**Examen du cours Recherche d'information**

**ENSEEIHT 3A– année 2021-2022**

Les documents sont autorisés.
Durée : 1h30

**Exercice 1 (8)**

1- Donner l'impact des procédures ci-dessous, en termes de rappel et de précision (indiquer si elles améliorent ou détériorent ces facteurs) :
   a) Utilisation de la lemmatisation des mots
   b) Utilisation des synonymes
   c) Suppression des mots vides
   d) Représentation des documents et des requêtes par des vecteurs (CLS) issus de BERT.
   e) Utilisation des bigrammes de mots dans les modèles de RI basés sur les modèles de langue de type

2- Répondre par Vrai ou Faux et justifier votre réponse
   a) Un document ne comportant aucun terme de la requête ne peut pas être pertinent pour cette requête
   b) La représentation en sac de mots permet de capturer (représenter) le sens des mots.
   c) La représentation en trigrammes permet de capturer la syntaxe.
   d) L'utilisation des représentations distribuées de mots (*Word embedding*) permet de sélectionner des documents pertinents même si ces derniers n'ont aucun terme en commun avec la requête
   e) Un système de recherche d'information basé sur un modèle de langue de type bigrammes renvoie deux listes différentes pour les requêtes suivantes : « Information retrieval » et « Retrieval information »

3- A votre avis, quel modèle de RI (parmi ceux étudiés en cours) serait le plus approprié pour rechercher des tweets ? Justifiez votre réponse ? On suppose qu'un tweet est formé d'un message où les mots se répètent rarement.

4- Les modèles BERT et GPT représentent les textes dans des espaces de termes de tailles réduites (≈ 30 mille termes pour BERT) et (≈ 50 mille termes pour GPT), or l'*Oxford English Dictionary* recence **plus de 200 mille mots.** Pensez-vous que les modèles BERT ou GTP éliminent plus de 150 mille mots ? sinon expliquez comment ces modèles s'y prennent pour représenter une bonne partie des mots du dictionnaire dans un espace réduit

**Exercice 2 (6)**

Nous disposons d'une collection comportant les 4 documents suivants:

D1. jean donne un livre à marie
D2. jean qui lit le livre travaille avec marie
D3. qui pense que jean travaille avec marie ?
D4. jean pense qu'un livre est un bon cadeau

**Questions**

1- On suppose que ces documents sont indexés avec suppression des mots vides (à, un avec, qui, que, qu', est, le) et racinisation à 7 caractères. Proposez une structure de fichier inversé (dictionnaire+*posting*) permettant de représenter ces documents indexés

2- Considérons la requête suivante : **q** (marie marie travail)
Donner les scores de pertinence des documents D1 et D2 vis-à-vis de la requête **q** pour les 4 modèles suivants :

- Le modèle vectoriel utilisant la pondération de type qqq.ddd=nnn.ltn,
- Le modèle probabiliste BIR,
- Le modèle de langue basé sur l'interpolation de Jelinek Mercer (JM) avec $\lambda$=0.5

**Exercice 3 (6pts)** :

La table ci-dessous liste les documents trouvés par un système de recherche d'information, Sys, en réponse à une requête, parmi les 10 documents d'une collection. Les documents retrouvés sont listés par ordre décroissant de leur pertinence (le premier document est D1 le dernier est D10). La valeur **1** de de la ligne Sys de la table indique que le système a effectivement sélectionné le document spécifié dans la colonne correspondante et la valeur **0** indique que le document n'a pas été pas retrouvé.
La dernière ligne « Pert » indique si le document est pertinent (noté **1**) ou non pertinent (noté **0**) pour la requête.

| docs | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 | D10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Sys | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |
| Pert | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |

**Questions**

1- Calculer la courbe de rappel-précision interpolée du système Sys.
2- Calculer la précision moyenne de Sys.
3- Calculer également sa R-Précision.