

Lundi 17 mai 2021

**Exercice 1 : Application simple (2pt)**

Soit un corpus de documents dont le vocabulaire est composé de 8 mots numérotés de A à H. Les documents de ce corpus sont les suivants :

- D1 = AAA
- D2 = BBB
- D3 = AABDE
- D4 = ABDFGH
- D5 = C
- D6 = B
- D7 = H
- D8 = DDDGGGH

Soit la requête  $Q=ABC$ .

**Question 1**

Rappeler à quoi correspond l'IDF d'un terme. Donner une formule simplifiée de l'IDF et le calculer pour les mots A, B, C.

**Question 2**

Calculer le score de pertinence de chacun des documents du corpus pour la requête Q selon le modèle vectoriel basé sur un produit scalaire. On considérera une pondération TF pour les termes des documents et une pondération IDF pour les termes de la requête.

**Question 3**

Sachant que les documents pertinents sont les documents D1, D3 et D4, calculer la précision, le rappel, le MRR et le NDCG au rang 5 de l'ordonnancement retourné en question 2.

Aide :

$$DCG_p = rel_1 + \sum_{i=2}^p \frac{rel_i}{\log_2(i)} \quad (1)$$

$$nDCG_p = \frac{DCG_p}{IDCG_p} \quad (2)$$

**Exercice 2 : loi de Zipf (2pt)**

Une des expressions formelles de la loi de Zipf est la suivante :

$$\text{frequence} = \frac{\lambda}{\text{rang}} (\text{avec } \lambda > 0) \quad (3)$$

**Question 1**

Quelle est l'intuition de la loi de Zipf? À quoi sert-elle?

**Question 2**

On suppose que le 50<sup>e</sup> mot le plus fréquent a une probabilité d'apparition de 0.02 dans une collection de 10 000 mots. Quel est le rang d'un mot qui apparaît 40 fois dans la collection?

**Question 3**

Si on considère que tous les mots d'une collection ont des fréquences différentes, quel est le nombre total d'occurrences si on considère  $\lambda = 36\,000\,000$ ? Indication : Posez juste la formule, sans faire de calculs si trop compliqué.

**Exercice 3 : Modèle de RI et vraisemblance (2pt)**

On considère une collection de documents  $D = \{d_1, \dots, d_N\}$  et une requête  $q$ . Pour chaque couple  $\{d, q\}$  on dispose d'un jugement de pertinence binaire  $R_d$ . Un document  $d_j$  est représenté par un vecteur binaire  $d_j = \{t_{j1}, \dots, t_{jn}\}$  avec  $n$  exprimant la taille du vocabulaire. On suppose que les termes sont indépendants et que la probabilité d'apparition du terme  $t_{ji}$  dans un document pertinent  $d_j$  pour la requête  $q$  suit une loi de Bernouilli de paramètre  $p_i$  :  $P(t_{ji}|R = 1, q) = p_i$ .

**Question 1**

Démontrer que la probabilité d'un document  $d_j$  pertinent pour la requête  $q$  est :

$$p(d_j|R = 1, q) = \prod_{i=1}^n p_i^{t_{ji}} (1 - p_i)^{1-t_{ji}}.$$

**Question 2**

On note  $D_r \subset D$  l'ensemble des  $N_r$  documents pertinents. Donner l'expression de la log-vraisemblance du modèle binaire  $p(d_j|R = 1, q)$ .

**Exercice 4 : RI neuronale (2pt)**

On s'intéresse maintenant aux modèles de recherche d'information faisant appel aux techniques de machine learning.

**Question 1**

Expliquer en quoi les modèles de "learning-to-rank" se différencient des modèles de RI classiques (modèles vectoriels, probabilistes, de langue, ...).

**Question 2**

Quelles sont les deux familles de modèles neuronaux qui ont été proposées à ce jour en RI? Expliquer les grandes lignes et leurs différences. Quels sont les avantages/limites de chacune de ces familles?

**Exercice 5 : classification de sentiments (4pt)****Question 1**

On hésite entre une représentation en sac de mots et une représentation en tri-grammes de lettres. Quels sont les avantages et inconvénients de chacune des représentations? (par exemple, en terme de taille, de bruit généré, d'interprétabilité...)

**Question 2**

Etant donnée la nature particulière de ce problème et en vous appuyant sur le projet, quel choix de représentation du texte feriez-vous et pourquoi ? Indiquez quelques pré-traitements qui vous semblent utiles et quelques-uns que vous éviteriez ici. Que dire des stop-words tels que *would*, *should* ou *not* ?

Utiliserez-vous la même représentation pour un problème de classification d'auteurs ?

**Question 3**

Classiquement, les données d'avis utilisateur collectées sur le web présente une échelle de notation sur 5 étoiles. Rappeler la procédure de binarisation classique de la problématique en justifiant très brièvement.

Les notes sur internet sont habituellement très favorables au produit, typiquement la distribution des notes s'apparente à quelque chose de la forme : [10, 15, 10, 35, 30]. Il y a donc un problème d'équilibre des classes sur le problème binaire. Quelles sont les conséquences de ce déséquilibre ? Comment y remédier du point de vue de l'implémentation, de la formulation et de l'évaluation ?

**Question 4**

Quels classifieurs sont classiquement utilisés pour classer ces données ?

**Question 5**

Comment faudrait-il procéder pour construire un classifieur à 5 classes distinguant chacune des 5 classes ? Que pourrait-on attendre de l'analyse qualitative associée à cette approche ? Est ce que cette approche serait légitime par rapport à une approche en régression ? Pourquoi ?

**Exercice 6 : sémantique (3pt)****Question 1**

Définir la sémantique du point de vue de l'informatique. Définir le fossé sémantique / *semantic gap* en même temps.

**Question 2**

Rappeler brièvement les philosophies générales des algorithmes PLSA et word2vec (philosophie = objectif + hypothèses pour arriver à cet objectif + éléments marquants de l'algorithme).

**Question 3**

Ces deux approches sont en fait assez différentes. Citer une ou deux applications qui sont liées à chaque algo (mais pas à l'autre). Expliquer brièvement pourquoi.

**Exercice 7 : Entités nommées (3pt)****Question 1**

Donnez deux raisons pour lesquelles un réseau de neurones récurrent est une approche qui convient bien à la reconnaissance d'entités nommées.

**Question 2**

Quelle est l'entrée d'un réseau récurrent pour la reconnaissance d'entités nommées ? Quelle est la sortie de ce réseau récurrent (avant classification) ?

**Question 3**

Les formats classiques IOB, IO, IOBES pour la reconnaissance d'entités nommées, ne sont pas très adaptés pour la reconnaissance des entités imbriquées, comme dans cet exemple vu en cours :

*fonction* *pers:sportif*  
Le **joueur de tennis américain** John McEnroe

Citer une façon possible de gérer ce problème.

### Exercice 8 : Cas d'usage (3pt)

Une entreprise soucieuse de sa E-réputation cherche à construire un dashboard (=un panneau d'indicateurs) pour suivre en quasi-temps réel les commentaires positifs et négatifs associés à différentes thématiques identifiées par le service marketing à l'aide d'un ensemble de mots clés.

#### Question 1

Quelles sont les différentes étapes générale requises ?

#### Question 2

Sur le plan NLP, comment construire les modèles capables d'estimer ces indicateurs ?

#### Question 3

Dans ce cadre général, comment détecter un incident imprévu (= une thématique non présente jusqu'ici) ?