**Instruction for compiling and running Homework 6**

* Files: Zip file comes with a JAVA project exported directly from eclipse and contains following files:
  1. *Source code of Lucene Implementation in Java:* following files can be found under src folder
     1. *Classifier.*java (Classifier program)
     2. *ModelGenerator.*java (Used for generating the model from training data)
     3. *ValueComparator.java (For comparing and sorting dictionary)*
  2. *Textcat –* The input files containing training, test and dev data.

* Steps to run HW\_6 Assignment:

1. Extract the Archived Project and Navigate to HW6 folder then run below command:
   1. To run ModelGenerator program

> java -cp bin/ ModelGenerator “textcat/train”

* 1. To run Classifier program

> java -cp bin/ Classifier “model-file” “textcat/dev”

> java -cp bin/ Classifier “model-file” “textcat/test”

* Outputs:

1. *model-file*: (the language model generated on training data after ModelGenerator.java)
2. *neg\_to\_pos\_ratio:* A list of the **20 terms** with the highest (log) ratio of negative to positive weight.
3. *pos\_to\_neg\_ratio:* A list of the **20 terms** with the highest (log) ratio of positive to negative weight.
4. *prediction-file:* The prediction file for test data.
5. *prediction-file-dev:* The prediction file for dev data.
6. *Percentage of data classified correctly in dev*
   1. *Positive reviews- 74%*
   2. *Negative reviews- 88%*
7. *Reviews classification in test data*
   1. *Positive reviews- 84*
   2. *Negative reviews- 116*

**References**:

* Stackoverflow: <http://www.stackoverflow.com/>
* Book- <http://www.ntu.edu.sg/home/gaocong/papers/wpp095-yuan.pdf>
* Stanford slides- <https://web.stanford.edu/class/cs124/lec/naivebayes.pdf>

Note:

1. Project done in a team consisting of Abhinav Maurya and Kartik Chanana
2. We have not considered the empty space while generating the model on training data.