**CS5560 Knowledge Discovery Management**

**LAB ASSIGNMENT #5**

**Name: Lava Kumar Surparaju**

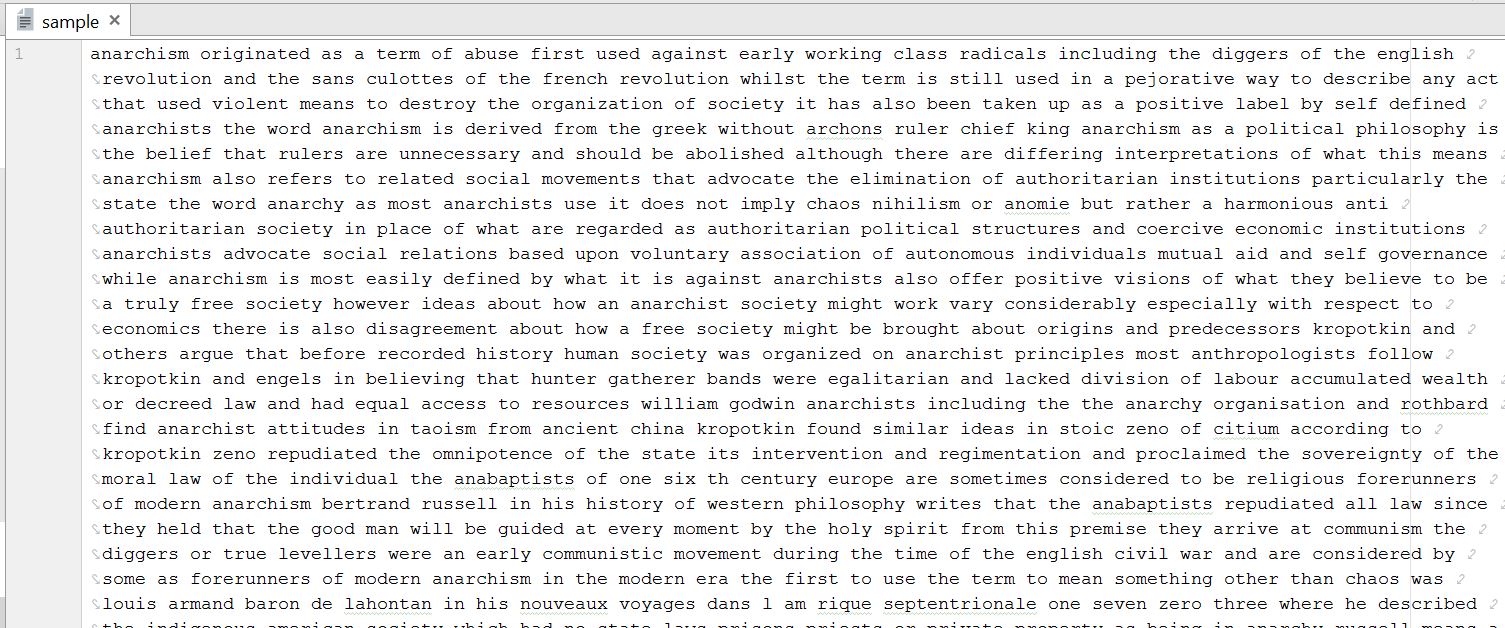
**Class ID: 27**

**Lab 5 Assignment:**

1. ***In Class Question***

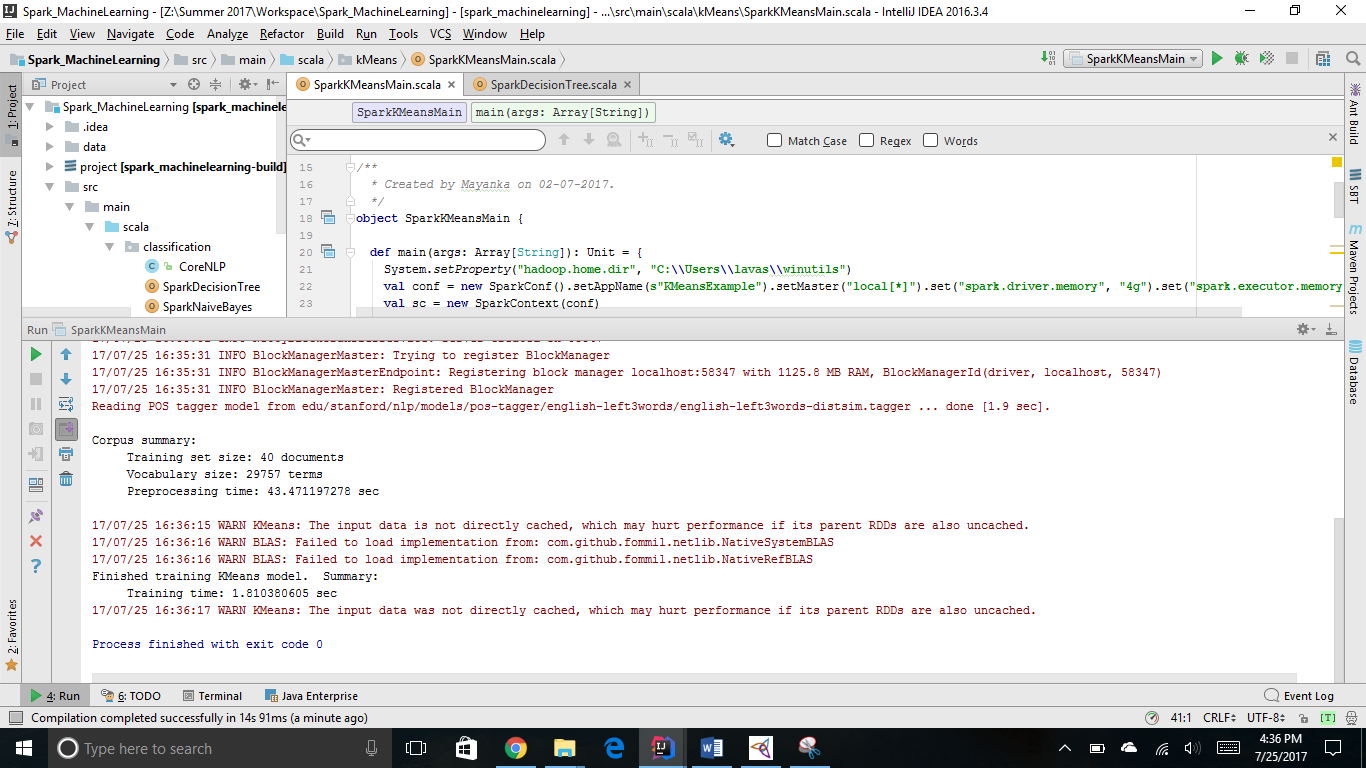
**Write a simple spark program to conduct clustering and classification techniques with the datasets (questions and answers) as part of your question answering system.**

