041eb01[1]  *University of Louisville*

*Computer Science and Engineering Department*

DATA MINING – CSE 632

Fall 2020

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

ASSIGNMENT: Project 1

In this project you will explore how topics have changed through time by gathering your own data on twitter and comparing it with historical data. You will practice multiple preprocessing technique on real-world twitter dataset. The project has three parts. Part 1 requires you to explore provided twitter data. In Part 2 you will collect your own data and perform preprocessing. In Part 3 you will summarize what you can find about how the data evolved over time.

**Part 1: Choose a Topic and Explore Data**

**1) Choose a Topic between “Coronavirus”, “Presidential Election”, or “Online Games”**

You can choose one of three topics for this project:

**a) “Coronavirus”**

**b) “Presidential Election”**

**c) “Online Games”**

In “Coronavirus”, you will be using the historical data “coronavirus.txt”, which is twitter data collected in July with keywords “coronavirus” and “covid-19”.

In “Presidential Election”, you will be using the historical data “election.txt”, which is twitter data collected in July with keywords “presidential election”, “2020 election”, “Trump re-election” and “Biden 2020”.

In “Online Games”, you will be using the historical data “game.txt”, which is twitter data collected in July with keywords “Online games”, “MOBA” (Massive Online Battle Arena), and “MMORPG” (Massively multiplayer online role-playing game).

Please state clearly in your report which topic you chose and why you choose it.

**2) Gather your own data**

A tutorial file “Tutorial on obtaining tweets.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. Once the data is downloaded, you can use the “twitterparser.py” to parse your twitter data into a more structured dataset. Since we are comparing current data with historical data, it might be easier for your analysis if you use the same keywords as the historical data (See Part 1.1). However, if you feel the keywords provided is not adequate, you can collect data based on your own keyword. However, if you do this, please explain what keywords you have used and why you have selected them. Also, later in your discussion you need to always remember that your dataset is collected differently than provided dataset. This might impact some of your analysis and discussion.

In your report, please explain in detail how your data is gathered, with discussions on:

1) The dates and times of when you collected your data

2) The keywords you used for data collection

3) The size of your data.

**3) Explore Provided Data and Your Gathered Data**

Another dataset previously gathered by us is provided datasets are under “Assignments->Porject 1” on blackboard. 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. Discussion the following aspects of the data 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?

**4) Hypothesis of What Has Changed Through Time**

Once you are familiar with the structure and contend of the data, **you need to state in your project three aspects of the data that you think might have changed between June and now**. For example, for coronavirus you can hypothesize that the tweet location about the disease has changed because there are now new hot-spot for the pandemic. For presidential election, you can hypothesize that the frequency of the tweets about election is higher than June because we are closer to the election. The hypothesis doesn’t have to be correct, since you will confirm it, or disprove it, with your own data gathering later in the project.

Write one paragraph for each hypothesis of change you decided to explore. State what your hypothesis is and explain why you think this hypothesis is important in each paragraph.

**5) Preprocessing the Data**

This is the most important part of the project. **You will perform at least 5 preprocessing techniques on both your dataset and the historical dataset**. Do not randomly choose any techniques to perform. Keep your end goal of visualizing your hypothesis in mind. Your preprocessing should help you to visualize and discuss your hypothesis later.

For each preprocessing technique, discuss the following in your project, each point should have at least one paragraph of discussion:

1. Explain what technique is used and how the technique works.

2. Detailed explanation on how the technique is applied. If the technique requires parameters, explain what the parameters are and how they are chosen.

3. Show and discuss the result of the technique. Do not copy and paste raw results or screen capture. Summarize your results into figures and explain what the figure say about your preprocessing results.

4. Explain how this technique helps you discuss your hypothesis later

**Part 2: 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 or Tableau Public . The project is not graded on tools used but on the discussion. Make sure you explain what each visualization shows in detail. Discuss the following aspects 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 changes between current data and historical data?

6. Is your hypothesis correct or not based on this visualization.

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 topic you have chosen and why you have chosen it.

3.2 **Data Gathering.** Explain how you gathered you own data. Follow instruction in part 1.2

3.3 **Data Exploration**. Exploring provided historical data. Discuss aspects of the dataset following instructions in Part 1.3

3.4 **Hypothesis.** Based on your data exploration result, clearly stated your three hypothesis. Follow instruction in part 1.4.

3.5 **Preprocessing**.

**Software tools**. Explain what tools you used for preprocessing.

**At Least Five preprocessing Techniques**. Explain in detail what techniques is applied and the results. Follow instruction in part 1.5

3.5 **Visualizing patterns:**

**Software tools**. Explain what tools you used for visualization

**Visualizations**. Explain in detail what each visualization mean. Follow instruction in part 3

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 date for Project 1:**