Assignment #1 – Data preprocessing and Sentiment Analysis

Due Date: Week #5 (**09/February/2022 at 11:59 pm**)

Purpose:
The purpose of this Lab assignment is to:

1. To acquire data using website scraping .
2. To carry out pre-processing steps and data augmentation.
3. To classify tweets using a lexical approach.

General Instructions:
Be sure to read the following general instructions carefully:

1. This assignment must be completed individually by all the students.
2. Only provide the requested screenshots and make sure to have a complete screenshot, partial screenshots will not earn any marks.
3. You will have to add all the analysis and screenshots in the Analysis report.
4. You will have to provide a **demonstration video for your solution** and upload the video together with the solution on **eCentennial** through the assignment link. See the **video recording instructions** at the end of this document.
5. In your 8-minute demonstration video you should explain your solution clearly, going over the main code blocks and the purpose of each method also demoing the execution of the code. YouTube links and links to google drive or any other media are not acceptable, the actual recording file must be submitted.
6. Any submission without an accompanying video will lose 25% of the grade.
7. Any submission without an accompanying Analysis report will lose 20% of the grade.

Assignment Pre-requisites:

1. Python
2. Word2vec
3. Datasets attached to this assignment as indicated in the exercises

Assignment Exercises

Exercise #1: Web Scrap (30%)

Exercise requirements:

1. Visit your Artificial Intelligence" program web page at Artificial Intelligence - Software Engineering Technology (Online) (Optional Co-op) (centennialcollege.ca)
2. Open (Inspect) the page's source code and explore the different elements.
3. Write a program that can fetch and printout the following information from the web page:
    a. The title of the website.
    b. All possible companies offering jobs under the "Companies Offering Jobs" heading.
    c. All possible careers you can pursue under the "Career Outlook" heading.
4. Export the fetched information into a file (text file, or .csv) name the file firstname_my_future.csv or .txt

Exercise #2:Text preprocessing and data augmentation (35%)

Scenario: you have been given a dataset of users' tweets for a particular topic. You have been asked to use this data for sentiment analysis to know whether users are satisfied with the related topic. However, you figured out that the dataset is not enough to build an accurate model.

Exercise requirements:

1. If your first name starts with A-M use the COVID19_mini.csv attached to this assignment, else use the Artificial_inteligence_mini.csv file.
2. Pre-processing: Apply the following:
    a. Load the data into a dataframe. Name it "firstname_df", and examine the data. You will notice that the file has a header and four tweets with their sentiments.
    b. Drop the user column.
    c. Use regular expressions or python string methods to get rid of the additional data at the begging and end of each tweet.
    d. Check the tweet data and identify, if you need to carry out any further pre-processing steps, you should at least do one or two more steps.
    e. List them in your analysis report under exercise #2.
    f. Cary out those steps in your code.
3. Apply the following data augmentation technique to expand the original dataset:
    Word embedding augmenter. Use "word2vec" as a model for the augmenter. After applying this augmenter, you should have a dataset (call it **"firstname_df_after_word_augmenter**"). The size of this dataset should be double the size of the original dataset. That is, the size of **firstname_df_after_word_augmenter** will be 2 X size of the original dataset (firstname_df).
    a. Augment using random insertion. Write a script to apply the following steps for each cleaned tweet in the dataset:
        i. Tokenize the cleaned tweet, if you haven't done so earlier.
        ii. Remove stop words if you haven't done so earlier
        iii. Per tweet choose two words randomly.

iv. Get synonyms of each of the words selected in step i
v. Select the most similar synonym, and replace the original word with the synonym to create a new tweet (You should not replace the original tweet, you need to add a new copy of each tweet using the selected synonyms to your dataframe as a row and maintain the original sentiment
vi. Export the new dataset (after applying random insertion) into a text file. Name it (**firstname_df_after_random_insertion)**

NOTE:

i. for words' synonyms, use word2vec of Google's "GoogleNews-vectors-negative300.bin.gz"

Exercise #3: Sentiment Analysis (35%)

In this exercise you will use a larger dataset of 100 tweets, and dictionaries of negative and positive words (lexicons), cited below, to calculate the sentiment score of each tweet and then compare your calculated scores to the original sentiment scores.

