COMP6125 – Knowledge Discovery and Data Analytics II
Final Project – Part 1 and Part 2 (Weighting: 50%)

Due Date: Sunday 19h December 2021 @ 11:55 p.m.

Instructions:

1) This project is to be done in groups of two to five.

2) Choose only one problem: Problem 1, Problem 2, or Problem 3

3) Submissions will be done via Google Forms at this address (https://forms.gle/ppnZHcJ6THghh1VG9) on or before the posted due date as a single Zip archive (not .rar). The archive must contain:

   a) The dataset in raw format (eg. .jpeg, .png, .csv, .txt) located in the folder **./data/**
   b) **Development Notebook.ipynb** (will contain data loading, exploration, training and model saving code. This spans Part 1 and Part 2 activities.)
   c) **Deployment Notebook.ipynb** (will contain code to load the model and to display a simple[1] interface that can be used to supply examples and receive the prediction of the class.)
   d) **Project Summary.pdf** containing:
      i. Listing of the group members
      ii. Description of the project (<200 words)
      iii. Data statement/Model Card[2]
      iv. Summary of the results
      v. Observations and plans for the next iteration. (This should answer questions such as: How well did the model do compared to what would be needed for production? How can this be achieved? Etc.)
      vi. Screenshots of the deployment notebook showing the widgets and predictions being made.

   IMPORTANT: *Ensure that the directory structure is maintained so that the notebooks can locate the files in the subdirectories when the archive is unzipped. Ensure to adhere to the file-naming conventions specified above.*

4) If there are any challenges, please contact me via email at djones.comp6125.2021@gmail.com as soon as possible.

[1] Simple means a heading, a description, textboxes or buttons and their corresponding labels.
[2] Mitchell et al. 2018. https://dl.acm.org/doi/abs/10.1145/3287560.3287596

Problem 1 – NLP – #JamaicaDecides

In September 2020, Jamaica held its most recent general elections. The National Electoral Commission wishes to investigate and report on the attitudes of the public towards issues and individuals.

You have been provided with a dataset of the twitter discourse before and after the election. From this, you are tasked with investigating the dynamics of the public discourse.

Your investigations yield various insights. Among them, you must include (but not be limited to):

1. How the sentiments of the public evolved over time (day by day)
2. Whether the public express consistent support for one party or do their attitudes vary depending on key events such as news, scandals, debates.
3. What percentage of the public appears to be aligned with each party
4. What percentage of the public appears to be neutral
5. What level of discussion/debates take place between individuals with opposing views
6. What level of influence do people have within the discourse

You have therefore been tasked with performing the following:

Part 1: Data Preparation and Ethics (Overall Course Weighting: 40%)

- Develop or use datasets or models to support your investigation. A data statement must be provided for any dataset that is used and a model card (Mitchel et al. 2018) for any model that that is used/developed. (10%)
- Development Notebook – Research Question X.ipynb *(30%)*
    - Present a thorough exploration of data with respect to each research question identified. You must include statistics on the (1) lengths of the texts, (2) the languages used
    - Use a mixture of classification and topic modelling techniques should be used.
    - Use graphs to illustrate the insights discovered.

Part 2: Modelling and Deployment (Overall Course Weighting: 10%)

- Documentation in PDF format
    - Develop a summary of the findings to be submitted to the executive committee of the National Electoral Committee.

Problem 2 – Computer Vision – Banking Machines

Since 2020, businesses around the world have been forced to limit face-to-face activities. In response, the International Commercial Bank (ICB) wishes to introduce new automated machines to accept cash and cheque deposits from clients. The hardware at their disposal is capable of taking pictures of the bank notes and sending that to the system for processing.

The chief technology officer (CTO) has decided that it would be best to first develop a proof of concept and refine the approach in subsequent iterations. In the first iteration you will develop a system that is capable of classifying just two different notes of currency based on a picture of each note.

