Recherche d'Information Web

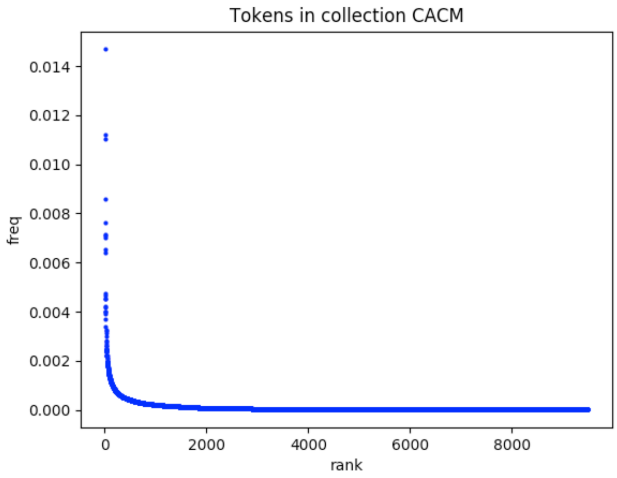
Réponses et résultats graphiques

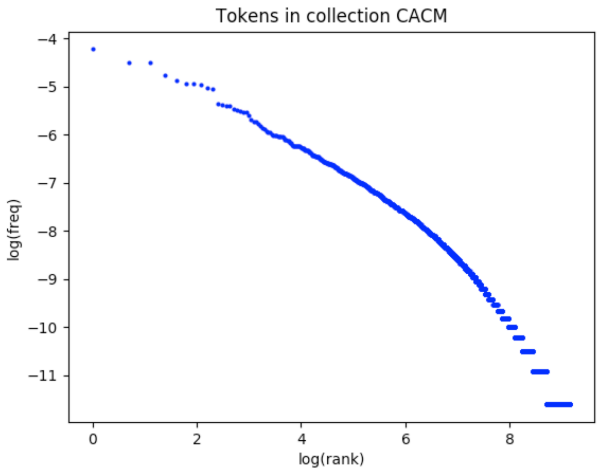
2.1 Traitements linguistiques

Le nombre de tokens et la taille du vocabulaire (nombre de mots différents) sont calculés tous les deux après les étapes de tokenization, de filtrage et éventuellement de normalisation. Ci-dessous les résultats des tests de traitement sur l'ensemble des deux collections (CACM et CS276), d'abord sans troncature des mots puis avec troncature.

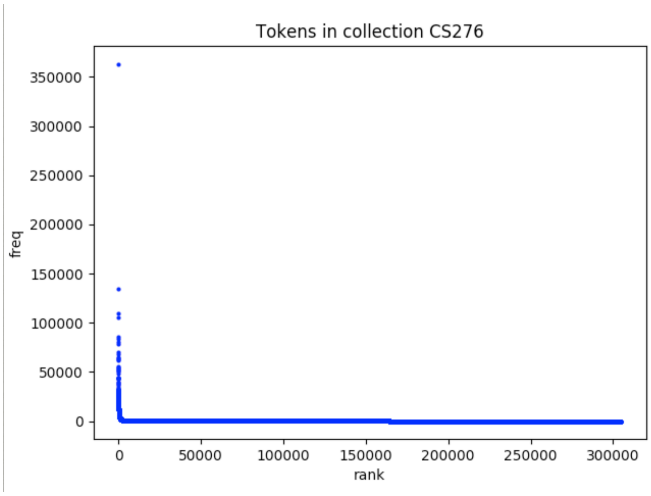
Cas sans normalisation

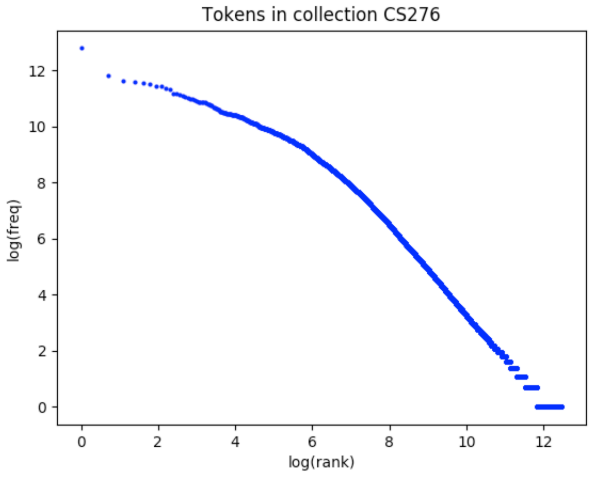
--- Collection CACM ---
Nombre de tokens : 110398
Taille du vocabulaire : 9496
Nombre de tokens pour la moitié de la collection : 55112
Taille du vocabulaire pour la moitié de la collection : 6899
Loi de Heaps : b = 0.459886135587183 , k = 45.53577842534476
Taille du vocabulaire pour une collection de 1 million de tokens : 26162