Exercise requirements:

i. If your first name starts with A-M use the COVID19_data.csv attached to this assignment, else use the Artificial_inteligence_data.csv file.
ii. Drop the user column.
iii. Use regular expressions or python string methods to get rid of the additional data at the begging and end of each tweet.
iv. Carry out some basic data exploration and note all the results in your analysis report.
v. Add a column to your dataframe to reflect the length of each tweet, name it "tweet_len".
vi. Load the positive and negative words lexicons into two dataframe you can find the files in lab week #2 exercises.
vii. Iterate through all of the words in each tweet and hit against the list of lexicons in the positive and negative word datframes. Since there are longer tweets, you need to normalize the number of positive and negative hits by the number of words in each tweet.
viii. Add two columns to your datafrme one to reflect the percentage of positiveness and the second to reflect the percentage of negativeness.
ix. Tag each tweet with a sentiment score i.e. add a column to your data frame name it "predicted_sentiment_score". Use the below rules for tagging:
    a. If both positive and negative percentages are equal to zero, or if both percentages are equal (have the same percentage) tag the tweet as neutral
    b. If the positive percentage is greater than the negative percentage then tag the tweet as positive.
    c. If the negative percentage is greater than the positive percentage then tag the tweet as negative.
x. Compare the original sentiments i.e. column #1 to the new predicted sentiments you just calculated. Calculate the Accuracy and F1 score and note them in your written analysis report.
xi. Write some conclusions in your report to indicate why you got these results and what you would suggest to improve.

Naming and Submission Rules:

1. You must name your submission according to the following rule: **YourFullname_COMP262_assignmentnumber**.Example: **AdamPerjouski_COMP262_assignment1**
2. Please add all the commands/instructions into a python script.

3. Upload the submission file on e-Centennial using the Assignment link(s).
4. In total you should submit the following:
    a. One demonstration video
    b. One output file for exercise #1
    c. One output file for exercise #2One analysis report covering all three exercises
    d. Python scripts for exercises 1,2 and 3

Rubric (applies to each exercise)

Evaluation criteria	**Not acceptable**	**Below Average**	**Average**	**Competent**	**Excellent**
	0% - 24%	**25%-49%**	**50-69%**	**70%-83%**	**84%-100%**
Requirements in exercises 50%	Missing all requirements required	Some requirements are implemented.	Majority of requirements are implemented but some are malfunctioning.	Majority of requirements implemented.	All requirements are implemented Correctly.
Instruction/ Code Documentation on python script 5%	No comments explaining code. Missing screenshots	Minor comments are implemented.	Some code is correctly commented.	Majority of code is correctly commented.	All code is correctly commented.
Written analysis Content 15%	Missed all the key ideas; very shallow.	Shows some thinking and reasoning but most ideas are underdeveloped.	Indicates thinking and reasoning applied with original thought on a few ideas.	Indicates original thinking and develops ideas with sufficient and firm evidence.	Indicates synthesis of ideas, in-depth analysis and evidences original thought and support for the topic.
Written analysis report format and organization 5%	Writing lacks logical organization. It shows no coherence and ideas lack unity. Serious errors. No transitions.	Writing lacks logical organization. It shows some coherence but ideas lack unity. Serious errors.	Writing is coherent and logically organized. Some points remain misplaced. Format is neat but has some	Writing is coherent and logically organized with transitions used between ideas and paragraphs to create coherence. Overall unity of ideas is present.	Writing shows high degree of attention to logic and reasoning of all points. Unity clearly leads the reader to the conclusion.

	Format is very messy.	Format needs attention, some major errors.	assembly errors.	Format is neat and correctly assembled.	Format is neat and correctly assembled with professional look.
Demonstration Video 25%	Very weak no mention of the code changes. Execution of code not demonstrated.	Some parts of the code changes presented. Execution of code partially demonstrated.	All code changes presented but without explanation why. Code demonstrated.	All code changes presented with explanation, exceeding time limit. Code demonstrated.	A comprehensive view of all code changes presented with explanation, within time limit. Code demonstrated.

Demonstration Video Recording

Please record a short video (max 8 minutes) to explain/demonstrate your assignment solution. You may use the Windows 10 Game bar to do the recording:

1. Press the Windows key + G at the same time to open the Game Bar dialog.
2. Check the "Yes, this is a game" checkbox to load the Game Bar.
3. Click on the Start Recording button (or Win + Alt + R) to begin capturing the video.
4. Stop the recording by clicking on the red recording bar that will be on the top right of the program window.

(If it disappears on you, press Win + G again to bring the Game Bar back.)

You'll find your recorded video (MP4 file), under the Videos folder in a subfolder called Captures.

Or

You can use any other video recording package freely available.

References:

i. https://twitter-sentiment-csv.herokuapp.com/
ii. Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews." ; Proceedings of the ACM SIGKDD International Conference on Knowledge ; Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle, ; Washington, USA,
iii. Bing Liu, Minqing Hu and Junsheng Cheng. "Opinion Observer: Analyzing ; and Comparing Opinions on the Web." Proceedings of the 14th ; International World Wide Web conference (WWW-2005), May 10-14, ; 2005, Chiba, Japan.