

Forestier Quentin & Herzig Melvyn

MAC – 18.12.2021

Évaluation des résultats obtenus pas Apache Lucene

Labo 4

# Question initiale

Parmi les analyseurs d’Apache Lucene, quel est le plus performant et le moins performant.

# Résultats

# Analyse

L’idéal pour un analyseur est d’être au plus proche du coin supérieur droit du graphique, c’est-à-dire, avoir une précision de 1 avec un rappel de 1. Cela signifierait que tous les documents retournés sont des documents pertinents.

La précision moyenne interpolée à onze points montre la précision maximale moyenne par pas de 10% de rappel. En d’autres termes, quel est la précision maximale moyenne lorsque 0% des documents pertinents ont été récupérés, puis 10%, puis 20% jusqu’à 100%.

## Plus mauvais analyseurs

L’analyseur avec les moins bonnes performances est le Whitespace Analyzer. Cet analyseur a le fonctionnement le plus basique des analyseurs de Lucene. Il se contente simplement de former des tokens en séparant le texte entre des [espaces blancs](https://lucene.apache.org/core/8_0_0/analyzers-common/org/apache/lucene/analysis/core/WhitespaceTokenizer.html). Il n’applique aucun stemming et conserve la casse. Ainsi, il faut comprendre que l’analyseur va valoriser certains termes alors que ça ne devrait pas être le cas, puisque certains de ces termes, comme des stop words, ne devraient pas influer sur le résultat. En conséquence, il a la précision la plus faible à chaque rappel. De ce fait, le White Space Analyzer est le moins bon analyseur suivi de près par le Standard Analyzer. Ce dernier, à la différence du Whitespace Analyzer, forme des tokens en se basant sur la grammaire et il est sensible à la casse. C’est un fonctionnement à peine plus élaboré que le Whitespace Analyzer, de ce fait il est légèrement plus performant que le Whitespace Analyzer mais moins bon que les meilleurs.

## Meilleurs analyseurs

Le meilleur analyseur est le English Custom Analyzer. C’est un analyseur spécialisé pour l’anglais, il applique le stemming, une liste de stopwords personnalisée et il n’est pas sensible à la casse. Ce type d’analyseur conserve le sens de base des tokens importants qui sont orthographiquement proches (p.ex. : Computers = computer). Ainsi, comme cet analyseur ne se concentre que sur les mots importants, il a la meilleure précision à chaque taux de rappel. Toutefois, il est suivi de près par le English Analyzer avec la liste de mot par défaut. Nous en déduisons que la liste de stopwords personnalisée était plus adaptée pour cette collection de document que la liste par défaut puisque c’est la seule différence entre ces deux analyseurs.

# Conclusions

Nous avons deux groupes de performances les analyseurs avec liste de stopwords (English Analyzers et English Custom Analyzer) et les analyseurs sans liste de stopwords (whitespace Analyzer et Standard Analyzer). Les analyseurs avec stopwords sont plus performants que les analyseurs qui n’en ont pas.

D’ailleurs, pour deux analyseurs les mêmes caractéristiques excepté une liste de stopwords divergente (English Analyzers et English Custom Analyzer), bien que proche, les deux analyseurs ont des résultats divergeant. Nous en concluons que l’un avait une liste de stopwords plus adapté à la collection, ce qui souligne l’importance d’une liste de stopwords judicieusement choisie.

En conclusions, les analyseurs proche du sens du texte ont un meilleure précision à chaque taux de rappel que les analyseurs « basiques ». Toutefois, si les English Analyzers ont un si bon résultats, c’est parce que l’ensemble des documents est en anglais. Si la collection était multilingue, les résultats auraient probablement été inversés.

Quoi qu’il en soit, l’écart entre meilleure et moins bon n’est pas extravagant, 10% de précision en moyenne. Au bout du compte, pour un rappel de 1, tout les analyseurs proposent des performances similaires, probablement parce que l’analyseur parfait n’existe pas.