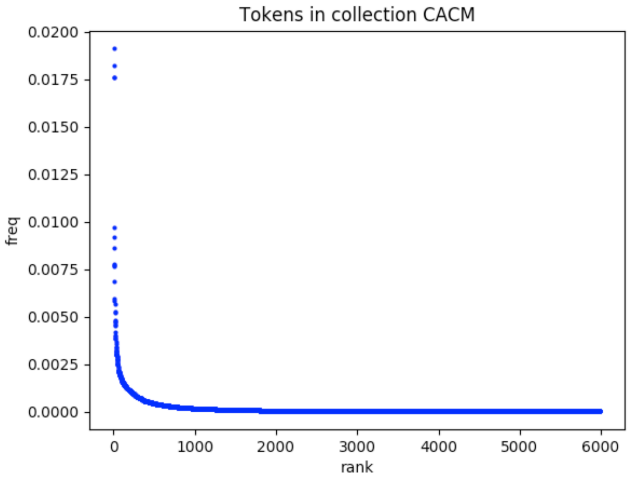
--- Collection CS276 ---
Nombre de tokens : 17484833
Taille du vocabulaire : 304871
Nombre de tokens pour la moitié de la collection : 9824138
Taille du vocabulaire pour la moitié de la collection : 178499
Loi de Heaps : b = 0.9285580609365727 , k = 0.05739636746106991
Taille du vocabulaire pour une collection de 1 million de tokens : 21391

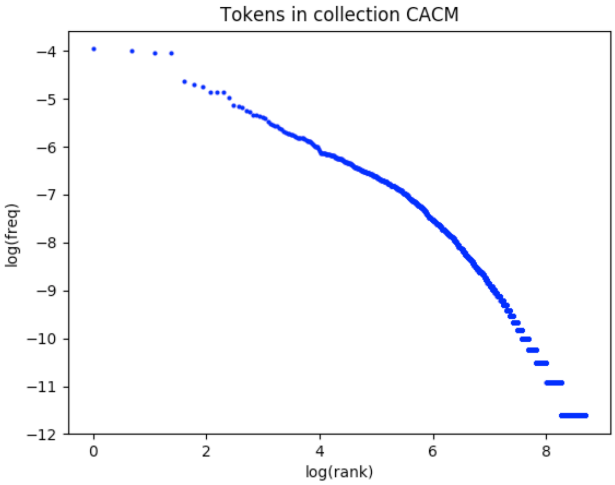




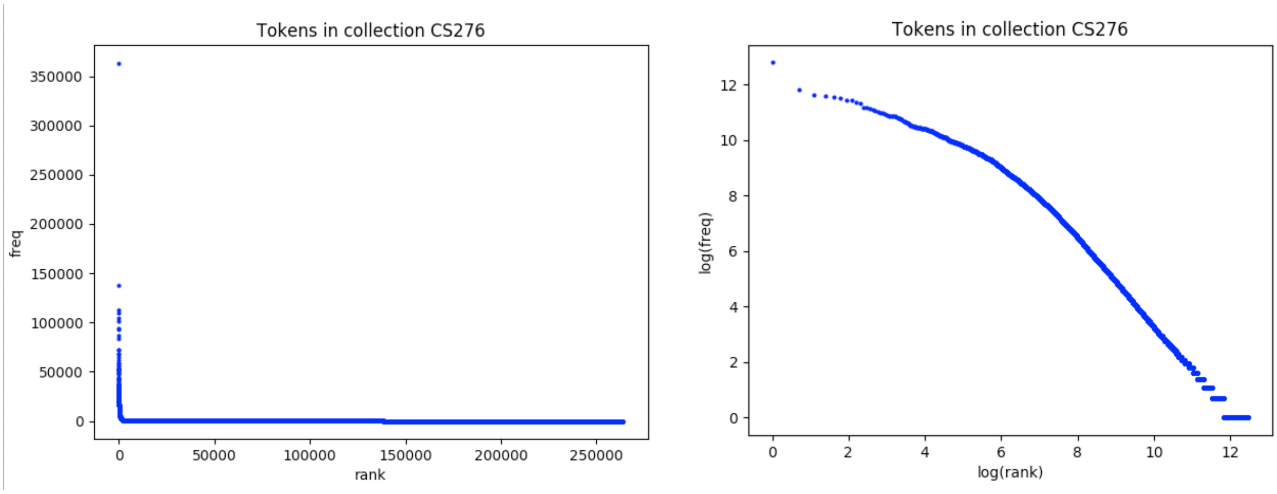
Cas avec normalisation

--- Collection CACM ---
Nombre de tokens : 110398
Taille du vocabulaire : 5992
Nombre de tokens pour la moitié de la collection : 55112
Taille du vocabulaire pour la moitié de la collection : 4376
Loi de Heaps : b = 0.4523953930659716 , k = 31.34435863927336
Taille du vocabulaire pour une collection de 1 million de tokens : 16238





--- Collection CS276 ---
Nombre de tokens : 17484833
Taille du vocabulaire : 263849
Nombre de tokens pour la moitié de la collection : 9824138
Taille du vocabulaire pour la moitié de la collection : 150243
Loi de Heaps : b = 0.9768107306242427 , k = 0.02221498335215894
Taille du vocabulaire pour une collection de 1 million de tokens : 16125



