UECS3213 / UECS3453 Data Mining
Assignment 3 (Group Project)

Instructor: Dr. Simon Lau Boung Yew
Major: AM/SE
SESSION: Jan 2019

Title: Twitter sentiment analysis

Learning Outcomes (Assessment = 30%)

At the end of this assignment, students are expected to achieve the following CLOs:

- **CLO2**: Create programming solutions using data mining techniques for given problem (10%)
- **CLO3**: Evaluate performance of data mining solutions for a given problem (10%)
- **CLO4**: Construct a data mining project as a team (10%)

Introduction

Twitter is an online news and social networking site where people communicate in short messages called tweets. (Read more about Twitter: https://www.lifewire.com/what-exactly-is-twitter-2483331). Tweeting is posting short messages for anyone who follows you on Twitter, with the hope that your messages are useful and interesting to someone in your audience. On Twitter, tweet messages sometimes can be positive/negative about something or they can be indifferent (neutral). This is known as the "sentiment" of the tweets. In order for the underlying sentiment of tweet messages to be inferred, we may perform Sentiment Analysis. Sentiment Analysis is the process of determining whether a piece of writing (product/movie review, tweet, etc.) is positive, negative or neutral. It can be used to identify the customer or follower's attitude towards a brand through the use of variables such as context, tone, emotion, etc. Marketers can use sentiment analysis to research public opinion of their company and products, or to analyze customer satisfaction. Organizations can also use this analysis to gather critical feedback about problems in newly released products. (Read more about sentiment analysis: https://monkeylearn.com/sentiment-analysis/)

This assignment is about practising the real task of Text Processing. The task is to build a model that will determine the tone (neutral, positive, negative) of the twitter text. To do this, you will need to train the model on the existing data (e.g. train.csv and test.csv from https://www.kaggle.com/c/twitter-sentiment-analysis2/data). The resulting model will have to determine the class (neutral, positive, negative) of new texts (test data that were not used to build the model) with maximum accuracy.

Instruction

- This is a **group assignment**.

- Each group MUST consist of **FOUR (4) or FIVE (5)** students. Only in very special cases a group would be allowed to form group with less than four or more than five members upon lecturer's approval. Students from different major (AM/SE) are allowed to join in the same team.

- Please register Group Members at this link: https://goo.gl/Eebrrk

- This assignment consists of **THREE (3)** sections:

  - **Section 1: Coding (10%)**

1. Download the data for analysis from https://www.kaggle.com/c/twitter-sentiment-analysis2/data or https://www.kaggle.com/kazanova/sentiment140

2. Write Python program by including relevant machine learning libraries such as scikit-learn etc. to perform classification of twitter text into neutral, positive or negative.

3. Write an interesting interface in the Python program to demonstrate the classification results and the relevant accuracy of it. You may also visually display the result using Python visual plot library such as matplotlib.

4. There are plenty of tutorials available online for your reference. For examples:

   a) https://www.geeksforgeeks.org/twitter-sentiment-analysis-using-python/

   b) https://www.analyticsvidhya.com/blog/2018/07/hands-on-sentiment-analysis-dataset-python/

   c) https://towardsdatascience.com/another-twitter-sentiment-analysis-bb5b01ebad90

   d) https://www.kaggle.com/nayonika/twitter-data-analysis

   e) etc

  - **Section 2: Technical Report (10%)**

Write a concise technical report of not more than 15 pages about

a) the problem(1 page)

b) the method / flow / algorithm / tools / software libraries that you use. Show snippets of your Python codes wherever relevant and explain.

c) the classification result and its performance measurement. You may consider to adopt one or more of the following performance measurement tools, whichever relevant:

   - Cross-Validation Accuracy
   - Accuracy [Num. of Correct Queries / Total Num. of Queries]
   - Precision, Recall and F1-Measure
   - Confusion Matrix
   - ROC curves

- Agreement rate = (No. of Expected Outcomes) / (Total No. of Runs)
- etc.

d) Perform some studies on what other sentiment analysis researchers have proposed to use in evaluating sentiment analysis performance. Present and justify your results.