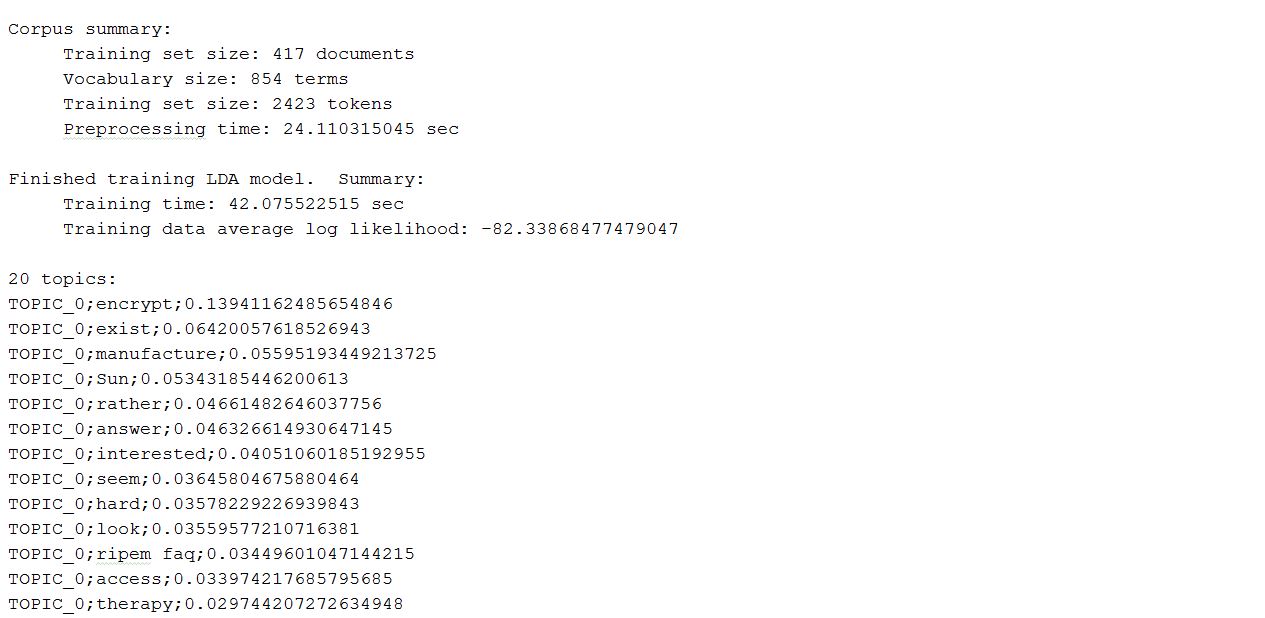
* 1. Dataset: Your own dataset



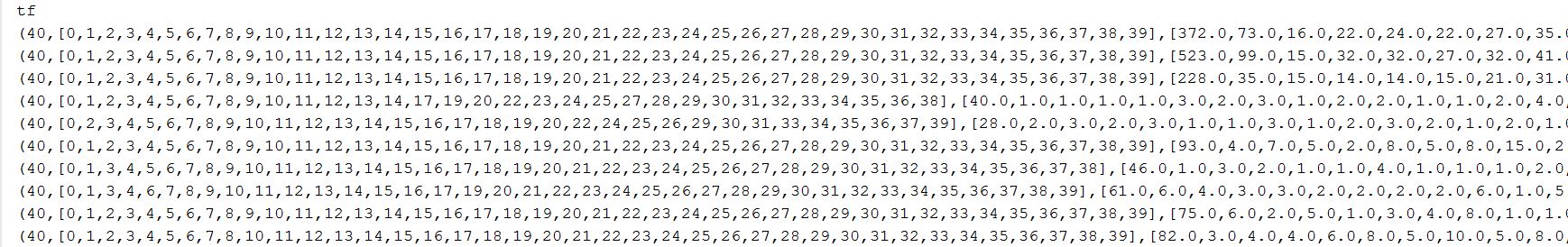
* 1. Conduct and compare K-Means vs LDA

The training time taken is way less in the case of K-means than in LDA.