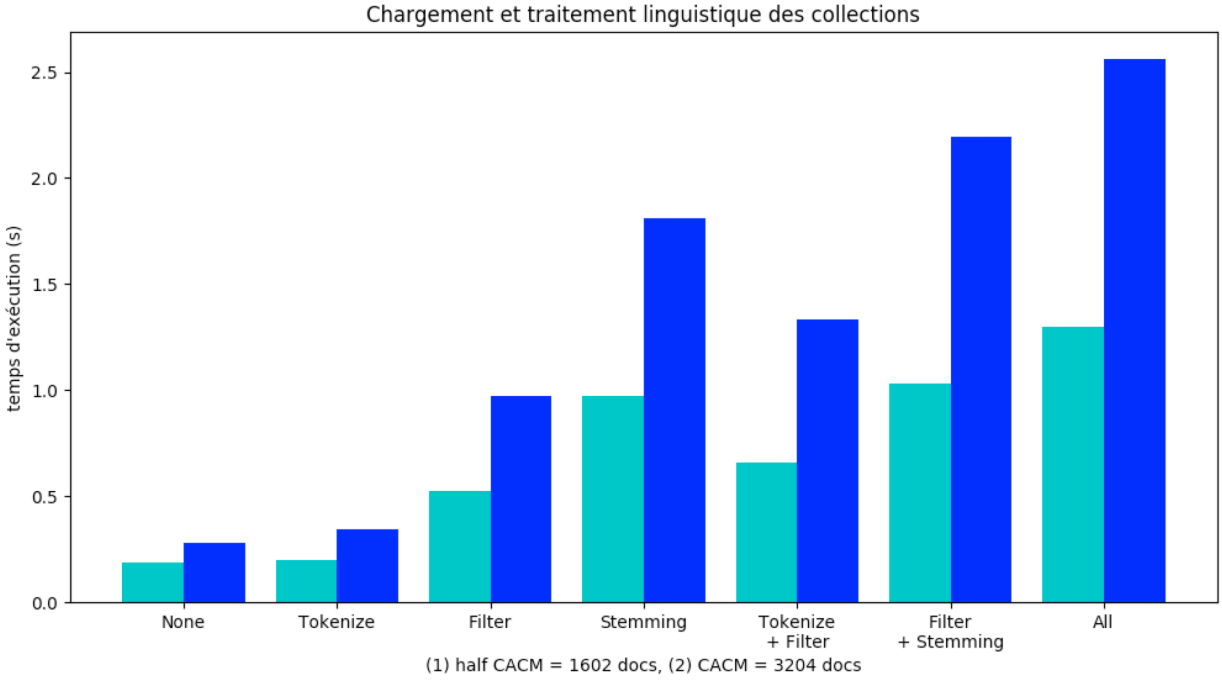
On observe naturellement une diminution de la taille du vocabulaire pour un même nombre de tokens lorsqu'on applique une troncature (en effet, des mots différents vont avoir la même racine, par exemple un nom et son pluriel).

On remarque aussi que les collections n'ont pas les mêmes coefficients de Heaps. La collection de l'Université de Stanford CS276 semble avoir un vocabulaire plus diversifié au sein du corpus que la collection CACM : le nombre de termes distincts est presque proportionnel au nombre de mots de la collection ($b > 0.92$).

2.3 Evaluation avec la collection CACM

Mesures de performance

Temps d'exécution des traitements linguistiques :



On teste différentes combinaisons de traitements linguistiques possible parmi la tokenization, le filtrage et la normalisation (aucun traitement signifie qu'on a juste séparé les mots déjà espacés).

Résultats détaillés :
- None: 0.277 (sec)
- Tokenize: 0.346 (sec)
- Filter: 0.972 (sec)
- Stemming: 1.809 (sec)
- Tokenize + Filter: 1.333 (sec)
- Filter + Stemming: 2.193 (sec)
- All: 2.561 (sec)

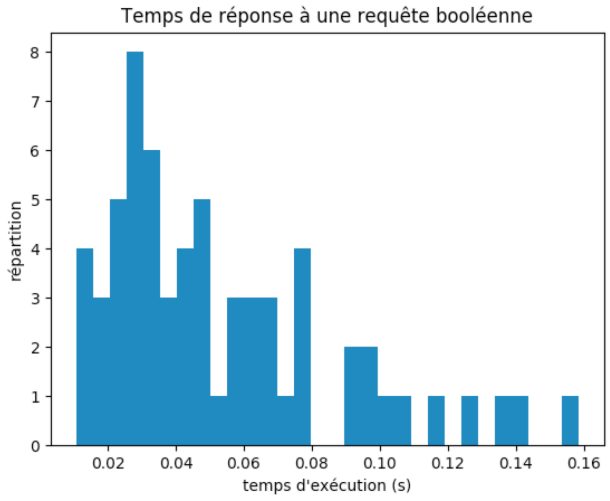
Temps de calcul pour l'indexation :

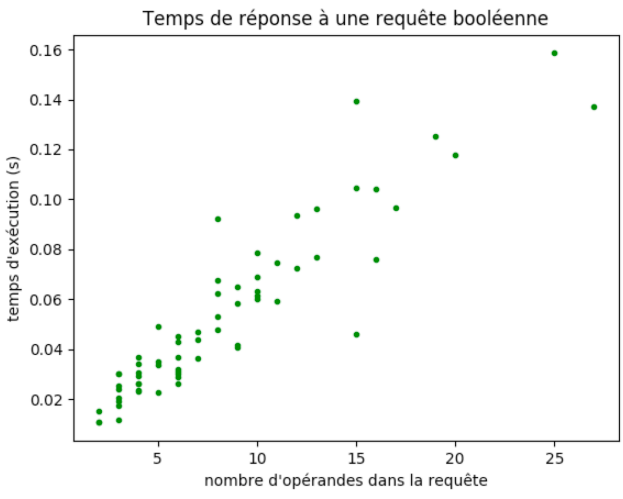
On teste la construction des deux types d'index (DocID index et Frequency index).
- DocID index: 2.674 (sec)
- Frequency index: 3.095 (sec)