You have therefore been tasked with performing the following:

Part 1: Data Preparation and Ethics (Overall Course Weighting: 25%)

- Develop a **dataset** of two different notes (18%)
    - At least 50 examples must be crowd-sourced (See Appendix 1 Crowdsourcing using Google Forms)
- *Development Notebook.ipynb*
    - Present a brief exploration/preview of your dataset in your Jupyter notebook. (2%)
- *Project Summary.pdf* inclusive of a data statement (5%)

Part 2: Modelling and Deployment (Overall Course Weighting: 25%)

- *Development Notebook.ipynb*
    - Develop a **model** that is capable of classifying two different notes. (15%)
    - Save your trained model to disk in *.h5* format (5%) (See Tensorflow Documentation on Saving and Loading Models here. If you are not using Tensorflow, use the save format recommended by that framework.)
- *Deployment Notebook.ipynb*
    - Create a presentation/deployment notebook that loads your trained model and presents a simple interface to allow a user to supply examples.
    - You will use Jupyter Notebook Widgets to create these items. (5%) (See Appendix 2 - Using Jupyter Widgets and Voilà to create a User Interface)

Problem 3 – NLP – Health System

Since 2020, health care systems around the world have been have been stretched to capacity. Many persons in the general public are uncertain as to whether they should seek urgent medical attention or treat their ailment at home. In response, some health care systems have set up interactive questionnaires that can provide this guidance.

The National Health Authority (NHA) wishes to explore providing a text-based service that is able to allow the public to describe symptoms and receive guidance. They wish to deploy this via SMS in order to make it as widely accessible as possible.

You have been asked to develop a proof of concept that is capable of classifying four different illnesses based on the description in layman's terms.

You have therefore been tasked with performing the following:

Part 1: Data Preparation and Ethics (Overall Course Weighting: 25%)

- Develop a **dataset** containing descriptions of four different illnesses (15%)
  - At least 50 examples must be crowd-sourced (See Appendix 1 Crowdsourcing using Google Forms)
- Development Notebook.ipynb
  - Present a brief exploration/preview of your dataset in your Jupyter notebook. (5%)
  - For each category/label/class, you must include a presentation of the statistics on the (1) lengths of the texts, (2) the languages used
- *Project Summary.pdf* inclusive of a data statement (5%)

Part 2: Modelling and Deployment (Overall Course Weighting: 25%)

- *Development Notebook.ipynb*
  - Develop a **model** that is capable of classifying two different illnesses[3]. (15%)
  - Save your trained model to disk in *.h5* format (5%) (See Tensorflow Documentation on Saving and Loading Models here. If you are not using Tensorflow, use the save format recommended by that framework.)
- *Deployment Notebook.ipynb*
  - Create a presentation/deployment notebook that loads your trained model and presents a simple interface to allow a user to supply examples. You will use Jupyter Notebook Widgets to create these items. (5%) (See Appendix 2 - Using Jupyter Widgets and Voilà to create a User Interface)

[3] Note that you should try to choose two illnesses that have clear differences between the symptoms. For example, it may be difficult for a classifier to distinguish between COVID19 and influenza.

