Sorbonne Université

Examen M1 DAC

Recherche d'information et Traitement automatique du langage

Lundi 16 mai 2022

— Les documents non autorisés
— Le barème indiqué est donné à titre indicatif

1 Exercice 1 : Questions de cours (2,5 points)

Question 1

Décrire les étapes du processus d'indexation.

Question 2

Quel est l'intérêt de la pondération TF-IDF ?

Question 3

Quel est le principe du modèle de langue ? Quelle est sa formulation obtenue par maximum de vraisemblance ?

Question 4

Expliquer le principe de la relevance feedback. Quelle est la différence avec la pseudo-relevance feedback ?

Question 5

Expliquer en quoi un modèle de recherche d'information peut être vu comme un modèle de classification, en vue d'utiliser des techniques de machine learning.

2 Exercice 2 : Représentation et similarité (2,5 points)

On considère une requête q contenant les termes $OS, Jaguar$ et trois documents ~~de même taille~~ d_1, d_2 and d_3 qui contiennent respectivement *Jaguar, Jaguar, Jungle, Jungle, Jungle,* et *Système d'exploitation, Jaguar, Mac, Système d'exploitation*, et *Jaguar, Bentley, Mercedes, Jaguar, Jaguar*. On considère l'abréviation *S.E.* pour *Système d'exploitation* et on suppose un vocabulaire $\mathcal{V} = bentley, jaguar, jungle, mac, mercedes, os, S.E.$.

Question 1

Donner les vecteurs tf (nombre d'occurrences) associés aux documents et à la requête.

Question 2

Ordonner les documents par rapport à leur score *produit scalaire* avec la requête. Commenter les résultats. Que pouvez-vous déduire du mot *Jaguar*.

Exercice TAL-2 : sémantique (~3 pts)

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 TAL-3 : knowledge graph (~3 pts)

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 ? Vous répondrez en terme d'entrées/sorties en ajoutant éventuellement un nom de modèle adapté.

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 TAL-4 : cas d'usage (bonus, ~2 pts)

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 crowdsourcing avec des systèmes tel que Amazon Mechanical Turk ; décrire les techniques de machine-learning à mettre en oeuvre ; décrire la méthodologie d'évaluation...

On souhaite maintenant augmenter la couverture des termes du vocabulaire en prenant en compte dans la représentation vectorielle : 1) le nombre d'occurrence de ce terme, et 2) les termes synonymes de ceux apparaissant dans les documents. Un moyen simple consiste à définir une matrice de similarité W entre les termes et de projeter les termes des documents et de la requête pour calculer la représentation vectorielle. Ainsi, un terme absent du document peut se retrouver avec un poids positif s'il est similaire à un ou plusieurs termes du document. Considérons la matrice W suivante (l'ordre des colonnes et des lignes suivent la définition du vocabulaire $\mathcal{V}$) :

$$W = \begin{pmatrix} 0,5 & 0,1 & 0 & 0 & 0,4 & 0 & 0 \\ 0,1 & 0,5 & 0,05 & 0,05 & 0,1 & 0,1 & 0,1 \\ 0 & 0,05 & 0,95 & 0 & 0 & 0 & 0 \\ 0 & 0,05 & 0 & 0,8 & 0 & 0,05 & 0,1 \\ 0,4 & 0,1 & 0 & 0 & 0,5 & 0 & 0 \\ 0 & 0,1 & 0 & 0,05 & 0 & 0,55 & 0,3 \\ 0 & 0,1 & 0 & 0,1 & 0 & 0,3 & 0,5 \end{pmatrix}$$

Question 3

Quelles sont les nouvelles représentations des documents et de la requête ?

Question 4

Calculer les nouveaux scores produits scalaires entre les documents et ordonner. Conclure.

Question 5

Si on suppose que les termes qui apparaissent dans les mêmes documents avec les mêmes fréquences sont sémantiquement similaires, donner un moyen simple de calculer la matrice de similarité W entre termes.

3 Exercice 3 : Etude de cas (4 points)

Mettons-nous dans le cas d'une application médiale, dans laquelle on gère des documents de diagnostics médicaux d'examens. Dans certains diagnostics, on peut lire : "tumeur à droite du poumon gauche". Dans d'autres : "tumeur à gauche sur le poumon droit". Supposons que l'on utilise un anti-dictionnaire avec les mots utilitaires habituels.

Question 1

Expliquer en quoi le modèle vectoriel vu en cours est inadapté à ce cadre. Donnez brièvement des exemples pour étayer vos arguments.

Question 2

Proposer une approche (pas trop détaillée !) qui permettrait de résoudre ce problème au niveau de l'indexation des documents. Votre proposition devra être en mesure de traiter les exemples cités ci-dessus.

Question 3

Discutez votre proposition en indiquant comment vous pourrez, lors de la recherche, être capable de rechercher des documents parlant de "tumeur à droite du poumon gauche" qui ne retourne pas ceux qui parlent de "tumeur à gauche sur le poumon droit".

Question 4

Discutez, pour votre proposition, la difficulté potentielle pour extraire les termes, pour la pondération, pour les index inversés, etc.

4 Exercice 4 : Formulation de requête / Rocchio (2 points)

L'hypothèse sous-jacente de l'algorithme de Rocchio est que pour une requête initiale q_0 et deux ensembles de documents pertinents $\mathcal{D}_p$ et non pertinents $\mathcal{D}_{np}$ par rapport à q_0, la nouvelle

requête q^* est celle qui vérifie la condition suivante :

$$q^* = argmax_q(s(q, \mathcal{D}_p) - s(q, \mathcal{D}_{np})) \quad (1)$$

Question 1

Montrer que si la fonction de score entre une requête et un document $s(q, ...)$ est la fonction produit scalaire, la condition précédente serait équivalente à :

$$q^* = \frac{1}{||\mathcal{D}_p||} \sum_{d \in \mathcal{D}_p} d - \frac{1}{||\mathcal{D}_{np}||} \sum_{d' \in \mathcal{D}_{np}} d' \quad (2)$$

Question 2

D'après ce résultat, justifier que la mise à jour du vecteur de la requête proposée par Rocchio est :

$$q^* = \alpha q_0 + \beta \frac{1}{||\mathcal{D}_p||} \sum_{d \in \mathcal{D}_p} d - \gamma \frac{1}{||\mathcal{D}_{np}||} \sum_{d' \in \mathcal{D}_{np}} d' \quad (3)$$

où α, β, γ sont des réels positifs.

Exercice TAL-1 : classification de sentiments (~4 pts)

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.