Temps de réponse à une requête booléenne :

On teste le système de recherche booléen sur les 64 requêtes préalablement définies dans le fichier query_bool.text :

Résultats :
- Average time: 0.053545884788036346 (sec)
- Median time: 0.043419480323791504 (sec)



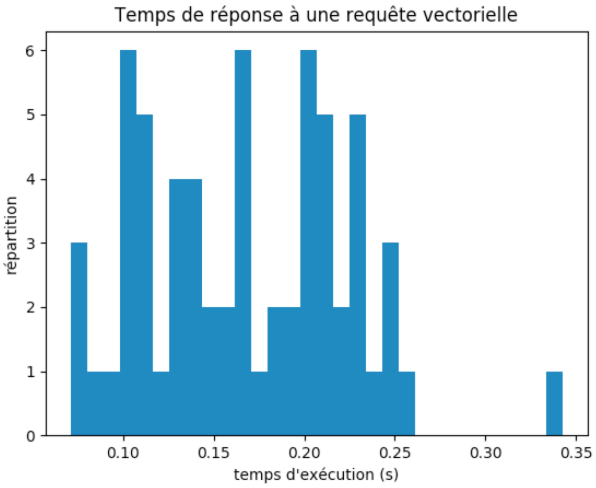


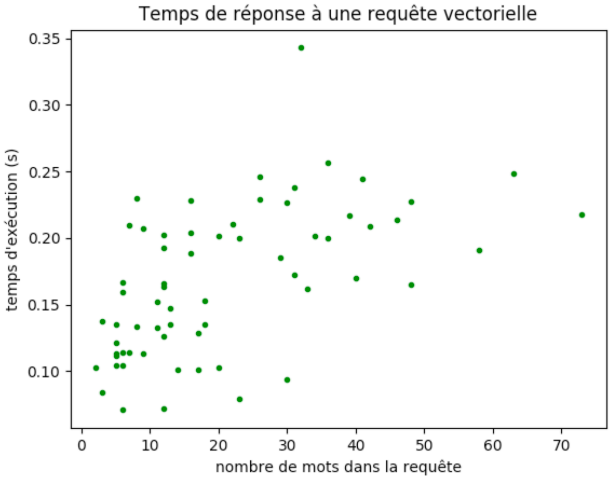
On observe pour les requêtes booléennes une corrélation relativement nette entre le nombre d'opérandes de la requête et le temps d'exécution

Temps de réponse à une requête vectorielle :

On teste le système de recherche vectoriel sur les 64 requêtes préalablement définies dans le fichier query.text :

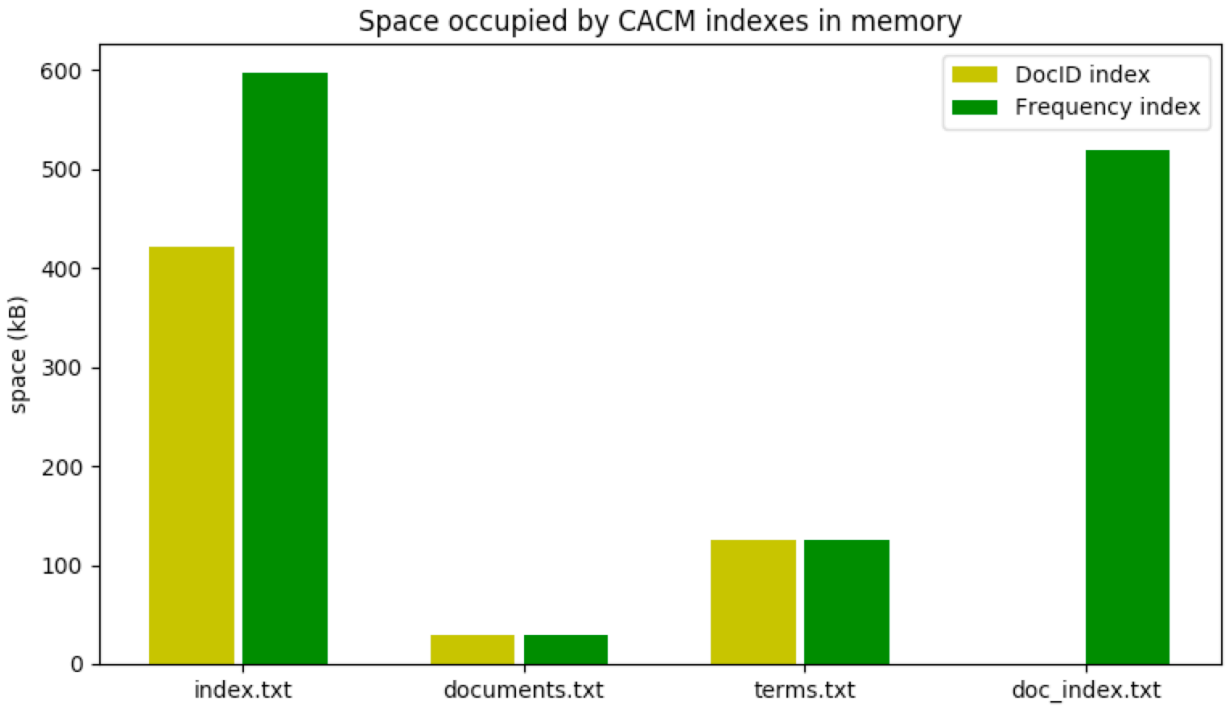
Résultats :
- Average time: 0.1674606166779995 (sec)
- Median time: 0.16536295413970947 (sec)





Les requêtes vectorielles sont plus longues à s'exécuter que les requêtes booléennes. Par ailleurs, on n'observe pas de corrélation claire entre le nombre de mots d'une requête et son temps d'exécution.