■ **Section 3: Group Presentation (10%)**

1. Perform a group presentation of not more than 15 minutes during Week 13/14 (presentation schedule by group will be announced later).
2. The presentation must include description of the problem, the program and demonstration of the functioning program.
3. All group members must be present and contribute to the presentation (whether as speaker or demonstrator).
4. Attire is formal.

Submission Procedure

1. Name the technical report (filename) as the following: **<Course>_<ID>_assign3.docx or .pdf** where <Course> and <ID> denote your course and university identity number of your group leader, respectively.

2. Zip and name the source codes of your project as the following: **<Course>_<ID>_assign3.zip** where <Course> and <ID> denote the course and university identity number of your group leader, respectively.

3. Name the power point presentation slides as the following: **<Course>_<ID>_assign3.ppt** where <Course> and <ID> denote the course and university identity number of your group leader, respectively. Include demo video, if any (not compulsory).

4. Submit the softcopy of the report via wble.utar.edu.my on or before **29 March 2018 (Friday) at 5.00pm sharp**.

References

[1] https://www.kaggle.com/c/twitter-sentiment-analysis2

[2] https://www.kaggle.com/kazanova/sentiment140

Marking Scheme

No.		Poor	Adequate	Proficient	Mark
Section 1: Coding					
1	Functionality & correctness	Program does not function.	Program functions partially.	Program fully functions.	
		0 - 2	**3**	**4-5**	
2	Performance analysis	No or wrong performance assessment	Partially correct and reasonable performance assessment	Correct and reasonable performance assessment	
		0 - 2	**3**	**4-5**	
				Subtotal	**/10**
Section 2: Technical Report					
1	Presentation & formatting	- Not formatted according to requirement - Not clean, untidy, too wordy, hard to read - The logical flow of the content is poor. - Not organized. - Poor grammar and language ability	- Formatted Moderately clean, tidy, concise, easy to read - The logical flow of the content is satisfactory. - Somewhat organized. - Satisfactory grammar and language ability	-Well formatted according to requirement -Very clean, tidy, concise, easy to read -The logical flow of the content is excellent. -Organized. -Well written with correct grammar, choice of words and sentence structure.	
		0 - 2	**3**	**4-5**	
2	Technical content	- Lack of clarity, invalid, incomplete - Copy and paste work	- Partially correct, valid and complete. - Lack of novelty	- Correct, valid and complete. - Original	

		0 - 2	**3**	**4-5**	
				Subtotal	**/10**
Section 2: Group Presentation					
1	Presentation and teamwork	- The flow of presentation is unsatisfactory. - The language used is nont clear and incorrect. - Poor intonation, voice projection, body language, pronunciation - Substantial delay in set up - The amount of work / effort is not acceptable. - Poor teamwork	- The flow of presentation is quite smooth. - The language used is partially clear and correct. - Satisfactory intonation, voice projection, body language, pronunciation - No delay in set up - The amount of work / effort is acceptable. - Some teamwork	- The flow of presentation is very smooth. - The language used is clear and correct. - Good intonation, voice projection, body language, pronunciation - No delay in set up - The amount of work / effort is excellent. - Good teamwork	
		0 - 2	**3**	**4-5**	
2	Demonstration & Q&A	- Program does not function - Not able to answer - Technically poor	- Program functions but demo is not smooth - Able to answer some questions - Moderately knowledgeable	- Program functions and demo is clear - Able to answer all questions - Technically knowledgeable	
		0 - 2	**3**	**4-5**	
				Subtotal	**/10**
				Total	/30%

UNIVERSITI TUNKU ABDUL RAHMAN

Assignment 3

Course Code:	UECS3213 / UECS3453
Course Name:	Data Mining
Lecturer:	Dr. Simon Lau Boung Yew
Academic Session:	2019/01
Title:	Assignment 3: Twitter Sentiment Analysis

Member ID	Member Name	Major (AM/SE)
	Mark	/100