

Labo 2 : Evaluation of results obtained from ElasticSearch

Peiris Sébastien & Raemy Mathis

Date: 17.11.2021

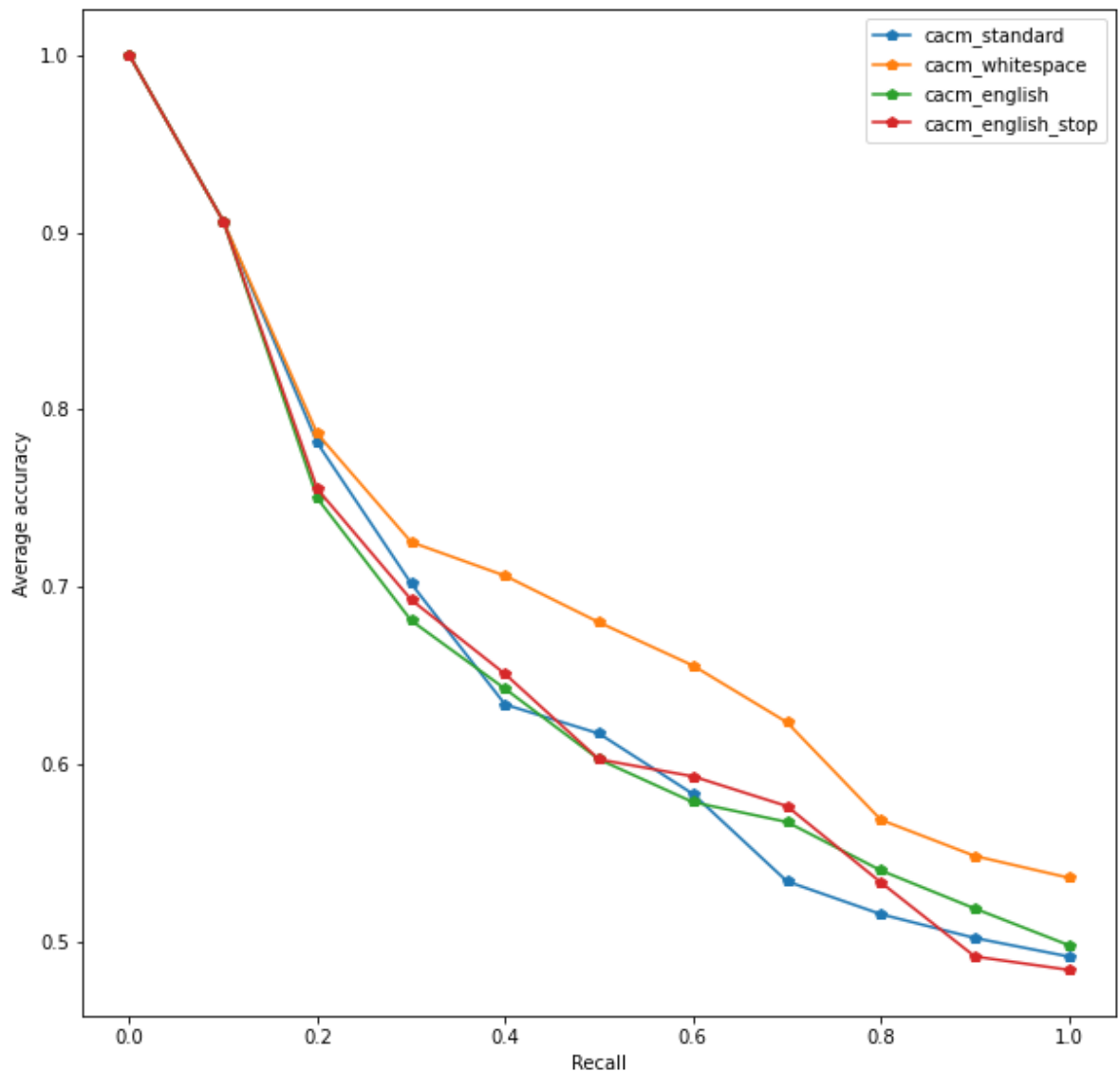
Introduction

Dans ce laboratoire, il nous a été demandé de réaliser une étude comparative de quatre analyseurs ("standard", "whitespace", "english" et "english avec la liste des mots communs") dans un système d'IR. Dans ce but, nous avons dans un premier temps calculé différentes métriques en fonction d'un jeu de requêtes donné.

Pour chaque requête, nous avons pu calculer la précision, le rappel ou encore la moyenne de ces métriques.

```
In [ ]: import matplotlib.pyplot as plt
```

```
In [ ]: # insert manually data
data = [
    {
        "index_name" : "cacm_standard",
        "avg_precision_at_recall_level" : [1.0,0.90625,0.7812500000000002,0.7500000000000002,0.7552083333333335,0.7500000000000002,0.7500000000000002,0.7500000000000002,0.7500000000000002,0.7500000000000002],
    },
    {
        "index_name" : "cacm_whitespace",
        "avg_precision_at_recall_level" : [1.0,0.90625,0.7864583333333336,0.7500000000000002,0.7552083333333335,0.7500000000000002,0.7500000000000002,0.7500000000000002,0.7500000000000002,0.7500000000000002],
    },
    {
        "index_name" : "cacm_english",
        "avg_precision_at_recall_level" : [1.0,0.90625,0.7500000000000002,0.7500000000000002,0.7500000000000002,0.7500000000000002,0.7500000000000002,0.7500000000000002,0.7500000000000002,0.7500000000000002],
    },
    {
        "index_name" : "cacm_english_stop",
        "avg_precision_at_recall_level" : [1.0,0.90625,0.7552083333333335,0.7500000000000002,0.7552083333333335,0.7500000000000002,0.7500000000000002,0.7500000000000002,0.7500000000000002,0.7500000000000002],
    }
]
for d in data:
    x = [j/10 for j in range(len(d["avg_precision_at_recall_level"]))]
    plt.plot(x,d["avg_precision_at_recall_level"], '-p', label=d["index_name"])
    plt.xlabel("Recall")
    plt.ylabel("Average accuracy")
    plt.rcParams["figure.figsize"] = (10/2.54,10/2.54)
plt.legend()
plt.show()
```



Conclusion

On peut voir que les 4 courbes ont globalement une forme similaire à ce que l'on peut attendre d'une courbe de **Recall-Precision**. En effet pour les 4 analyzers on voit une régression de la précision en fonction du rappel.

Nous avons également remarqué que la courbe **cacm_whitespace** se démarque légèrement vers le milieu de la fonction. Selon nous, elle devrait plutôt se démarquer vers le bas. En effet à cause des mots utilisés dans presque tous les textes en anglais (stop_words) on devrait obtenir un nombre total de document beaucoup plus grand. Comme le nombre de documents est plus élevé, cela devrait donc avoir tendance à faire **baissier** la précision et **augmenter** le recall.

On a également remarqué que les deux courbes **cacm_english** et **cacm_english_stop** se ressemblent car ils ont certainement une base de mots bannis similaire.