NLP Project Description Spring 2024

Deadlines are listed below

1 Overview

This document contains the requirements of the NLP project and the final report for Spring 2024. Natural Language Processing is one of the most active research fields worldwide, new problems and challenges are introduced every day which motivates the need for creative ways to deal with them aiming to solve the problem and eliminate it or reduce its effect.

In this project, you are required to target one of the active tasks that appear in real applications listed in Section2.

You can work in teams of max two while specifying the work done by each team member and their contribution, link for submitting the teams: https://forms.gle/p1b688gnjDLuGLrA8

The deadline for submitting the team is Sunday 3^{rd} of March 2024 at 11:59 pm.

Note: You are also required to share your GitHub project link so we can track your progress as it is part of the evaluation. Kindly make sure that it is visible by adding me, *mayaarosama* to the project or making the repo public.

2 Project Requirements and Milestones

2.1 NLP Problems

In this project, your team will get the chance to work on solving one of the main tasks in NLP and IR by working with real collected datasets. For each problem, you may use one of the given datasets, or if you would like to use a different dataset please propose it before the 3^{rd} of March 2024.

1. Generation model to generate song lyrics by artist name and genre.
   - 57,650 Spotify Songs
   - 500 Greatest Songs of All Time
2. Generating replies based on movie scripts,The Terminator Franchise Movie Scripts
3. Question Answering System that would help the user decide on a product to buy, Amazon's 500 Bestsellers in Laptop Gear 2024
4. Question Answering System that would maintain chat history in consideration
   - Glassdoor Data science Jobs - 2024
   - Customer Support Dataset
   - LinkedIn Job Postings 2023
5. Question Classification and Answering System
   - Andriod VS IOS Dataset
   - Email Thread Summary Dataset
6. Summarizing model
   - Towards Data Science

- Environment News Dataset
- Email Thread Summary Dataset

7. Machine Translation
    - Environment News Dataset
    - FastText Translation Data
    - Translated Dataset Augmentation

2.2 Milestones and Deliverables

This project is divided into three main milestones, each milestone is worth 10% for which you will submit a report and your code for the given task. In milestone 1, the objective is to allow you to study and get an overall idea of the given problem and present the literature review in your report. For the technical part in milestone 1, you will apply data analysis and provide a review of the dataset you are working with. By milestone 1, your report should contain i) an introduction and motivation section that does not exceed 1 page, ii) a literature review section of at least two pages containing recent work, done on the given problem, and iii) presenting your data analysis and insights over the given dataset while discussing the limitations.

In milestone 2, you will build a neural network model, without using any pre-trained model, to solve the given problem. (Continuing on the same report) In the report, you should explain the methodology that you followed and present the training and evaluation results. The evaluation will be based on the architecture of the network, not on the performance of the model.

In milestone 3, you may use a pre-trained model and apply fine-tuning and/or transfer learning to improve the performance and evaluate the system on natural text unseen during the training phase. In the report, you should i) update the methodology section with the chosen pre-trained model architecture, ii) present the training and evaluation results, and iii) discuss your findings and conclusions.

3 Timeline

Kindly note that the submission should be done on your GitHub submitted repo, the last commit before the deadline is the one that your milestone will be evaluated on.

- This document is to be posted on Tuesday 26^{th} of February 2024 and the deadline for the final project and presentation report is on the 18^{th} of May 2024 at 11:59 pm.
- Deadline for team submission is Sunday 3^{rd} of March, 2024 at 11:59 pm.
- Reporting issues regarding team submission by Thursday 7^{th} of March via GUC email: mayar.osama@guc.edu.eg with subject *NLP Project Team Issue*
- Final teams announcement Sunday 10^{th} of March, 2024
- Milestone 1 deadline is 11^{th} of March, 2024 at 11:59pm
- Milestone 2 deadline is 13^{th} of April, 2024 at 11:59pm
- Milestone 3 deadline is 18^{th} of May, 2024 at 11:59pm
- Final Presentation will be held on the 20^{th} of May, 2024. The teams' specific times and locations are to be announced by Sunday, May 19^{th},2024.

4 Useful Links

- 10 Leading Language Models For NLP In 2022
  https://www.topbots.com/leading-nlp-language-models-2020/

- An Extensive Guide to collecting tweets from Twitter API v2 for academic research using Python3
  https://towardsdatascience.com/an-extensive-guide-to-collecting-tweets-from-twitter-api-v2-for-academic-research-using-python-3-518fcb71df2a
- How To Extract Data From The Twitter API Using Python
  https://towardsdatascience.com/how-to-extract-data-from-the-twitter-api-using-python-b6fbd7129a33
- Huggingface https://huggingface.co/models?sort=downloads
- Best 25 Datasets for NLP Projects
  https://www.kaggle.com/discussions/general/150720#845341
- ALUE: Arabic Language Understanding Evaluation https://aclanthology.org/2021.wanlp-1.18.pdf
- Reading list for Awesome Sentiment Analysis papers
  https://www.kaggle.com/getting-started/150145
- Papers with codes where you would find papers, state-of-the-art, datasets, etc..
  https://paperswithcode.com/
- Arabic News Articles Dataset
  https://www.kaggle.com/datasets/haithemhermessi/sanad-dataset?select=Tech
- Twitter Data set for Arabic Sentiment Analysis Data Set
  https://archive.ics.uci.edu/ml/datasets/Twitter+Data+set+for+Arabic+Sentiment+Analysis
- Arabic-named-entity-recognition https://github.com/EmnamoR/Arabic-named-entity-recognition
- https://huggingface.co/datasets/Fatima-Gh/GLARE
- https://archive.ics.uci.edu/ml/datasets/Twitter+Data+set+for+Arabic+Sentiment+Analysis
- https://arbml.github.io/masader/