Mini Project: Classification of Covid-19 Tweets

Please read all of the guidelines carefully before submitting the lab. ☺ There are ***100 points*** in total. **You can work alone or in a group of two in this project.**

Due date: September 25th, 2022 (Sunday), 11:59 PM. Late submissions will be penalized as per syllabus by 10 points / day. No submissions will be accepted three days after the deadline.

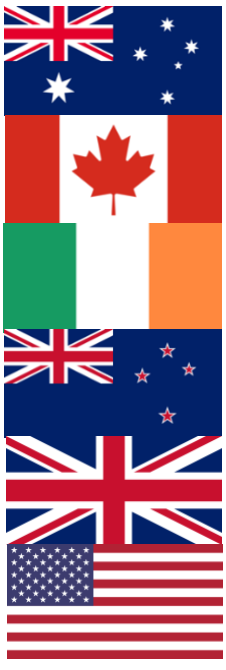
Deliverables:

1) The code of the project in **.ipynb** format
2) The lab report written with ***LaTeX*** and exported in **.pdf** format (one file)

Guidelines – Before You Start

1) You will be using the **`Python`** programming language for this project. You need to write your codes in an empty **.ipynb** file.
2) Make sure that you provide many comments to describe your code and the variables that you created.
3) Please use the ***IEEE*** journal template on ***overleaf.com***. Here is the link: https://www.overleaf.com/latex/templates/preparation-of-papers-for-ieee-sponsored-conferences-and-symposia/zfnqfzzzxghk
   To be able to work on ***overleaf.com***, you will need to register first (you can also compile your ***LaTeX*** file locally.)
4) For some of the code, you may need to do a little bit of "*Googling*" or review the documentation.

What is the Covid-19 Tweets Data?



This dataset consists of Covid-19 related tweets posted by users coming from six English-speaking countries: Australia, Canada, Ireland, New Zealand, the United Kingdom, and the United States. You will have:

- Tweets about Covid-19 coming from six countries collected in 2020[1]
- Each country is represented with 40,000 tweets in the training set and 10,000 tweets in the test set.[2]
- In total, there are 240,000 tweets in the training set (and 60,000 tweets in the test set).

Mini Project – Data Dictionary

The data has been provided in the assignment folder online (**training_data.csv** and **test_data.csv**). Open the CSV files and take a look at them before starting. Individual features (columns) of the dataset have been described below:

[1] Earliest tweet is dated April 17, 2020, and the latest tweet is dated December 10, 2020.
[2] The observations in the training set are ordered by **`country`** label, the test set is randomly ordered.

text: The text of tweet (including emojis, htmls, hashtags)

reply_to_screen_name: The Twitter screen name of the user the owner of the tweet is replying to (if any)

is_quote: A Boolean variable that indicates if the owner of the tweet is quoting someone else's tweet

is_retweet: A Boolean variable that indicates if the owner of the tweet is retweeting someone else's tweet

hashtags: A list of hashtags included in the tweet

country: The label of the country in which the tweet was posted (found only in the **`training`** dataset)[3]

Id: An index number associated with tweets (found only on the **`test`** dataset)

Before you start with the questions below, create a new and empty folder called **covid19tweets_mini_project**. Call the file where you will write your code **covid19tweets_mini_project**, as well.

Part I: Descriptive Analysis (30 points)

In this part of the analysis, you will be exploring some NLP (natural language processing) techniques to better understand the data. For *graduate* students, each question in Part I will have a weight of 5 points.

a) Create a **table** that contains information on minimum, average, median, and maximum for the following: tweet length (#characters and #words) (`text` column), hashtag length (#characters and #words) (`hashtags` column) (Add your table to the report at the end.) (**10 points** for undergraduate students, **5 points** for graduate students)

b) Find the top ten most commonly used hashtags (`hashtags` column) in the training dataset by calculating the frequency. Then, create a stacked bar chart (one stacked bar per country) which shows the distribution of these ten most commonly used hashtags for each country. Do you observe any patterns? Write your findings in the report. (Add the stacked bar chart to the report at the end.) (**10 points** for undergraduate students, **5 points** for graduate students)

c) Using the **lda_tutorial.pdf** file in the assignment folder, perform a **Latent Dirichlet Allocation (LDA)** analysis to extract the topics in the ***text*** column in an unsupervised manner. Set the number of clusters/topics to **10 (ten)** (you can adjust other settings to obtain the results that you think work the best). What are your observations? Does each cluster seem to form a meaningful subset? What do they seem to represent? (Add the clusters and your observations to the report at the end.) (**10 points** for undergraduate students, **5 points** for graduate students)

Graduate students must also complete the following questions.

d) Using the code in the following link[4], perform **Non-negative Matrix Factorization** for topic analysis. Again, like in question c), set the number of clusters/topics to 10 (ten) and extract the

[3] Countries are labeled as ***'us'*** (United States), ***'uk'*** (United Kingdom), ***'canada'*** (Canada), ***'australia'*** (Australia), ***'ireland'*** (Ireland), and ***'new_zealand'*** (New Zealand).