Occupation de l'espace disque par les différents index :



Espaces occupés en mémoires par les index de type DocID :
- documents.txt: 29823 (bytes)
- index.txt: 431491 (bytes)
- terms.txt: 128992 (bytes)

Espaces occupés en mémoire par les index de type Frequency :
- doc_index.txt: 531420 (bytes)
- documents.txt: 29823 (bytes)
- index.txt: 610795 (bytes)
- terms.txt: 128992 (bytes)

L'espace mémoire occupé est très important. Sans surprise, l'index frequency est plus volumineux en espace que l'index docID.

Mesures de pertinence

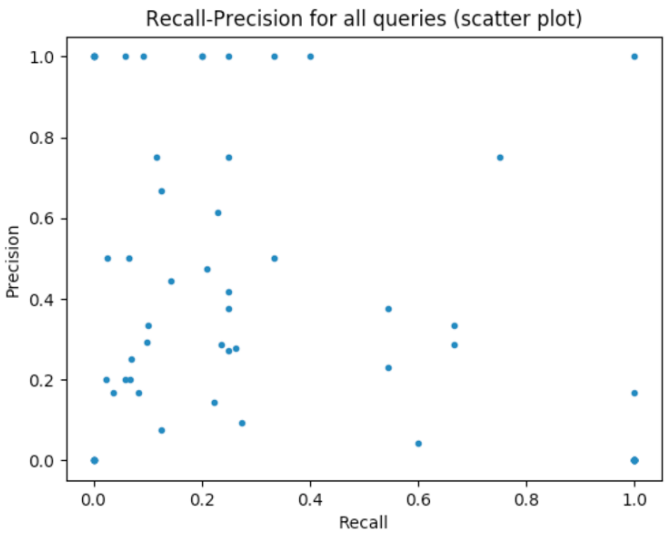
Pour le modèle booléen (sans classement) :

- Rappel-Précision

On mesure le rappel et la précision de façon globale (sur l'ensembles des résultats renvoyés) pour chacune des 64 requêtes du fichier *query_bool.text*.

Résultats :
Global average precision: 0.377155963077439
Global average recall: 0.36257277618744294

Ni le rappel, ni la précision ne sont très bons.

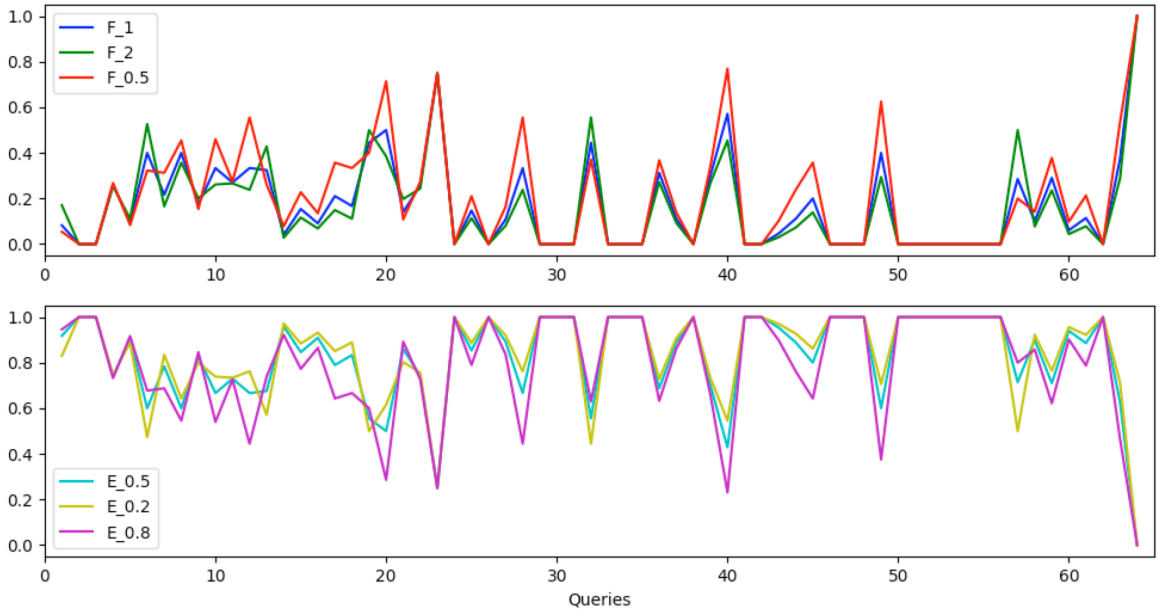


- E-mesure, F-mesure

A l'aide du rappel et de la précision, on calcule pour chaque requête les E-mesure et F-mesure avec différents paramètres apha (et donc différents bêta)

