Sorbonne Université
Examen M1 DAC
Recherche d'information et Traitement automatique du langage

Lundi 20 mai 2019

Exercice 1

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 :

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

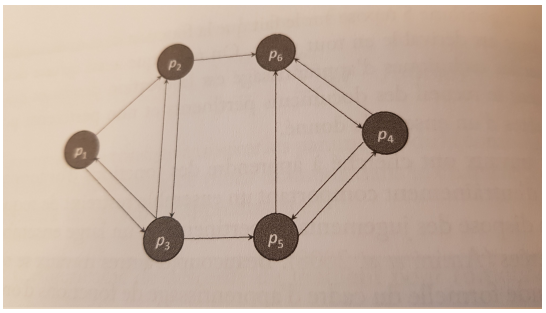
$$nDCG_p = \frac{DCG_p}{IDCG_p} \tag{2}$$

Exercice 2

On considère une collection de 6 documents reliés par des liens hypertextes.Le graphe dirigé de ces 6 documents est présenté en Figure 1.

Question 1

SANS FAIRE DE CALCULS, donner un ordonnancement de ces documents qui serait obtenu après application de l'algorithme de PageRank. Justifier votre réponse.



Graphe dirigé de la collection

Exercice 3

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

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

Question 1

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

Question 2

On suppose que le 50^e 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 4

On considère une collection de documents $D = \{d_1, ..., D_N\}$ 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 $d_1 = \{t_{j1}, ..., t_{jn}\}$ 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 5

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, ...). Donner une version possible de la vraisemblance en learning-to-rank.

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 chacunes de ces familles ?

Exercice 6 : classification de sentiments

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, 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* ou *should* ? Utiliseriez-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 (passage à un problème à 2 classes).

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

Dans le cadre d'une collaboration avec un linguiste, nous voulons construire un corpus d'adjectifs associés à une polarisation. Nous voulons exploiter le corpus de revues étiquetées pour y arriver. Proposer une procédure en détaillant les étapes par lesquelles vous passeriez.

Question 6

La société CA cherche à analyser la polarité des contributions sur Facebook la concernant pour mieux cerner les communautés d'opinions. Quels sont les problèmes qu'elle va rencontrer ? Proposer rapidement une ou deux idées pour faire face à ces problèmes.

Exercice 7 : sémantique

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 et le fonctionnement des algorithmes PLSA/LDA et word2vec (philosophie = objectif + hypothèses pour arriver à cet objectif + éléments marquants de l'algorithme). Donner les principaux hypers-paramètres associés aux deux modèles. Quelles sont les sorties associées aux deux modèles.

Exercice 8 : knowledge graph

On considère le texte suivant :

> `Dr` House (House, M.D., puis House) est une série télévisée américaine en 177 épisodes de 43 minutes et répartis sur huit saisons. Elle a été créée par David Shore et sa diffusion s'est déroulée du 16 novembre 2004 au 21 mai 2012 sur le réseau Fox.
> (Wikipédia)

Nous souhaitons développer un algorithme de construction de graphe de connaissances. L'idée est de transformer le texte en une série de triplets : { (`Dr House`, *TYPE*, `série TV`), (`Dr House`, *NATIONALITE*, `américaine`), ...}

Question 1

Pourquoi être intéressé par une telle transformation ? (rapidement)

Question 2

Quelles sont les problématiques en jeu ?

Question 3

En imaginant que vous deviez concevoir un tel système, comment procéder ? Donner quelques grandes étapes pour la construction des jeux de données d'apprentissage permettant d'entrainer les modèles attaquant les problématiques mentionnées dans la question précédente.

Exercice 9 : cas d'usage

Dans le temps qui vous reste après les nombreuses questions précédentes, imaginez une idée clé permettant de créer une start-up florissante [basée sur le NLP].

Après avoir énoncé cette idée, décrire les différentes étapes pour créer le système visé : les données nécessaires, celles à collecter sur le web, celles qui peuvent être obtenues en crowd-sourcing avec des systèmes tel que Amazon Mechanical Turk ; décrire les techniques de machine-learning à mettre en oeuvre ; décrire la méthodologie d'évaluation...