Assignment Number 1

Z534 – Search

Part 1: Answers

1. Number of documents in the corpus are: **84474**
2. Different fields can be used for different purposes for indexing. Some fields can be used for indexing while some for tokenizing. We can use String field when we need to index the field rather than tokenizing like using the word as an id. Text field is indexed as well as tokenized and this field can be used for body text.

Part 2: Answers

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	84054
Simple Analyzer	Yes	34843730	No	No	932081
Stop Analyzer	Yes	25089642	No	Yes	932048
Standard Analyzer	Yes	25405918	No	Yes	1098687