RITAL – Examen réparti – 13/05/24 – 60 pts

Durée : 2h

Seuls documents autorisés : Aucun document autorisé
– Barème indicatif –

1 Partie Traitement du Langage (TAL)

Exercice 1 (10 points) – Modèle Bag of Words (BoW)

Q 1.1 La mise en place du modèle Bag of Words (BoW) comporte deux étapes principales pour encoder un texte brut sous forme vectorielle : 1) pré-traitements, et 2) construction du vecteur BoW. Préciser les choix et variantes possibles ainsi que les implications de ces choix en fonction des tâches de classification de sentiments et reconnaissance de locuteur étudiées dans le projet.

Q 1.2 On considère les deux configurations de BoW suivantes :

— Dictionnaire de bi-grammes produisant un vecteur de taille 10^5.

— Dictionnaire d'unigrammes avec un filtrage du vocabulaire donnant un vecteur de taille 100.

On dispose de 1000 exemples d'entraînement pour prédire la catégorie de documents à partir des deux représentations BoW précédentes. Quel modèle d'apprentissage préconisez-vous dans les deux cas ? Justifier.

Q 1.3 Dans la tâche de prédiction de locuteur du projet, donner en les justifiant les métriques d'"évaluation à utiliser. Préciser la façon dont la fonction objectif d'entraînement doit être adaptée. Enfin, décrire un ~~pré-traitement~~ post sur les prédiction pour améliorer les performances.

Exercice 2 (5 points) – Clustering

Q 2.1 Donner deux différences principales entre la méthode LSA et l'algorithme du K-Means, en termes i) d'expressivité du modèle de clustering, et ii) de contraintes appliquées sur l'ensemble des vecteurs ("champs lexicaux") appris.

Q 2.2 Est-il intéressant d'appliquer LSA avant le K-Means ? Si oui justifier. Sinon, expliquer.

Q 2.3 Dans la méthode PLSA, est-il possible d'obtenir une formule analytique pour l'estimation des paramètres par maximisation de vraisemblance ?

Exercice 3 (7 points) – Séquences

On considère un problème d'étiquetage de phrases de type "Part of Speech" (PoS) tagging.

Q 3.1 On modélise la séquence des mots dans une phrase par une chaîne de Markov.

— Quels sont les états et observation pour ce problème ?

— Quels sont les paramètres de la chaine de Markov ?

— Comment estimer les paramètres de la chaîne de Markov ? Existe-t-il une formule analytique pour cela ? Justifier.

Q 3.2 Peut-on utiliser un modèle BoW pour la tâche de PoS ? Si oui, préciser. Sinon, justifier.

Q 3.3 Donner deux avantages concernant l'utilisation d'un un champ aléatoire conditionnel (CRF) par rapport à un modèle de Markov pour la tâche PoS.

Exercice 4 (8 points) – Réseaux de neurones

Q 4.1 Décrire l'hypothèse distributionnelle pour l'entraînement d'embeddings vectoriels.

— Comment est-elle mise en oeuvre dans Word2Vec (CBoW et Skip-gram) et Glove ?

— Pourquoi ces méthodes sont elles dites auto-supervisées ?

Q 4.2 Quelle est la spécificité architecturale des transformers ? Que permet-elle ?

Q 4.3 Expliquer le principe de l'"In Context Learning" avec les Large Language Models (LLMs). Cette étape nécessite-t-elle une mise à jour des paramètres du modèle ?

2 Partie Recherche d'information (TAL)

Exercice 5 (6 points) – Loi de Zipf

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

$$frequence = \frac{\lambda}{rang} (avec\ \lambda > 0) \quad (1)$$

Q 5.1 Quelle est l'intuition de la loi de Zipf ? A quoi sert-elle ?

Q 5.2 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 ?

Q 5.3 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 6 (4 points) – Modèles probabilistes

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_j, dans un document pertinent d_j pour la requête q suit une loi de Bernouilli de paramètre p_i : $P(t_{ji}|R = 1, q) = p_i$.

Q 6.1 Démontrer que la probabilité d'un document d_j pertinent pour la requête q est : $p(d_j|R=1.q) = \prod_i^n {}_1 p_i^{t_{ji}} (1-p_i)^{1-t_{ji}}$.

Q 6.2 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 7 (4 points) – Formulation de requête / Rocchio

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 :

$$q^* = argmax_q(s(q,\mathcal{D}_p) - s(q,\mathcal{D}_{np})) \quad (3)$$

Q 7.1 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 :

$$q^* = \frac{1}{||\mathcal{D}_p||}\sum_{d\in\mathcal{D}_p} d - \frac{1}{||\mathcal{D}_{np}||}\sum_{d'\in\mathcal{D}_{np}} d' \quad (4)$$

Q 7.2 D'après ce résultat, justifier que la mise à jour du vecteur de la requête proposée par Rocchio est :

$$q^* = \alpha q_0 + \beta\frac{1}{||\mathcal{D}_p||}\sum_{d\in\mathcal{D}_p} d - \gamma\frac{1}{||\mathcal{D}_{np}||}\sum_{d'\in\mathcal{D}_{np}} d' \quad (5)$$

où $\alpha.\beta.\gamma$ sont des réels positifs.

Exercice 8 (5 points) – Modèles neuronaux

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

Q 8.1 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.

Q 8.2 Quelles sont les grandes 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 9 (11 points) – Etude de cas

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.

Q 9.1 *Expliquer en quoi le modèle vectoriel vu en cours est inadapté à ce cadre. Donnez brièvement des exemples pour étayer vos arguments.*

Q 9.2 *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.*

Q 9.3 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".

Q 9.4 Discutez, pour votre proposition, la difficulté potentielle pour extraire les termes, pour la pondération, pour les index inversés, etc.