Résultats :
Average E-measure (for alpha=0.5): 0.8288119666797837
Average F1-measure: 0.17118803332021618
Average F2-measure: 0.1618746106021235
Average F0.5-measure: 0.20911473236096256

Voici le résultat détaillé pour les 64 requêtes. Les mesures Eα et Fβ se comlètent à 1 lorsque l'on a : $\alpha = 1/(\beta^2 + 1)$

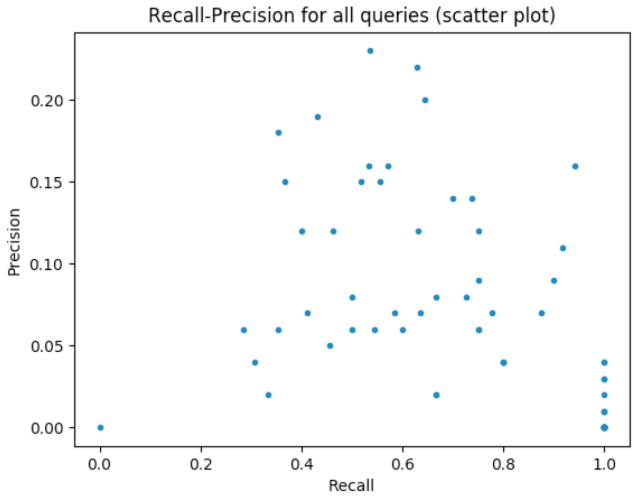


Pour le modèle vectoriel (avec classement) :

On reproduit d'abord les mêmes mesures globales que pour les requêtes booléennes pour comparer (mesures appliquées à l'ensemble des résultats, sachant que ces résultats sont tronqués à 100)

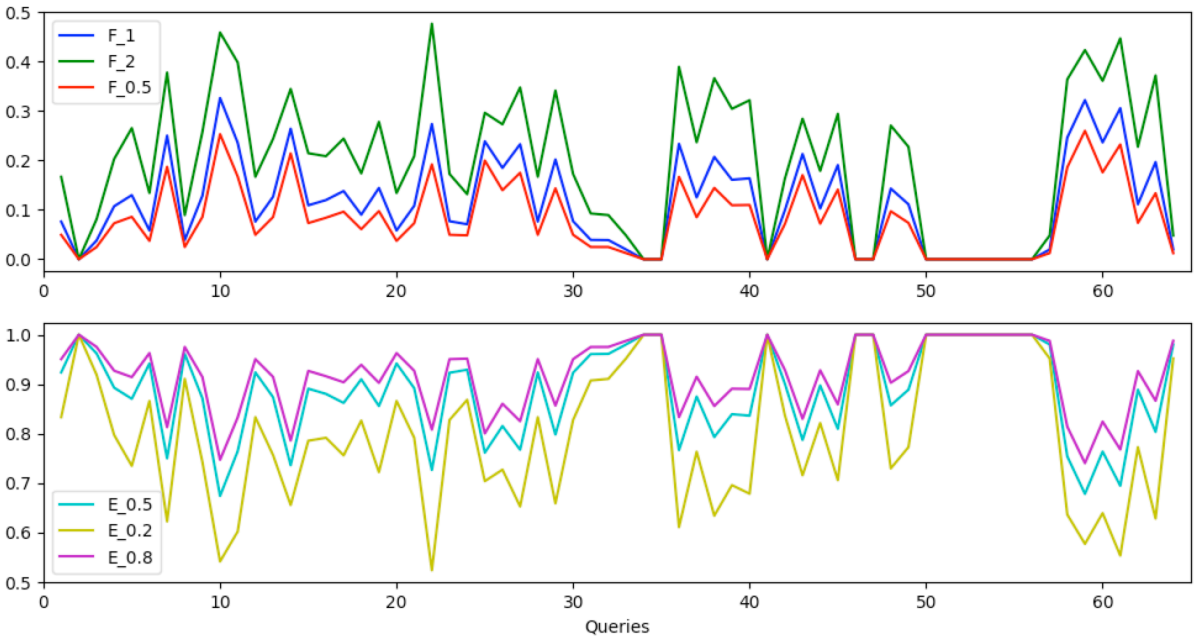
- Rappel-Précision

Global average precision: 0.06984375000000001
Global average recall: 0.7205591286178544



- E-mesure, F-mesure

Average E-measure (for alpha=0.5): 0.8850710923983752
Average F1-measure: 0.11492890760162489
Average F2-measure: 0.1964365966402771
Average F0.5-measure: 0.08265571590256003

