king')
```

Question 5

Puisque nous utilisons un espace vectoriel, on peut faire de l'arithmétique entre les représentations.

À quoi pourrait correspondre l'expression "king" - "man" + "queen" ?

Question 6

Pour des raisons d'implémentation, on utilisera plutôt le code suivant:

```
model.most_similar(positive=["king", "woman"], negative=["man"])
```

Obtient-on le résultat espéré ?

Question 7

Dans les deux phrases suivantes, quelle sera la représentation du mot *bank*:

- "We went to the river bank."
- "I will go to the bank to make a deposit."

Question 8

Comment représenter un mot inconnu ? (*out of vocabulary*)

Question 9

Une façon possible de représenter une phrase est de moyenner tous les mots qu'elle contient puis de normaliser le vecteur obtenu.

Écrivez une fonction calculant la représentation d'une phrase.

Question 10

Écrivez une fonction calculant la similarité entre deux phrases.

Question 11

À partir du dataset suivant, affichez la matrice de similarité de ces phrases.

```
data = [
    'the road is straight',
    'the black cat plays with a ball',
    'a big dog with a ball',
    'dog and cat are together',
    'traffic jam on the 6th road',
    'white bird on a big tree',
    'a big truck',
    'two cars crashed',
    'two deers in a field',
    'I like ridding my bike',
    'a lion in the savane',
    'a motorcycle rides on the road',
```

```
    'a mouse bitten by a cat',
    'two pigs in the mood',
    'take a plane is sometimes slower than taking train',
    'take the highway'
]
```

Exercice 2 - *Modèles de langue*

En terme d'implémantation la référence est la bibliothèque `transformers` dévelopée par Huggingface.

Question 12

Voir cette page pour un tutoriel https://huggingface.co/learn/nlp-course/chapter2/1.

Exercice 3 - *Sacs de mots et BM25*

On va de nouveau utiliser `Gensim` et son implémentation des sacs de mots.

Question 13

Voir cette page pour l'utilisation de la représentation TF-IDF https://radimrehurek.com/gensim/models/tfidfmodel.html

Question 14

Construisez un mécanisme de recherche d'information complet, du prétraitement du texte à la rehcerhce en elle-même.