Recherche d'Information

SIHAMDI Mostefa, BOUSBA Abdellah

UE RITAL 2020-2021, Encadrante : Laure Soulier

M1 DAC

Répartition du travail :

Nous avons équitablement repartie le travail de chaque TME d'une façon que chacun implémente des fonctionnalités indépendantes et ensuite les regroupé ensemble, on a réussi à finir le travail demandé dans chaque TME avant le début de l'autre, en ajoutant une séance de révision et test générales a la fin.

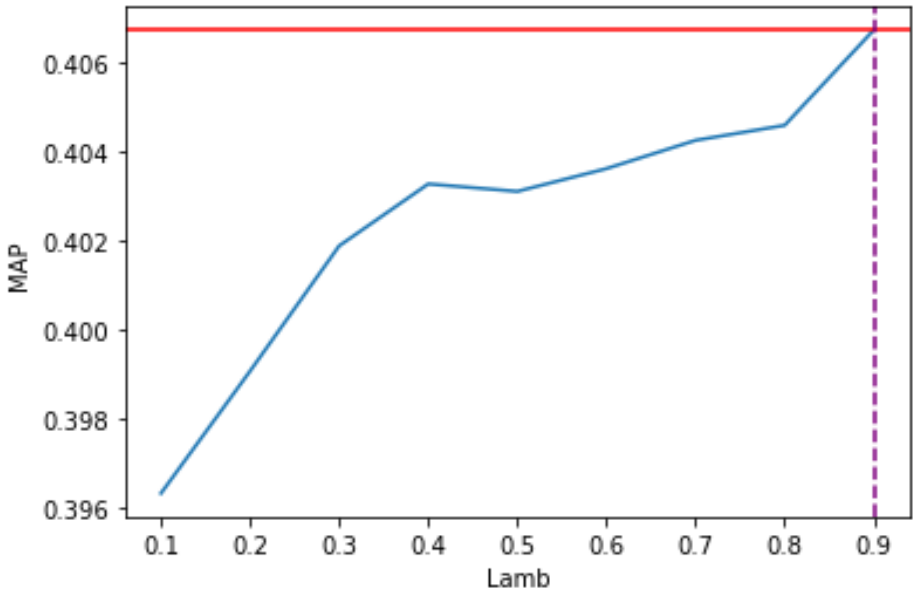
Tâches réalisées et bonus :

Nous avons réussi à faire tout le travail demandé en plus les bonus qui concerne l'optimisation des paramètres des modelés, cross validation, l'optimisation des paramètres de pagerank, et le test de similarité. Nous n'avons malheureusement pas réussi a implémenter l'interaction avec l'index sans le charger en mémoire, mais on a ajouter une fonctionnalité de save/load pour ne pas indexer les documents à chaque fois.

Expérimentations sur la base cisi :

Modele	P_10	Recall_10	F_measure	MAP	Recip_rank	NDCG
Vectoriel1	0.10803	0.36439	0.11970	0.38868	0.26337	0.43919
Vectoriel2	0.11696	0.36702	0.12970	0.39823	0.31450	0.45576
Vectoriel3	0.12946	0.37111	0.14337	0.39950	0.33987	0.46108
Vectoriel4	0.14732	0.37496	0.15788	0.40587	0.32185	0.47229
Vectoriel5	0.11964	0.36620	0.12791	0.35924	0.31138	0.45351
Modele Langue	**0.16339**	0.37863	0.17059	**0.41357**	0.36139	0.49431
Okapi BM25	0.16071	**0.37970**	**0.17189**	0.41288	**0.36150**	**0.49527**

