Z534:Assignment 1:

Aishwarya Dhage

adhage@iu.edu

Date:26 September 2017

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?**

   StringField stores data as a single string. TextField stores data as chunks i.e. tokens. In our example, we want docno to be stored as a single string so we are using StringField for DOCNO, for rest all fields i.e. DATALINE, BYLINE, HEAD and TEXT we want data to be stored as tokens and so we are using textfield.

Comparison Of different analyzers-

	tokenization	How many tokens are there for this field	Stemming applied	Stop words removed?	How many terms in dictionary
keyword	no	84474	no	no	84054
simple	yes	34843730	no	no	932081
stop	yes	25089642	no	yes	932048
standard	yes	25405918	no	yes	1098687