Assignment 1 Solutions:

Sanjana Pukalay(spukalay)

README:

generateIndex.java :

1. change path values for variable indexPath (set directory where you want index to be created)
2. change path values for variable docsPath (set the directory which has the corpus files)

indexComparison.java :

1. change path values for variable corpusPath (set the directory which has the corpus files)

1. *How many documents are there in this corpus?*

   Solution:

   Total number of documents in the corpus:364

   File output snippet `for` generateIndex.java

```
Total number of documents in the corpus:364
Number of documents containing the term "new" for field "TEXT": 129
Number of occurrences of "new" in the field "TEXT": 241
Size of the vocabulary for this field: 25014
Number of documents that have at least one term for this field: 363
Number of tokens for this field: 111489
Number of postings for this field: 80093

******* Printing Vocabulary-Start**********
0.008   0.047   0.43    0.5     00      01      0159    0199    03
*******Printing Vocabulary-End**********
```

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 ?*

   Solution:

   TextFields comes with a tokenizer and text analysis, it breaks the indexed content into separate tokens. There is no mandate for an exact match here, every individual word/token is matched separately and than decided whether the whole document is included in the response.

   Unlike TextFields, StringFields don't have any tokenization or analysis applied, and gives results for exact matches only. A StringField with analysis or filters applied, you can be implemented using a TextField with a KeywordTokenizer.

Solution 2)

Analyzer	Tokenization Applied?	No of Tokens	Stemming Applied?	Stop words removed	Terms in Dictionary	Unique Terms After Tokenization	Terms Afte Tokenizati
KeywordAnalyzer	No	1	No	No	150877	364	1
SimpleAnalyzer	Yes	152998	No	No	150877	143544	152998
StopAnalyzer	Yes	110515	No	Yes	150877	110514	110515
StandardAnalyzer	Yes	111489	No	Yes	150877	111488	111489

Code output snippet for indexComparison.java

```
total count>>150877
Number of tokens for analyzer Keyword count: 1
Number of tokens for analyzer Simple count: 152998
Number of tokens for analyzer Stop count: 110515
Number of tokens for analyzer Standard count: 111489
```