Université de Médéa.
Faculté des sciences et de la technologie.
Départ : Génie électrique & Informatique.
2ere Année Master Informatique (ISTW)
Matière: RI

Mini-Projet1

Partie A.

I. Création d'index

1- Créer une fonction d'initialisation du Moteur de recherche **Whoosh** https://pypi.org/project/Whoosh/
2- Créer des objets pour : créer, ouvrir et fermer un fichier indexé (**Index**) par Woosh
3- Créer une fonction qui permet d'ajouter un fichier à l'index (avec titre, contenu, chemin) (**MAJ**).

II. Recherche dans l'index

Pour préparer les requêtes, il faut un analyseur et un parseur lié à l'index.

1. Créer un objet de gestion de vos requêtes et une fonction qui permet d'initialiser le moteur de requête.
2. Créer une fonction qui permet d'interroger votre index et d'afficher le résultat à l'aide d'une requête et récupérer les X (ex :20) premiers résultats.
3. Faire la recherche en modifiant le mode d'appariement (Probabiliste, Vectoriel, LSI..)

Partie B : Evaluation (sur un corpus Arabe)

Cette partie est subdivisée en 3 phases comme suit :

0. Faire une recherche manuelle des documents pertinents relativement à 3 requêtes (pour chaque étudiant) à partir du corpus et préparer votre fichier texte nommé par le N° de la requête contenant les noms de ces fichiers pertinents.

Phase1 : Evaluation

1. Faire la recherche par Woosh pour chaque requête (Modèle Vectoriel et LSI) et puis marquer les documents pertinents dans la liste des documents retournés (dresser un **Tableau** [N°, Nom de document, Pertinent (O/N)].
2. Calculer la précision et rappel aux points des documents pertinents (utiliser le **Tableau** précédent).
3. Calculer AvrPrec (précision moyenne).
4. Faire le **Tableau** des précisions interpolées aux 11 points de rappel (0.0, 0.1, 0.2, ..., 1.0)
5. Faire la courbe.
6. Calculer F-Mesure.
7. Calculer la précision aux points (P5, P10, P15, P20) (Dresser le **Tableau**).

Phase2 : Reformulation Automatique

Etapes pour la reformulation de la requête :

Télécharger l' ArabicWordNet (AWN) sur :

http://globalwordnet.org/arabic-wordnet/awn-browser/

ou

http://sourceforge.net/projects/awnbrowser/

I- Reformulation Automatique

1. Prendre les requêtes Qi initiales
2. Reformuler en utilisant le ArabicWordNet (AWN)
    i. Tockeniser la requête en Mots
    ii. Prendre chaque mot Mi (non vide) et chercher les Synset (un synset est un ensemble de mots sémantiquement liés) dans l'AWN, Sj{…..,Mi,…}
    iii. Valider et Prendre uniquement le synset lié au domaine de la requête et éliminer les autres synsets
    iv. Refaire l'étape 2 et 3 avec tous les mots de la requête
3. Lemmatiser tous les mots (Mi+Mi') de Q' (Q+ Tous les mots des synset)
4. Mettre la requête lemmatisée Q' sur le moteur Woosh
5. Faire la recherche par Woosh (Modèle Vectoriel et LSI) et marquer les documents pertinents dans la liste des documents retournés (**Tableau** [N°, Nom de document, Pertinent (O/N)]
6. Calculer la précision et rappel aux points des documents pertinent (Dresser le **Tableau**).