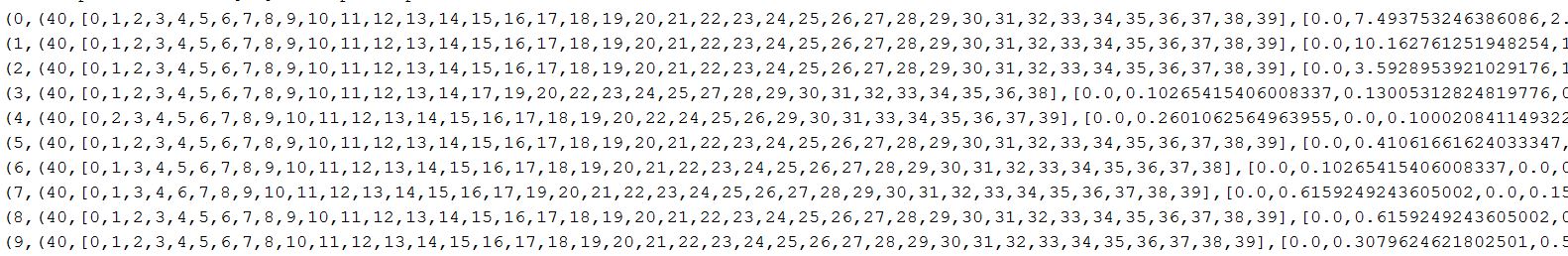




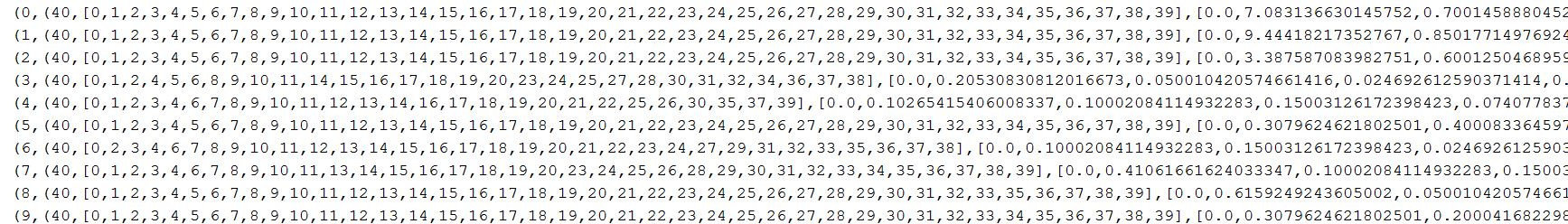
* 1. Conduct and compare Different Feature Vectors
     1. FV1: Data => NLP => TF==> Feature Vector



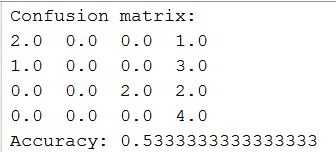
1. FV2: Data => NLP => TFIDF => Feature Vector

