Name: Akshay Naik
Email ID: aunaik@iu.edu

Z534
Search

Assignment 1

1. How many documents are there in this corpus?

Ans. There are in all 84474 documents in this corpus.

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?

Ans.

- StringField indexes entire string as a single token, that is, these fields are never tokenized
- StringField is generally used for fields like id, zip code and country where we have to index them as a single token
- TextField, on the other hand, tokenizes the string into different parts called as tokens
- TestField is generally used on body field that contains the bulk of a document's text

Comparison between different Analyzers

Analyzer	Tokenization applied?	How many tokens are there for this field?	Stemming applied?	Stop word removed?	How many terms are there in the dictionary
KeywordAnalyzer	NO	84474	NO	NO	84054
SimpleAnalyzer	YES	34843730	NO	NO	932081
StopAnalyzer	YES	25089642	NO	YES	932048
StandardAnalyzer	YES	25405918	NO	YES	1098687