Assignment 1

Question1:

Total Number of Documents in this corpus: 84474

Question2:

String field cannot have tokenization or analysis filters applied over them. Only an exact match will be provided.

However, Text Field has a tokenizer and analysis attached where indexed content is tokenized and in case there is no match, the tokenized content is matched upon to evaluate if the document can be added to the results.

String field is used store the document numbers. Hence an exact match is needed to index them. However, text field is used in match the actual text contained in the document and hence a partial match would suffice to match the text.

Results:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Analyzer | Tokenization | Tokens/Field | Stemming | Stop words | Number of terms in dictionary |
| Key Word Analyzer | No | 84474 | No | No | 84061 |
| Simple Analyzer | Yes | 3733014 | No | No | 169981 |
| Stop Analyzer | Yes | 26216475 | No | Yes | 169948 |
| Standard Analyzer | Yes | 26649680 | No | Yes | 233384 |