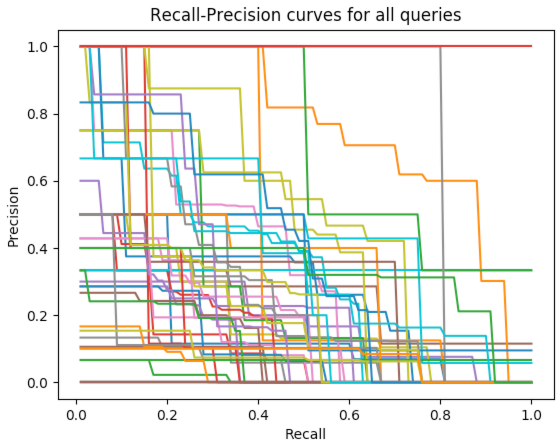


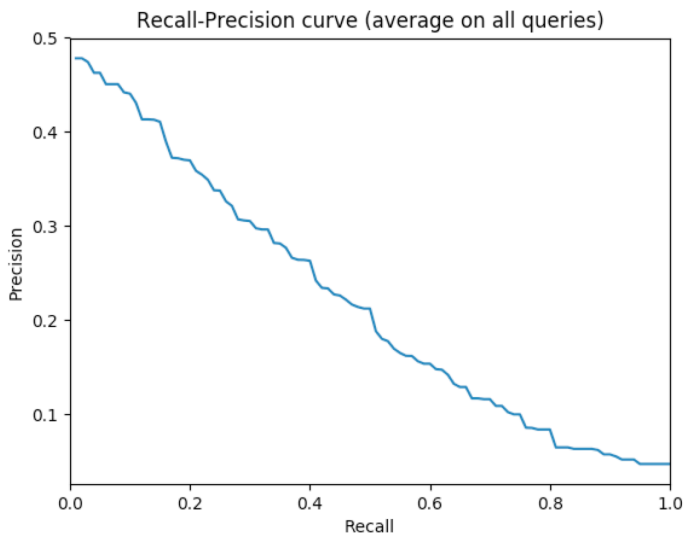
On observe que le rappel est globalement bien meilleur que pour les requêtes booléennes mais qu'à l'inverse la précision est beaucoup moins bonne (ce qui n'est pas étonnant sachant que les requêtes vectorielles renvoient beaucoup plus de résultats au total). C'est pourquoi on obtient une F-mesure toujours plus grande lorsque β augmente, car cela revient à privilégier le rappel.

Ensuite on effectue des mesures de pertinence spécifiques aux résultats avec classement

- Courbes Rappel-Précision

A gauche les courbes interpolées pour chacune des 64 requêtes, à droite la courbe moyennée :





- R-Précision moyenne

Average R-Precision: 0.25549007930853984

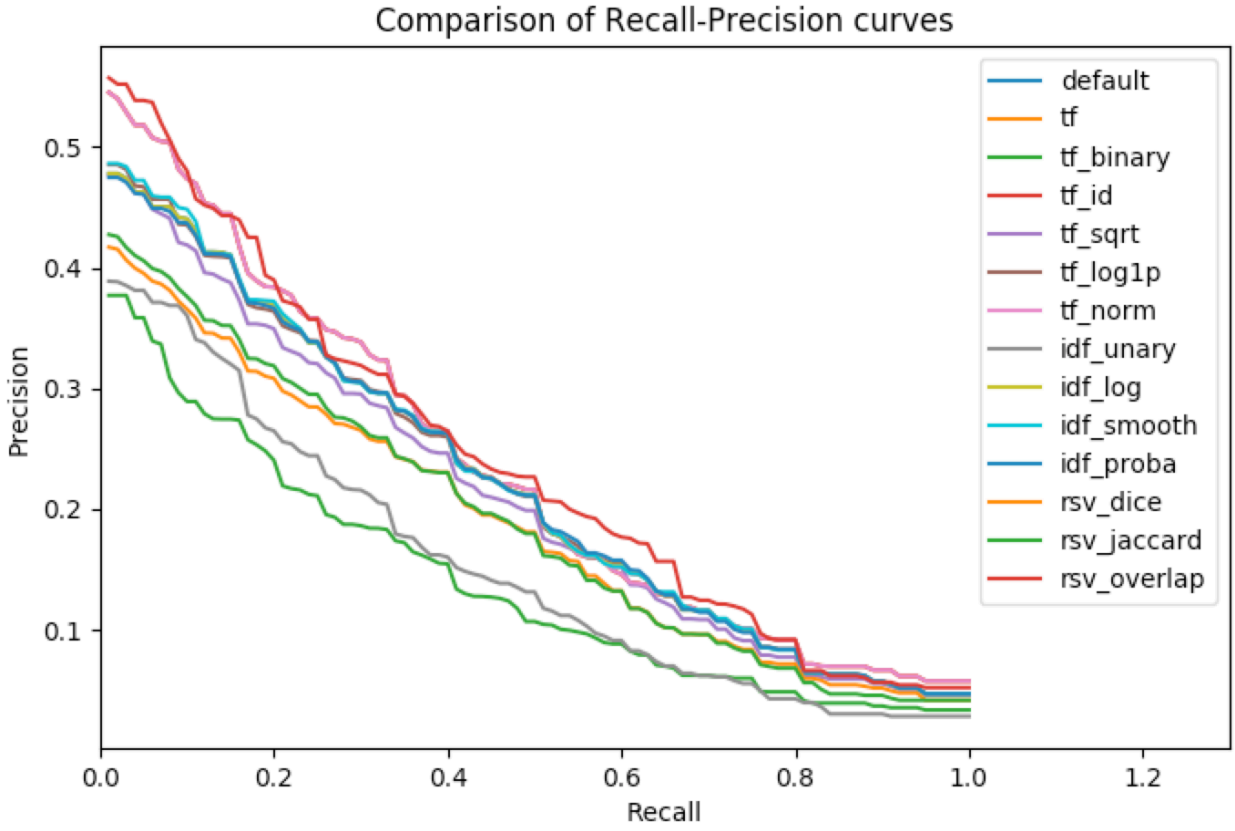
- Mean Average Precision

Mean Average Precision: 0.2203448291569245

Comparaison des différents modèles de pondération dans le système vectoriel :

Le modèle par défaut a été défini comme tf=tf_log, idf=idf et rsv=rsv_cos
Le « meilleur » modèle est celui qui donne la meilleure MAP dans ce cas précis et a été trouvé à la main en calculant pour toutes les possiblités. Dans le tableau ci-dessous, on fait varier seulement un paramêtre à la fois en gardant les deux autres par défaut.