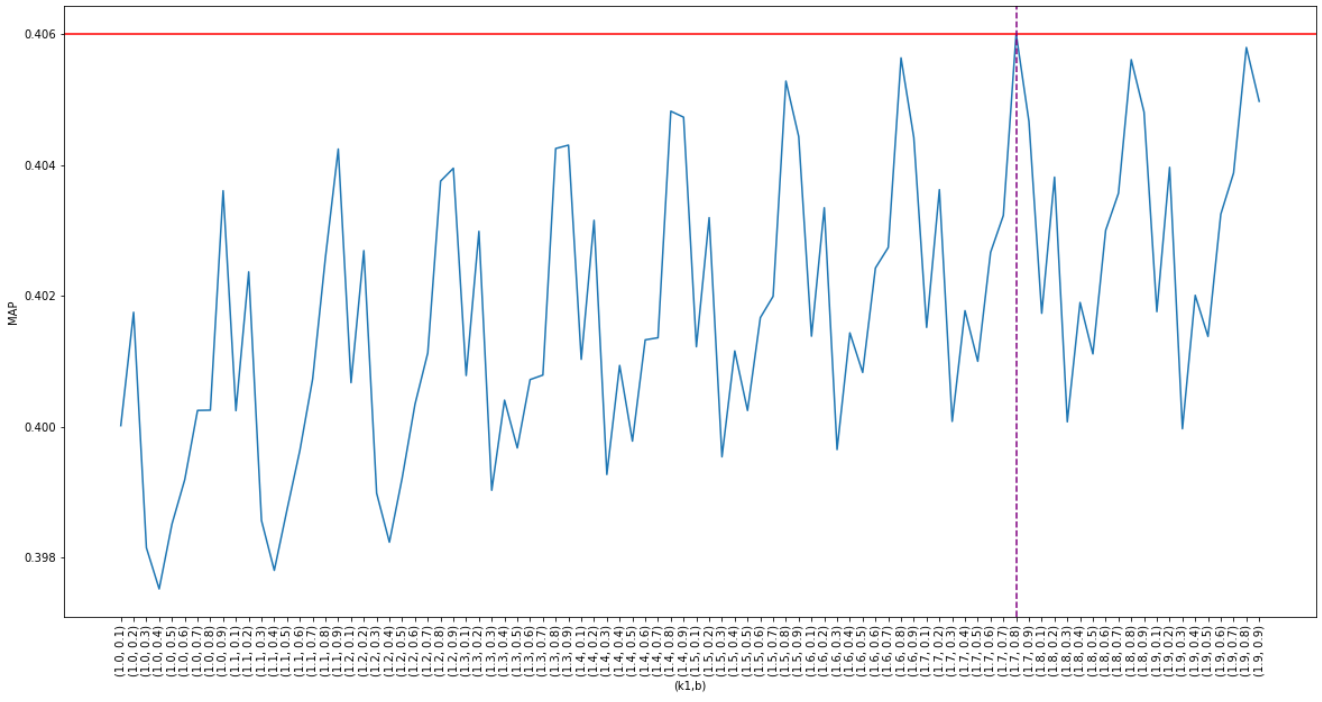
GridSearch Modèle Langue :



Avec un lamb = 0.9 on a 45% de MAP sur les données test

MAP cross_validation : 0.41653522231954493

GridSearch Okapi BM25 :

