**Assignment-5**

The data is downloaded from the Internet using the link given in the assignment.

There are 2 files docword.nips.txt and vocab.nip.txt.

A Java class (Nips.java) is written which generates the text files based on the requirements given in the question. This java code generates 1500 txt files located in “output” folder. Download the files given in the submission folder to your local computer. Change the paths given in the Java class and specify the output folder to run the program.

**Input File to Java Class:** docword.nips.txt, vocab.nip.txt

**Output Folder:** output

These output files are given as input to Mahout for performing LDA.

**Important Notes:**

1. Out of the 1500 files generated, only 15 have been used as an input to mahout because of slower performance of system.
2. Command# 12 and 13 are run in mahout distribution-0.9 where as other commands are run in mahout distribution-0.7

The following commands are then executed in sequence to perform LDA using Mahout.

// Start Hadoop

1. $cd $HADOOP\_HOME
2. $./bin/start-all.sh

// Make assignment3 folder in Hadoop directory

1. $./bin/hadoop fs –mkdir assignment5

// Copy the input folder to assignment3

1. $/bin/hadoop fs -put /Users/himansubadhai/Documents/input/assignment5/input assignment5

// Create sequence directory in Hadoop using mahout

1. $cd MAHOUT\_HOME
2. $./bin/mahout seqdirectory -i assignment5/input -o assignment5/out-seqdir -c UTF-8 -chunk 5

// Create vector in Hadoop using mahout

1. $./bin/mahout seq2sparse -i assignment5/out-seqdir -o assignment5/vectors -wt tf

// Use rowid to create matrix

1. $cd $MAHOUT\_HOME
2. $./bin/mahout rowid -i assignment5/vectors/tf-vectors -o assignment5/matrix

// Move docIndex to assignment5 folder

1. $./bin/hadoop dfs -mv assignment5/matrix/docIndex assignment5/docIndex

// Perform LDA using Mahout

1. $./bin/mahout cvb -i assignment5/matrix/matrix -o assignment5/LDA-out -dict assignment5/vectors/dictionary.file-0 -dt assignment/LDA-topics -mt assignment5/LDA-topic-model -k 10 -x 20 –ow

// Vector Dump to generate output files

1. ./bin/mahout vectordump -i assignment5/LDA-out -d assignment5/vectors/dictionary.file-0 -dt sequencefile -p true -o /Users/himansubadhai/Documents/input/assignment5/term-topic.txt -sort assignment5/LDA-out -vs 10
2. ./bin/mahout vectordump -i assignment5/LDA-topics -d assignment5/vectors/dictionary.file-0 -dt sequencefile -p true -o /Users/himansubadhai/Documents/input/assignment5/doc-topic.txt -sort assignment5/LDA-topics -vs 10

**Final Output:**

doc-topic.txt: gives top 10 words for each document

term-topic.txt: gives top 10 topics and top 10 words in each topic