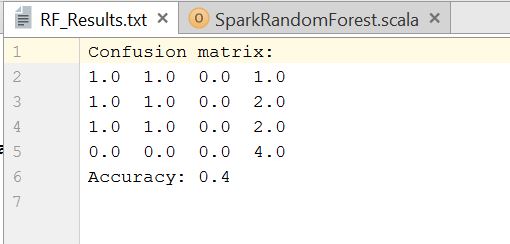


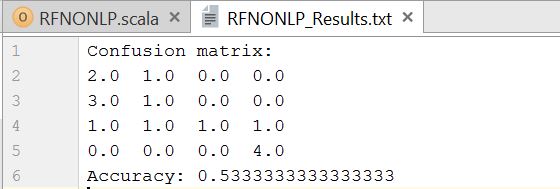
1. FV3: Data => TFIDF => Feature Vector

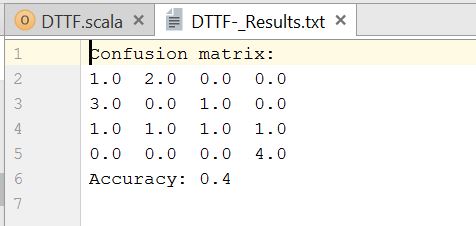


1. Conduct and compare the following Classification Algorithms using different feature vectors (FV1, FV2, FV3)







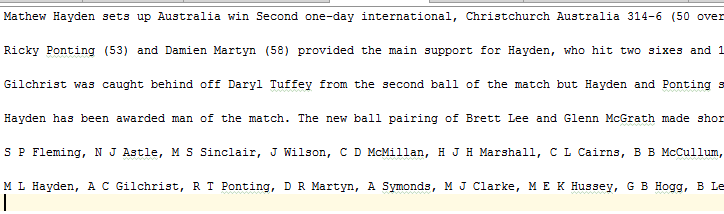


1. ***Take home Question***

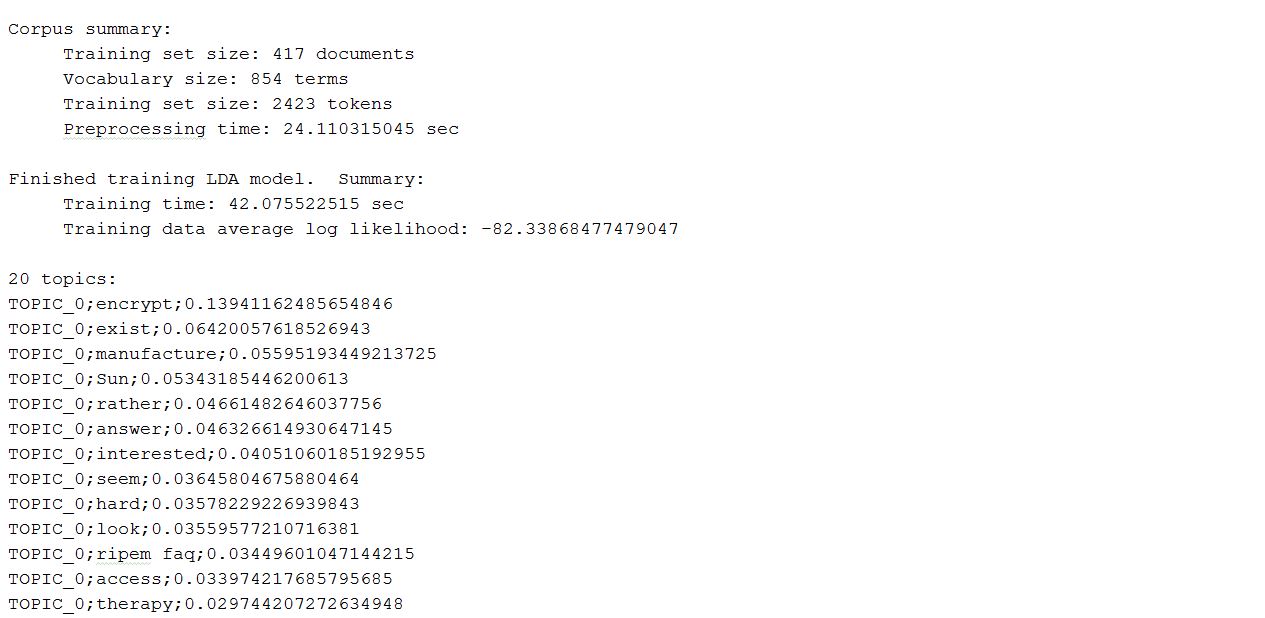
**Enhance your question answering system (from lab 4) using the classification/clustering techniques with your datasets (questions and answers) by following steps in task(1).**

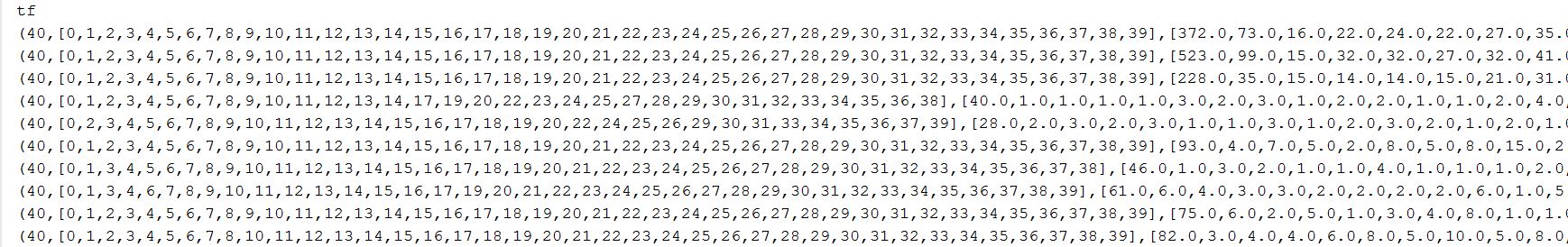
* 1. Using two datasets
     1. **Your own data**
     2. **Yahoo! Answer data (**The [training data](https://umkc.box.com/s/pph3fjv865wostjfc4zbnnzxl16m9r41) contains 2,698 questions, already labeled with one of the following 7 categories. The [test data](https://umkc.box.com/s/83yzy207p1u90mvhwft1bjpno18d70qf) contains 1,874 questions that are unlabeled). Questions into one of the following 7 categories:
        1. Business&Finance
        2. Computers&Internet
        3. Entertainment&Music
        4. Family&Relationships
        5. Education&Reference
        6. Health
        7. Science&Mathematics
  2. Report on the results on clustering your questions/answers datasets: K-Means vs LDA
  3. Report on the results on the different Feature Vectors used in classification on your questions/answers datasets
     1. FV1: Data => NLP => TF==> Feature Vector
     2. FV2: Data => NLP => TFIDF => Feature Vector
     3. FV3: Data => TFIDF => Feature Vector
  4. Report on the results and comparative evaluation on the following Classification algorithms using different feature vectors (FV1, FV2, FV3)
     1. Random Forest
     2. Naïve Bayes
     3. Decision Tree
  5. Report your insights on each of the task.