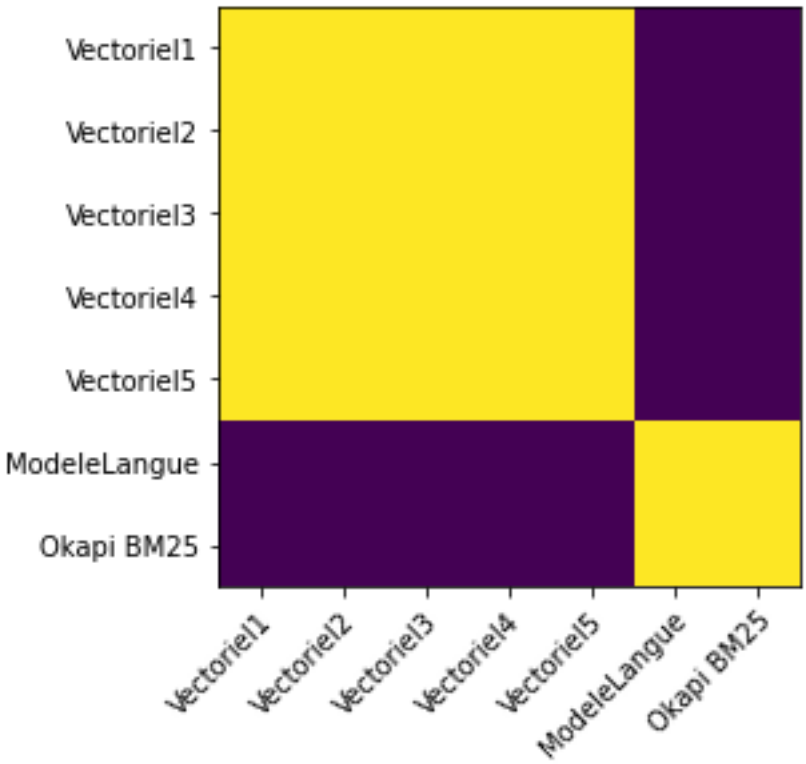


Avec un k1 = 1.7 et b = 0.8 on a 44% de MAP sur les données test

MAP cross_validation : 0.4132762898639264

Similarité entre modèle :

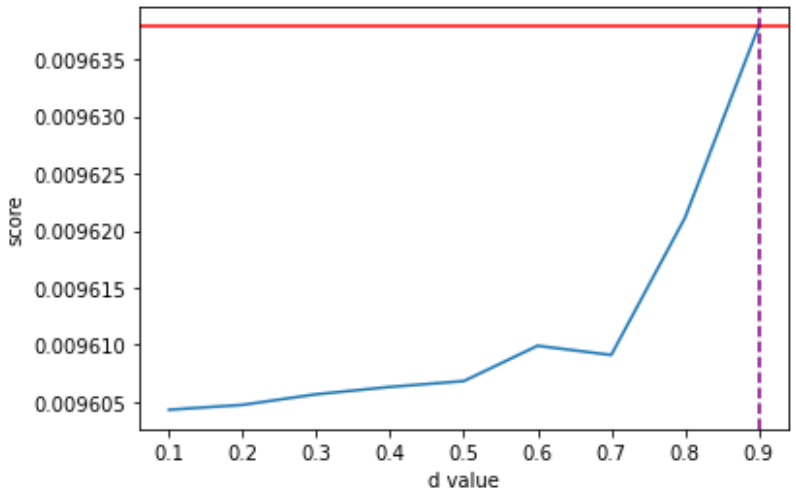
On remarque plus ou moins deux groupes, les modèles probabilistes qui sont similaire, l'autre groupe est les modèles vectoriels.



PageRank :

Prenant le modèle Okapi pour illustrer un exemple :

On realise un gridsearch pour avoir le meilleur parametre d :



Un exemple des documents retourner par l'algorithme pour la requête traité précédemment d = 0.9:

['830', '1421', '449', '175', '603', '625', '825', '530', '754', '812', '553', '643', '577', '608', '522', '526', '1216', '79', '644', '649', '73', '527', '660', '528', '1427', '755', '484', '579', '274', '997', '650', '333', '445', '523', '176', '785', '1435', '877', '874', '940', '1436', '941', '594', '517', '780', '546', '332', '390', '826', '628', '1079', '1374', '995', '1434', '57', '606', '38', '626', '752', '576', '508', '637', '878', '1282', '992', '534', '634', '993', '802', '589', '1144', '652', '790', '829', '895', '572', '1195', '641', '361', '52', '966', '446', '150', '422', '722', '1395', '645', '1266', '500', '709', '574', '2188', '1547', '1709', '2197', '2332', '2348', '382', '819', '499']