# MIE1513 Final Project

Submission date: April 20, 2020 at 23:59pm

## Introduction:

In this project you will analyze data from the reddit.com, one of the most popular online social platforms, and apply some of the techniques covered in the course.

In Reddit, posts are organized into user-created boards called "subreddits", which cover a variety of topics like news, science, movies, video games, music, etc. Users can up-vote or down-vote a post, as well as commenting on a post.

In this project you will analyze a sample of posts (without their comments) from a small subset of very active subreddits. The key tasks in this project are:

a. Exploratory data analysis.
b. Automatically classifying posts to their subreddit.
c. Analyzing the sentiments in reddit post titles.

Following is information of the deliverables and marking scheme, academic integrity, and the submission process. Then, we provide detailed explanation for each of the parts on the project.

**Deliverables and marking scheme:**

- The main deliverable for this project is a Jupyter notebook containing **both your code and text cells (use markdown for styling)** with answers.
- Your notebook should be **fully reproducible on Google Colab**, i.e., we should be able to run your notebook from top to bottom without errors.
- Notebooks will be graded manually based on **both the code and text cells**.
- The marking scheme is as follows:

| | |
|---|---|
| Part 1: Loading the Data [code provided] | 0 pts |
| Part 2: Exploratory Data Analysis | 10 pts |
| Part 3: Classification of Reddit Posts | 15 pts |
| Part 4: Sentiment Analysis in Reddit | 5 pts |
| General: Report Structure & Style | 5 pts |
| **Total:** | **35 pts** |

**Academic Integrity:**

- The submitted code and answers should be written by you, represent work that you have done, based on your ideas.
- You are allowed use material from you own previous labs and assignments.
- You are allowed to use public resources such as books and online resources, however you must cite them in your work if you use any of the ideas presented there.
- You are not allowed to share code or ideas with classmates.

- In you submitted notebook, you will be asked to declare the following:
  "In submitting this assessment, I confirm that my conduct during this project adheres to the Code of Behaviour on Academic Matters. I confirm that I have not acted in such a way that would constitute cheating, misrepresentation, or unfairness, including but not limited to, using unauthorized aids and assistance, impersonating another person, and committing plagiarism. I pledge upon my honour that I have not violated the Faculty of Applied Science & Engineering's Honour Code during this assessment."

**Submission Procedure:**

- Submission will take place through Quercus using the dedicated submission box "Final Project Submission".
- You need to upload the .ipynb file of your notebook after running it from top to bottom (including the output of all cells). Make sure your notebook includes the cell with integrity declaration.

**How to ask questions:**

- Questions related to the project should be posted as **PRIVATE POSTS on Piazza**.
- **DO NOT use public posts** on piazza for the questions related to the project.

# Part 1: Loading the Data [code provided]

a. **Load the provided dataset (reddit_data.json):**
   - Run the corresponding cells in your notebook.
   - The json files include sample of posts from January 2019 from the top 25 subreddits.
   - The provided data has the following fields:

| | |
|---|---|
| id | Unique identifier for the post |
| author | The author of the post |
| created_utc | UTC timestamp of post |
| subreddit | The subreddit where the post is posted |
| title | The title of the post |
| num_comments | The number of comments to the post |
| score | Score of the post – sum of 'ups' and 'down', i.e., if there are 5 'ups' and 2 'downs' the score is 3. |
| selftext | Body of the post [optional] |
| url | Link attached to the post (e.g., video, image, news article) [optional] |

b. **Generate your unique subset of the data:**
   - Each student will analyze a different subset that includes the posts of 4 subreddits.
   - The provided function **`getMySubreddits(data, my_str)`** generates your unique set of 4 reddits.
   - Update the variable **`unique_string`** to be your UofT email address.
   - Run the corresponding cells in the notebook to generate your subset of the data in the variable **`data`**.

# Part 2: Exploratory Data Analysis (EDA) [10 pts]

For the Natural Language Processing tasks, we usually need more time to clean and explore the data. In the first part of your project, you need to do a thorough EDA to understand your data better. All of your EDA should be in the project report with visualizations (tables, figures) and corresponding explanation paragraphs. Note that a good EDA report should be structured instead of listed experiments. To be specific: you need to do EDA column by column (data columns), and for each column, you need to provide a list of analyses that could be referred to in your next tasks.

The data provided has multiple columns, including raw text as well as meta-data such as author, URL and number of comments.

a) We do not know which of those columns are informative for us (e.g., in order to do better classification in Part 3). Thus, the first challenge is to identify the importance of *each* column. For example,
   - Is the URL a good indicator of this classification task?
   - Should we do further processing on the URL to make it cleaner before doing classification?
   - Does there exist any author who is very productive and has a strong preference for specific topics?
   - Is there any correlation between each pair of columns? Can we remove redundant columns?

b) As the raw data is very rough in terms of containing many random symbols such as dollar sign ($) and stars (*). If you choose to remove them, provide some insights on why those symbols are not useful.

c) Data may have outliers and may also be imbalanced. Can you tell if the data is balanced or imbalanced? Can you visualize it?

The above are just a few examples, but you can do far more than this to improve your classification performance. Again, in the classification task, any modification in your model should have corresponding EDA evidence support.

# Part 3: Classification [25 pts]

- In this part, the goal is to automatically classify subreddit labels based on the rest of the information provided.
- There are many classification algorithms that are available for you to choose from, including Naive Bayes, Logistic Regression, SVM etc. Some of these were covered in class, however you are free to use other algorithms in Sklearn packages (in which case you will have to try to generalize your knowledge to these classifiers).
- There is no expected threshold for the performance, but you should demonstrate effort to achieve good performance.

In this part you need to make sure that:

1. You need to know which metric is the most suitable one for this task and why. Propose at least three metrics and provide a short interpretation and justification of your chosen metric.
2. You need to define and justify the evaluation method. E.g. how do you split data? Do you need a cross-validation?
3. You need to compare at least three classifiers, at least three different feature set, show hyper parameters tuning results.
4. Compare models in a systematic way (Table or Plots). Do not write results in plain text. See one of Wuga's older papers as a simple reference.
5. Try to analyze and characterize the performance of your model (e.g., where your model is doing well and where it does not).

# Part 4: Sentiment Analysis [5 pts]

The target of this part is to automatically classify subreddit labels based on the rest of the information provided.

a. Use Vader to compute for each post the sentiment of title field.
b. Analyze the distribution of sentiments for each of your 4 subreddit. What can you learn from the results?
c. Define a threshold for posts that are clearly positive and clearly negative according to Vader (e.g., > 0.75 and <-0.75, but this should be informed based on the distribution of values in your data) and assign a label for each post accordingly, while removing posts that are not clearly negative or positive.
d. For each of the four subreddits, compute the top words that tend to occur in (i) positive titles (ii) negative titles, based on the threshold for positive/negative in (c).
    a. Analyze the results: do you notice any subreddit-specific patterns?
    b. Explain the method/metric you used in your analysis and justify why it was appropriate here.

# General: Report Structure and Style [5 pts]

These points are given to:

- Clean and documented code
- Clear, concise, and complete answers
- Every result is interpreted or analyzed
- Every choice is explained and justified
- Proper use of visualization when appropriate
- Correct and clear English