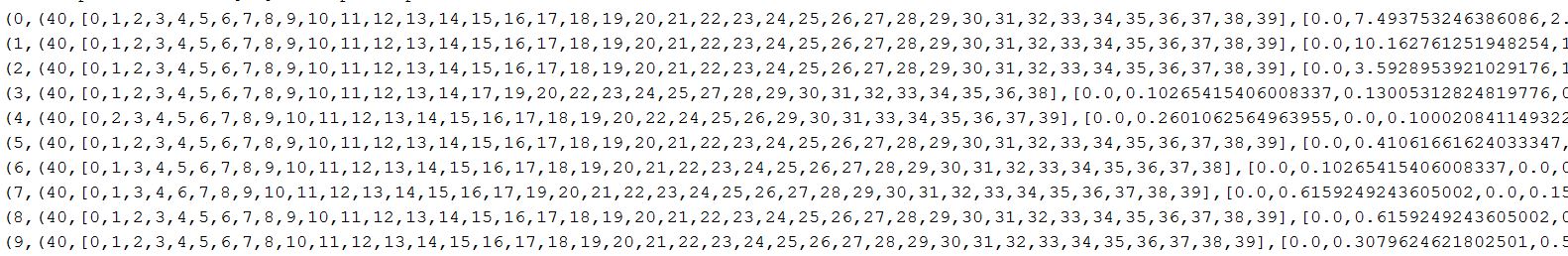
**Input:**

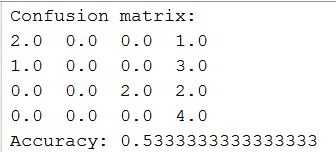
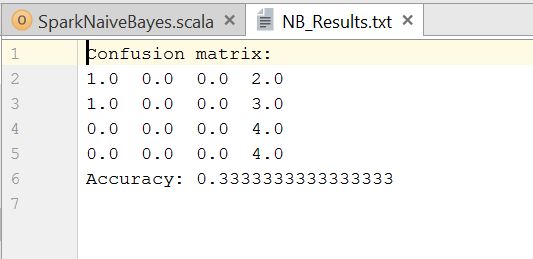
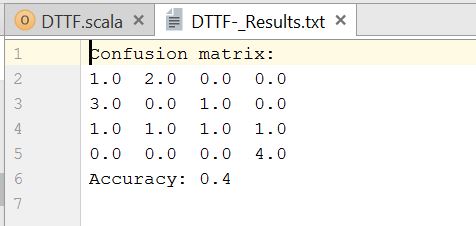


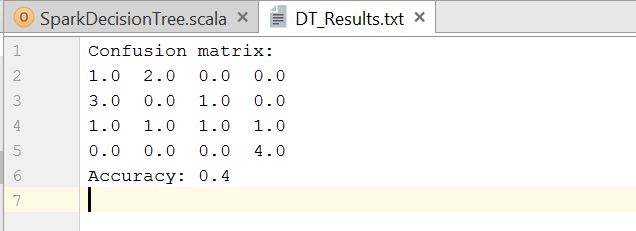
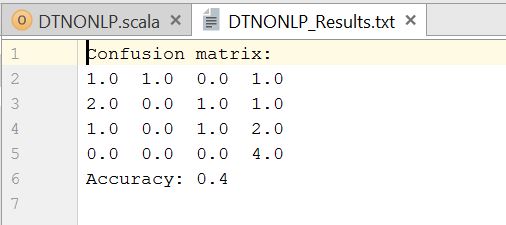
**Output:**







**Questions & Answers:**