	R-Precision	Mean Average Precision
default	0.255490	0.220345
FUNCTION TF:		
tf_log	default	default
tf	0.249084	0.234640
tf_binary	0.154283	0.140858
tf_id	0.249084	0.234640
tf_sqrt	0.230775	0.211099
tf_log1p	0.255188	0.220138
tf_norm	0.249084	0.234640
FUNCTION IDF:		
idf	default	default
idf_unary	0.178095	0.153894
idf_log	0.255490	0.220345
idf_smooth	0.251535	0.221507
idf_proba	0.252970	0.220256
FUNCTION RSV:		
rsv_cos	default	default
rsv_dice	0.203154	0.188718
rsv_jaccard	0.202235	0.190233
rsv_overlap	0.262539	0.240959
best	0.262539	0.240959



Rapide comparaison avec un modèle incluant une normalisation

Temps de calcul pour l'indexation :

Sans stemming
- DocID index: 2.674 (sec)
- Frequency index: 3.095 (sec)

Avec stemming
- DocID index: 3.846 (sec)
- Frequency index: 4.837 (sec)

Temps de réponse à une requête booléenne :

Sans stemming
- Average time: 0.053545884788036346 (sec)
- Median time: 0.043419480323791504 (sec)

Avec stemming
- Average time: 0.03949350118637085 (sec)
- Median time: 0.0298306941986084 (sec)

Temps de réponse à une requête vectorielle :

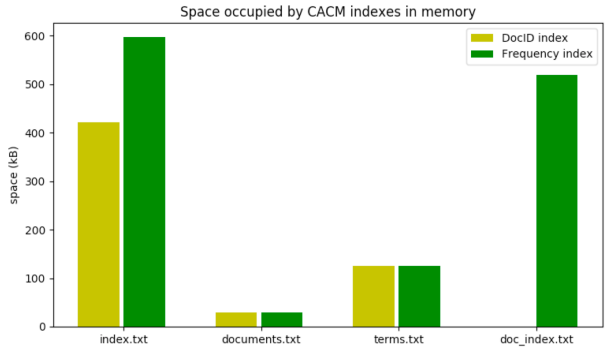
Sans stemming
- Average time: 0.1674606166779995 (sec)
- Median time: 0.16536295413970947 (sec)

Avec stemming
- Average time: 0.22213900461792946 (sec)
- Median time: 0.22826409339904785 (sec)

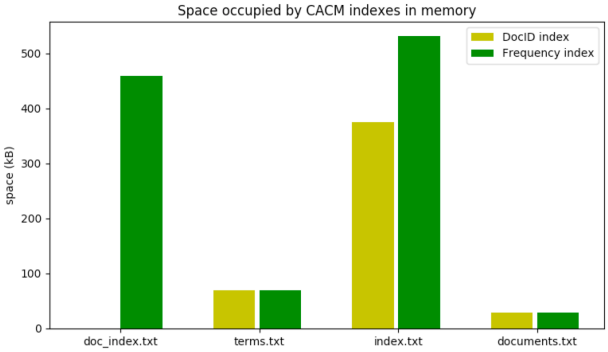
Le temps de réponse n'est pas systématiquement plus court. En effet, si le parcourt des index et des dictionnaires sera plus rapide du fait de la taille réduite de ces derniers, les moteurs de recherche (notemment le modèle vectoriel) sont également susceptibles de trouver plus de documents associés à un mot tronqué donc le traitement qui suivra sera plus long.

Occupation de l'espace disque par les différents index :

Sans stemming



Avec stemming



Pertinence du modèle booléen :

Sans stemming
Average precision: 0.377155963077439
Average recall: 0.36257277618744294

Average E0.5-measure: 0.8288119666797837
Average F1-measure: 0.17118803332021618
Average F2-measure: 0.1618746106021235
Average F0.5-measure: 0.2091147323609625

Avec stemming
Average precision: 0.2631000759786992
Average recall: 0.45770715530190753

Average E0.5-measure: 0.8162968461346211
Average F1-measure: 0.18370315386537908
Average F2-measure: 0.20079261745410654
Average F0.5-measure: 0.1959980840534209

Pertinence du modèle vectoriel :

Sans stemming
Average precision: 0.06984375000000001
Average recall: 0.7205591286178544

Average E0.5-measure: 0.8850710923983752
Average F1-measure: 0.11492890760162489
Average F2-measure: 0.1964365966402771
Average F0.5-measure: 0.0826557159025600

Average R-Precision: 0.25549007930853984
Mean Average Precision: 0.22034482915692

Avec stemming
Average precision: 0.08031249999999998
Average recall: 0.7665646629624026

Average E0.5-measure: 0.8693900097117698
Average F1-measure: 0.13060999028823014
Average F2-measure: 0.21970497503267228
Average F0.5-measure: 0.0947043388475795

Average R-Precision: 0.26635041603954734
Mean Average Precision: 0.25024775877569