041eb01[1]  *University of Louisville*

*Computer Engineering and Computer Science Department*

DATA MINING – CECS 632-01

Fall 2019

==============================================================

ASSIGNMENT: Project 1

In this project you will gather and explore your own data from twitter. The project has three parts. Part 1 requires you to select opposing topics that you are interested in, and start gathering data using provided twitter data tools. In Part 2 you will dive into your collected data, understand the data structure, understand data types, preprocess and clean the data if needed. In Part 3 you will summarize what you can find about your topics from the tweets you collected.

**Part 1: Selecting Topics and Gathering Data**

**1) Select a topic with two opposing opinions**

Select two popular topics that are currently related and of opposing view. For instance, you can have topics “pro gun-control” versus “anti gun-control”, or “raise tax for the rich” versus “reduce tax for the rich”. There are some requirements for your topics:

1. No topics that promote racism, anti-Semitism and obscenity. If you are not sure, email the teaching assistant (mehmetakif.gulum@louisville.edu) and ask if the topics are OK.

2. Topics must be popular enough to generate enough data. If after two hours you only obtained 100 tweets, then the topic might not be popular enough. You need topics with at least 1000 tweets per hour. A quick way to tell is if after an hour your data file is at least 2-3MB large, then the topic is OK.

**2) Gathering Twitter Data**

A tutorial file “twitter.docx” is provided that explains step-by-step how to use the twitter data gathering tool. The tool uses keyword to filter real time tweets and download them into a data file. You therefore need to think of some keywords for your two topics so that you can collect the data. There are some tips when gathering the data:

1. Have a test run for an hour or so to see if your topics are popular enough.

2. You want to collect two opposing topics separately so that you can find different patterns later.

3. You want to aim for at least 100MB worth of total data to have a meaningful project discussion.

4. You might want to gather data in a fixed interval, such as 2 hours every two days, so that you can have discussion on temporal aspect of your topics.

**Part 2: Exploration**

Now that you have the twitter dataset, you need to look at what the dataset structure is. Try to understand what each pre-defined feature represents and see if anything requires preprocessing. Also look into unstructured part of the data (e.g. the actual tweets) and think about any feature you need to extract. Answer the following questions in your project report:

1. What feature has missing data? How much data is missing? How do you deal with missing data?

2. What feature is potentially noisy? What level of ambiguity is the data? How do you deal with such noise?

3. What feature might need normalization? What feature might need discretization?

4. Are there any feature that needs to be created from unstructured data?

4. Are there any other preprocessing steps needed for the data?

**Part 3: Discussing the Data**

After the data is preprocessed, you should try to have three different types of visualization about some patterns discovered about your topics. Patterns should be spatial, temporal, demographic aspect of the data. Visualization using tweet text summary is also OK, but advanced text mining is not required. Any visualization tool is OK, such as excel. The project is not graded on tools used but on the discussion. Make sure you explain what each visualization shows in detail. Answer the following questions for each visualization:

1. Why is this visualization created?

2. How is it created?

3. What does each element of the visualization mean?

4. What pattern does this visualization show?

5. What does this pattern say about the characteristics of the two opposing opinions?

6. What other keyword can you find among tweets that say about your opinions?

Please read “Required Structure for the Project 1 Documentation”, “Required Structure for the Test Data Result” and “Required Materials for Project 1 Submission” below.

NOTE: *Many software tools give large number of output results and graphical interpretations (tens of pages of graphs, tables, diagrams!). Don’t use them in your report if you don’t understand these values and graphs, and if you don’t have enough knowledge to make interpretation and give appropriate discussion about these results. Just attachment to the project of these kind of ”software listings” without interpretation will have negative influence on the final project grade.*

*Interpretation and your analysis should be explained in detail, rather than just listing the techniques and their results.*

==============================================================

Required Structure for the Project 1 Report

1. *Cover page*: Course, title of the project, student name, and date.
2. *Abstract*: Short description of the problem and the solution given in the project applied techniques and obtained results (one page max).

## **Project:** *TITLE*

3.1 **Problem description**. State what the selected topic and opinions are and why they are important. Describe the keywords you used for data gathering, and why these keywords are related to the topics

3.2 **Data Gathering**. Explain in detail how the data is gathered, including time interval and total amount of data obtaind

3.3 **Software tools** used for predictive data mining: Description of a tool or tools that will be used (if comparative analysis was made). Give a short explanation of applied algorithms. Make emphasis on the parameters you will tune during the mining process.

3.4 **Preprocessing**:

a. Answer in detail required questions in part 2

b. Explain in detail how preprocessing is carried out. For instance, explain how missing data is handled

3.5 **Visualizing patterns:**

a. Have at least 3 visualizations.

b. Explain how each visualization is created.

c. Answer in detail required question about each visualization.

3.6 **Conclusion**: Do your findings reflect what news says or what your real-life experiences tells you? What can you learn from your findings?

***Appendix:***

Example citation format:

MLA format: Fayyad, Usama M. "Data mining and knowledge discovery: Making sense out of data." *IEEE Expert: Intelligent Systems and Their Applications* 11.5 (1996): 20-25.

APA format: Fayyad, U. M. (1996). Data mining and knowledge discovery: Making sense out of data. *IEEE Expert: Intelligent Systems and Their Applications*, *11*(5), 20-25.

Harvard format: Fayyad, U.M., 1996. Data mining and knowledge discovery: Making sense out of data. *IEEE Expert: Intelligent Systems and Their Applications*, *11*(5), pp.20-25.

===============================================================

**Due data for Project 1:**