Assignment 1: Indexation

Z534: Search, Fall 2015

Wensi Wang

Task 1:

1. How many documents are there in this corpus?

84474.

2. Why different fields are treated with different kinds of java class? i.e., StringField and

TextField are used for different fields in this example, why?

We use StringField to create index and tokenize a field as whole, such as field <DOCNO>, <BYLINE>, <DATELINE>, <HEAD> which do not have too much data in each field.

TextField will analyze, tokenize and then index a field, without considering term vectors, such as field between <TEXT> and </TEXT>.

Task 2:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Analyzer | Tokenization applied? | How many tokens are there for this field? | Stemming applied? | Stop words removed? | How many terms are there in the dictionary? |
| Keyword Analyzer | No | 84474 | No | No | 84061 |
| Simple Analyzer | Yes | 37436094 | No | No | 169981 |
| Stop Analyzer | Yes | 26322425 | No | Yes | 169948 |
| Standard Analyzer | Yes | 26755630 | No | Yes | 233384 |