Forestier Quentin & Herzig Melvyn

MAC – 22.11.2021

Indexing and Search with Apache LuceneProtocoles applicatifs

Labo 3



Table des matières

[Familiarisation avec Lucene 2](#_Toc88492277)

# Familiarisation avec Lucene

**Quels sont les types des *fields* dans l’index :**

* StringField
* LongPoint
* TextField

**Quels sont les caractéristiques des types précédents :**

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Indexing | Storage | tokenization |
| StringField | Oui | Oui | Non |
| LongPoint | Sous forme de long | Non | Non |
| TextField | Oui | Non | Oui |

**Est-ce que la démo de la ligne de commande retire les stopwords :**

Non, le queryParser est fait avec un StandardAnalyzer et son constructeur vide qui n’utilise aucun stop words.



Pour les appliquer il aurait fallu les lui passer à la construction.

**Est-ce que la démo de la ligne de commande utilise un filtre de stemming :**

Non, le queryParser est fait avec un StandardAnalyzer qui n’utilise pas le filtre PorterStemFilter.

Comme le prouve notre test, apples n’est pas transformé en apple. De plus, ils ne retournent pas les même documents.

Une image contenant texte

Description générée automatiquement

**Est-ce que la démo de la ligne de commande est non sensible à la casse ?**

Oui car le StandardAnalyzer utilise un LowerCaseFilter.

Une image contenant texte

Description générée automatiquement

Comme nous pouvons le voir, les deux requêtes retournent les mêmes documents.

**Est-ce important de faire le stemming avant ou après le retrait des stopwords ?**

D’une manière générale, nous estimons qu’il est plus sage de retirer les stopwords avant de stemmer. Dans le cas inverse, nous n’excluons pas qu’un mot stemmé deviennent un stopword et soit retiré de la requête.