Phase3 : Reformulation Automatique + Validation utilisateur

Refaire les étapes précédentes : (pour chaque requête)

1. Prendre les requêtes Qi .
2. Reformuler en utilisant ArabicWordNet (AWN)
    i. Tockeniser la requête en Mots
    ii. Prendre chaque mot Mi (non vide) et chercher les Synset (un synset est un ensemble de mots sémantiquement liés) dans l'AWN, Sj{…..,Mi,…}
    iii. Valider et Prendre uniquement les mots valides (liées au domaine de la requête) dans le synset aussi lié au domaine de la requête et éliminer les autres synsets, OU modifier le poids des mots de la requête
    iv. Refaire l'étape 2 et 3 avec tous les mots de la requête
3. Lemmatiser tous les mots (Mi+Mi') de Q' (Q+ mots choisis par l'utilisateur)
4. Mettre la requête lemmatisée Q' sur le moteur Woosh
5. Faire la recherche par **Woosh** (Modèle Vectoriel et LSI) et marquer les documents pertinents dans la liste des documents retournés (dresser un **Tableau** [N°, Nom de document, Pertinent (O/N)]
6. Calculer la précision et rappel aux points des documents pertinent (Dresser le **Tableau**).

Phase4 : Conclusion -Tableaux récapitulatifs-

Dresser les tableaux résultats comme l'exemple suivant

1-

Requête	Mots de requête1	Mots de requête2	Mots de requête3
Requête originale	M_1 M_2 M_3...	M_1 M_2 M_3...	M_1 M_2 M_3...
Requête Reformulée automatique	M_1 M'_1 M''_1 M_2 M'_2 M_3	M_1 M'_1 M''_1 M_2 M'_2 M_3	M_1 M'_1 M''_1 M_2 M'_2 M_3
Requête Reformulée automatique + validation de l'utilisateur	M_1 M'_1 M_2 M'_2 M_3	M_1 M'_1 M_2 M'_2 M_3	M_1 M'_1 M_2 M'_2 M_3

2-

Mots des Requêtes Q1 et Q2	AWN
	Existe en AWN
M1	Oui
M2	Oui
M3	Non
Mi	Oui
	.

3-

	Précision Interpolée Requete1					
Rappel	Sans reformulation (TP2)		Reformulation automatique		Reformulation automatique + validation de l'utilisateur	
	Modèle1 Vectoriel	Modèle2 LSI	Modèle1 Vectoriel	Modèle2 LSI	Modèle1 Vectoriel	Modèle2 LSI
0,00	1.00	1.00	1.00	1.00	1.00	1.00
0,10	1.00	0.94	0.876		0.976	
...	...			...		...
1.00	0.000		0.000		0.100	

	Précision Interpolée Requete2					
Rappel	Sans reformulation (TP2)		Reformulation automatique		Reformulation automatique + validation de l'utilisateur	
	Modèle1	Modèle2	Modèle1	Modèle2	Modèle1	Modèle2

	Précision Interpolée Requete3					
Rappel	Sans reformulation (TP2)		Reformulation automatique		Reformulation automatique + validation de l'utilisateur	
	Modèle1	Modèle2	Modèle1	Modèle2	Modèle1	Modèle2

	Précision Interpolée Requete1 + Requete2 + Requete3					
Rappel	Sans reformulation (TP2)		Reformulation automatique		Reformulation automatique + validation de l'utilisateur	
	Modèle1	Modèle2	Modèle1	Modèle2	Modèle1	Modèle2

Conclusion

4-

Précision	Sans reformulation			Reformulation automatique			Reformulation automatique + validation de l'utilisateur		
	Modele1 (Q1+Q2+Q3) /3	Modele2 (Q1+Q2+Q3) /3	Système	Modele1 (Q1+Q2+Q3) /3	Modele2 (Q1+Q2+Q3) /3	Système	Modele1 (Q1+Q2+Q3) /3	Modele1 (Q1+Q2+Q3) /3	Système
P@5	0.600	0.650							
P@10	0.650	0.650							
P@15	0.570	0.570							
P@20	0.450	0.500							
P@100	0.290	0.310							
AvrPrec 1,	0.350	0.380							
AvrPrec 2	0.43	0.32							
MAP	0.39	0.35							
Relevant Doc	73	73							
returned System	143	143							
returned Rel Doc	64	69							69