Appendix

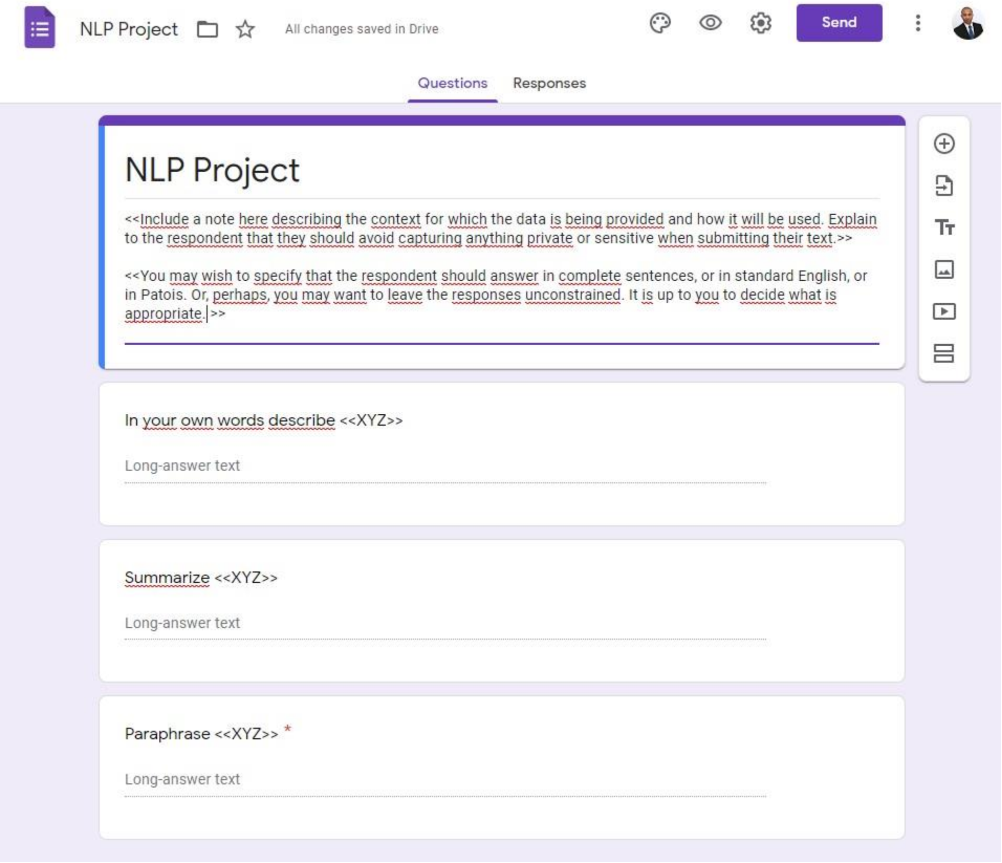
1. Crowdsourcing using Google Forms

Crowdsourcing data is typically done with a purpose-built platform like Amazon Mechanical Turk and others, as discussed in the course. However, for the purpose of this activity a platform like Google Forms should do just fine.

Screenshot A below shows a starter setup for collection of text. Be sure to inform respondents as you would with any other form of data collection. The screenshot provides examples of prompts that could be used to obtain natural sounding, varied responses.

Screenshot B below shows a starter setup of a form that can be used to collect images. A Google Forms requires users to be signed in to upload files to your form. Therefore, if you will be using Google Forms limit the collection to a people whom you know and with whom you have mutual trust. **IMPORTANT: Ensure that the file type is constrained to image file types.**

Screenshot A



NLP Project

<<Include a note here describing the context for which the data is being provided and how it will be used. Explain to the respondent that they should avoid capturing anything private or sensitive when submitting their text.>>

<<You may wish to specify that the respondent should answer in complete sentences, or in standard English, or in Patois. Or, perhaps, you may want to leave the responses unconstrained. It is up to you to decide what is appropriate.>>

