# Group 18:

1. Python version =2.7.11+
2. Genism python library should be installed on the system.
3. The data should be in this format as shown below:

   ***Baller to Batsman(comma)  run(comma) text commentary for this ball.***

   Example corpus:

   *Nehra to Charles, 1 run, short of a length on leg stump, Charles stays in his crease and turns the ball off his pads behind square.*
   *Nehra to Gayle, no run, pitches short of a length outside off and seams away from the left-hander, Gayle lets it go.*
   *Nehra to Gayle, no run, Gayle has a half-waft at a fuller ball that shapes away from the left-hander, inside-edge past the stumps!*
   *Nehra to Gayle, no run, short ball, it rises on the batsman, Gayle shapes to pull but bails out of the shot.*
   *Nehra to Gayle, FOUR, too short and this time it sits up, Gayle stays in his crease and pulls to the midwicket boundary, no fuss.*
   *Nehra to Gayle, 1 run, a fuller ball on middle and off, Gayle tucks it through square leg.*

4. Corpus file:
   The text commentaries for India-West Indies T-20 match was collected from :
   http://www.espncricinfo.com/icc-world-twenty20-2016/engine/match/951371.html?innings=1;view=commentary

   They are present in : ”IndiaBalling.txt” and “IndiaBatting.txt” in the Data folder.

5. **Steps to Execute the Code:**
   **I $python G18/code/preprocessor.py –f ../Data/IndiaBatting.txt**

   It does the preprocessing and generates “intermediate.txt”, which will be used in ‘’myProg.py”.

   To run “myProg.py”:

   **ii $ python myProg.py**

It has three functions:

I “saveModelParams()” reads the “intermediate.txt” and generates the Document-Term matrix which contains the Tf-Idf scores. This is the model stored in the file “modelParams.txt”.

ii “loadModelParams()” load this model and perform K-means clustering on it.
It then saves the obtained clusters in file “clustersWithPointId.txt”. IT also prints SSE errors corresponding to that clustering.

iii “LDA” now reads this file “clustersWithPointId.txt” and performs the summarization.
The obtained result is present in “result.txt”.
The format of result.txt is:

*Clusterid*
*Text commentaries in this cluster*
*topicid : score for the word1*word1 +……+ score for the wordi*wordi*
---