[4] https://scikit-learn.org/stable/auto_examples/applications/plot_topics_extraction_with_nmf_lda.html

topics in an unsupervised manner. Adjust any parameters as you see fit. Analyze the results. Do you see any similarities or differences with respect to your results in c) ? Explain. (**5 points**)

e) Write a ***'text cleaner'*** function that does the following: (i) remove stopwords[5], (ii) remove all words that are shorter than 3 characters, (iii) remove all links (starting with *http*), (iv) remove emojis, (v) remove punctuation. Attach the code you wrote to the ***lemmatizer.py*** file in the project folder. Run the lemmatizer function and create 'cleaned and lemmatized' version of ***text*** column. (You can name the new column as ***text_clean***). After the cleaning, expand the table you have created in Part I, a) by calculating minimum, average, median, and maximum for the newly created ***text_clean*** column (#characters and #words). (**5 points**)

f) Extract the unique set of all different hashtags found in the training dataset (i). Then, create one separate data subset for each country by splitting the data using the `country` column (ii). As next, for each country, create a vectorized format of the count values for all different hashtags you have extracted in (i) (you should obtain one count vector per country with all unique hashtags). Using ***cosine similarity***[6], compute the pairwise similarity values between **the** different count vectors you have created for each country. Finally, create a heatmap[7] of the pairwise similarity values you have calculated. What are some of your observations? Are any two countries more similar to each other than others? If yes, explain. If not, explain, as well. Are there any very large or very small values? (Add the graph you created to the report at the end.) (**5 points**)

Part II: Model Creation and Prediction (40 points)

For this part of the analysis, you will need to train a model that classifies the tweets in your training dataset according to '**country'** labels, report the **Accuracy** of your best model (i), and the **confusion matrix** (ii) that you will create. **You will mainly be graded on the Accuracy of your model** (more information provided below). Some guidelines (please also review the presentation on Mini-Project uploaded on BlackBoard):

- You can use **any** classification model (including, but not limited to, logistic regression, decision trees, neural networks etc.)
- You can use **any** feature engineering method to transform your dataset, such as:
    - Dimensionality reduction methods such as PCA, t-SNE, spectral embedding
    - Logarithmic, polynomial, and other transformations
    - Different vectorization techniques
    - Different weighting strategies
- You are free to create a new column (or a stream of data) based on the existing columns and use your new column as an independent variable.
- You are welcome to use **any** external dataset to enrich your training and test datasets.

[5] For more information: https://www.geeksforgeeks.org/removing-stop-words-nltk-python/
[6] More information: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html
[7] Documentation: https://seaborn.pydata.org/generated/seaborn.heatmap.html

Model Evaluation

Your submission will be evaluated using the ***Accuracy*** cost function. If you are unsure, please review what ***Accuracy*** means before starting on the project. **Please also report all of your code in the .ipynb file and your confusion matrix both in the .ipynb file and in the report. Please also report *Accuracy* in your code.**

Please use your actual name on Kaggle! [Or, please indicate your nickname in the report!]

Make sure that all of your code is running!

Save the code file you have created as "**covid19tweets_lab.ipynb**" in the folder you have created at the beginning.

Part III: Creating the lab report (30 points)

Write a report (minimum 2 pages) that includes the names of you (or you and your group member), all of your findings and the visuals that you created. The report that you will write should use the *IEEE format* and include the following sections:

Abstract: A short summary of your report (This part should include a very brief summary of your methods and analysis and the answer to why you think what you have done is important)

Introduction: A summary of what you expected and did, and two-three of your most significant findings (please use some numerical results here)

Data: Introduce your descriptive findings about the dataset here

Methods: Provide a description of your strategy and the steps you took to improve your prediction model (this includes the steps you followed for data-preprocessing, setting up the model, and checking the strength of the model)

Results: A detailed discussion on the results you obtained. What is your Accuracy value? Evaluate and criticize yourself / your team.

Save the project report as **covid19tweets_mini_project_report.pdf**.

Final step:

Send your code as an **.ipynb** file and the report in the **.pdf** format through *BlackBoard*.

General Rules and Grading

You will be graded based on the following criteria:

- Code: Cleanliness/understandability (i), executability (ii), format (iii)
- Ranking: Ranking in the **Kaggle** competition
- Lab Report:
    - *Introduction* (i), *Data* (ii), *Methods* (iii), *Results* (iv)
    - Flow, readability, level of detail, quality of visuals/tables, adherence to the guidelines