In your own words describe <<XYZ>>

Long-answer text

Summarize <<XYZ>>

Long-answer text

Paraphrase <<XYZ>> *

Long-answer text

Screenshot B

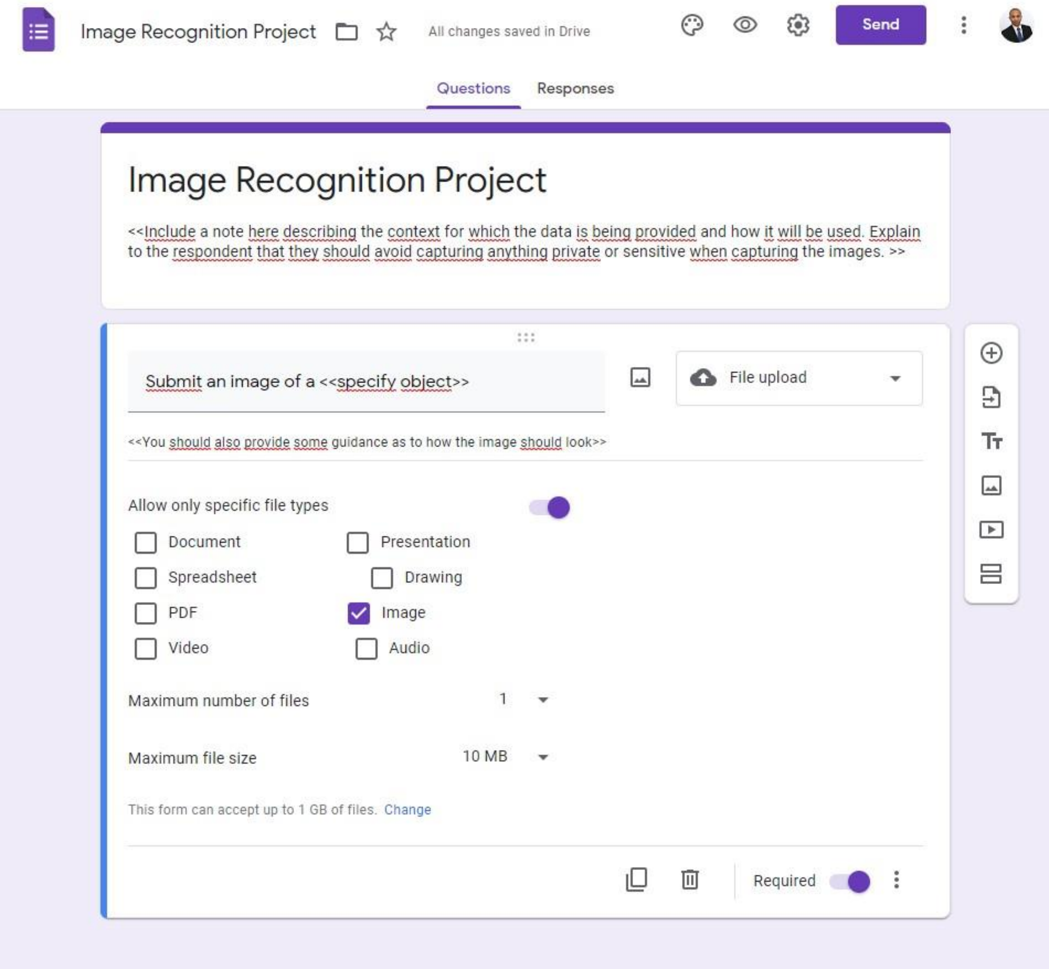


Image Recognition Project

<<Include a note here describing the context for which the data is being provided and how it will be used. Explain to the respondent that they should avoid capturing anything private or sensitive when capturing the images. >>

Submit an image of a <<specify object>>

File upload

<<You should also provide some guidance as to how the image should look>>

Allow only specific file types ●

□ Document	□ Presentation
□ Spreadsheet	□ Drawing
□ PDF	☑ Image
□ Video	□ Audio

Maximum number of files 1

Maximum file size 10 MB

This form can accept up to 1 GB of files. Change

Required ●

2. Using Jupyter Widgets and Voilà to create a User Interface

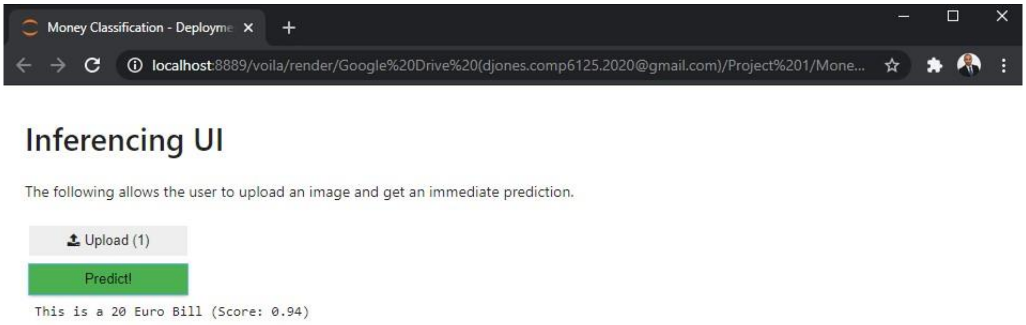
The screenshot below shows what a simple UI created with Voilà and Jupyter Widgets could look like. There is an article on Voilà here and you can find Jupyter Widgets here.

Jupyter Widgets provides controls like buttons and text boxes that can allow the user to add data as variables that you can process in your code. Voilà takes an existing Jupyter Notebook and displays all the non-code content. So, to make a clean web page using Voilà you just need to remove unwanted text. One strategy could be to replace the text display blocks with inline comments in the code blocks.

Note that Voilà is not available on Google Colab. The Jupyter Widgets will function as you intend but there will be no Voilà to convert it to a simple web page. For this, you will have to run your notebook locally. You may do this after you have obtained a trained model. Note that Google Colab runs Tensorflow 2.3 so your local environment will need the same version to avoid problems with loading your saved models.

Be sure to consider who the target user is when you are wording the text to be displayed. For example, instructions might be useful.

Screenshot 3


Money Classification - Deployme
localhost:8889/voila/render/Google%20Drive%20(djones.comp6125.2020@gmail.com)/Project%201/Mone...
Upload (1)
Predict!

Inferencing UI

The following allows the user to upload an image and get an immediate prediction.

```
This is a 20 Euro Bill (Score: 0.94)
```

3. Text Classification

You may use any of the libraries, frameworks and methods that we covered in the lectures of the course. You are encouraged to use SVM and Naïve Bayes approaches as a baseline but use Transformers (like BERT) for the final solution. These are state of the art and highly applicable to real-world problems.

You will have to determine what pre-processing steps are relevant based on your exploration of the data.