Assignment 2: Machine Learning

MSc Big Data Analytics and MSc Computer Science
Ramakrishna Mission Vivekananda Educational and Research Institute

You have to work on a problem of opinion mining. The description about the data and the tasks are given below. Subsequently, the guidelines of submission are provided. **The full marks of this assignment is 50**.

Problem Statement

The data set contains opinions of 38 different people on the following question:

"*What qualities do you think are necessary to be the prime minister of India?*"

The data set is already supplied to you in an excel file. Each row from second row onward in the excel file indicates the comment of a particular person. The task is to find the significant comments from this data. You have to use Wordnet, an English lexical database to accomplish the task. Some experts had already reviewed the data set and found significant qualities manually from the data. The human coding is also sent to you. The objective is to analyze the performance of your method using the human coding. You may discuss whether the method using Wordnet is able to identify the significant opinions or not.

Write a report in PDF that should properly explain the problem statement, the proposed methodology and analysis and conclusion about the results. The results should be composed in an excel file by following the sample output already sent to you. You must send the report, the code and the result file in the same email.

Create a zip file by combining these three files and the name of the zip file should be "*firstname_lastname_assignment2.zip*".

Methodology

A few suggestions regarding the methodology is given below. **However, any other efficient trick is highly appreciated.**

1. Form the vocabulary of the data set. You may consider only the nouns and adjectives and simply ignore the other words, since the objective is to find the qualities of the PM. Find the frequency of each word and order the words in order of decreasing frequency.

2. Find the synonyms for each word (using Wordnet synsets) and include them in the vocabulary.

3. Find the derivationally related forms of each word, if any and add them to the vocabulary.

4. Find similar words from the vocabulary and merge them to form a cluster. Repeat this step until no more merge takes place. You can also fix a particular criterion to stop merging clusters after certain steps.
5. Recalculate the frequency of each cluster and order them in decreasing order of frequency.

You may consider only the top frequency clusters. Compare the words in these clusters with the human coding and subsequently analyze the results.

Submission Guidelines

Write a report to describe the given tasks. The report should be submitted in PDF. The report should contain the following information:

A) The first page should contain the following information:

Name

Registration No/ Roll No

University Name

Program Name - BDA/CS

Problem Release Date: 09/04/2018

Date of Submission:

B) From the second page onward you have to provide the following information.

- A suitable title
- **Problem Statement:** This section should address the problem. Describe about the data set here. You may also discuss about Wordnet in few sentences.
- **Methodology:** Explain about the methodologies
- **Analysis of Results and Conclusion:** Significant findings and scopes of future works may be explained here in few sentences.

Other Relevant Information

- You may send the codes in separate files along with the report, but this is not compulsory.
- **Deadline of the submission is 21/04/2018, 23:59 IST**.
- Multiple submissions are allowed within the deadline, but only the last submission will be graded. You should send the submission to welcometanmay@gmail.com.
- **For everyday that your submission is late your score gets multiplied by 0.8.**