CSCE 5290: Natural Language Processing

Project name: "Identify Quora Questions That Have Similar Meaning"

Team: Thanh Le - Linh Ha

Github Link: https://github.com/haroldle/NLP_Class_project

1. Goals and Objectives:

• Motivation

My group's project is "Identify Quora Questions That Have Similar Meaning". Quora is a questions answering platform in which 67,000 questions are asked per day[3]. Within that number, there are various similar questions or questions carrying similar meaning in different words. As a Quora users, we understand the convenience and efficience in finding correct answers without providing exact sequences of words. Thus, in this project we will build a model to achieve the top 15 leaderboards in the Kaggle Quora question pair similarity contest.

• Significance

People ask questions every day on the internet. Some of the common sites where people ask questions are Stackoverflow, Quora, or maybe sometimes Reddit. On those sites, many questions are similar in context, which made the writers answer multiple times. This issue costs the users more time than they need, and in some case, they could not find a satisfy answer because they did not put correct sequence of words. Besides that, having multiple similar

questions and answers increases unnecessary storage as well as the budget for maintenance. Having a machine learning (ML) model that can group similar questions that can help people to find the answers easily and reducing the cost in up keeping the storage.

• Objectives

This project's domain is in semantic textual similarity. The semantic textual similarity is about determining whether two pieces of information are similar; In this case, we are trying to compare the similarity between two questions. We are going to have three objectives for this project.

- Firstly, understanding, visualizing, and transforming text data. For the first objective the goal is to get us familiar with the dataset and experimenting which feature transformation techniques work well on this dataset.
- Second goal is to build a ML model for based comparison, with some advice in this research paper[2]
- Thirdly is to re-train a pre-built transformer model to make a comparison and tuning.

• Features

We will use the Quora Question Similarity dataset, which is on Kaggle. This dataset has 4 feature columns:

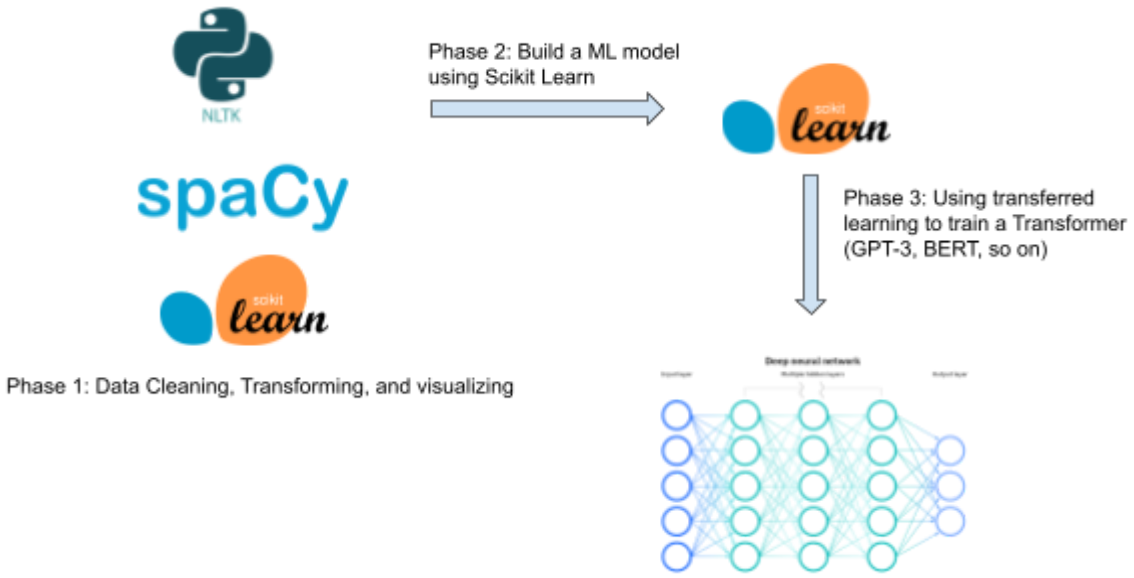
" **id** - the id number of the train pair

qid1, qid2 - unique ids of each question

question1, question2 - the full text of each question

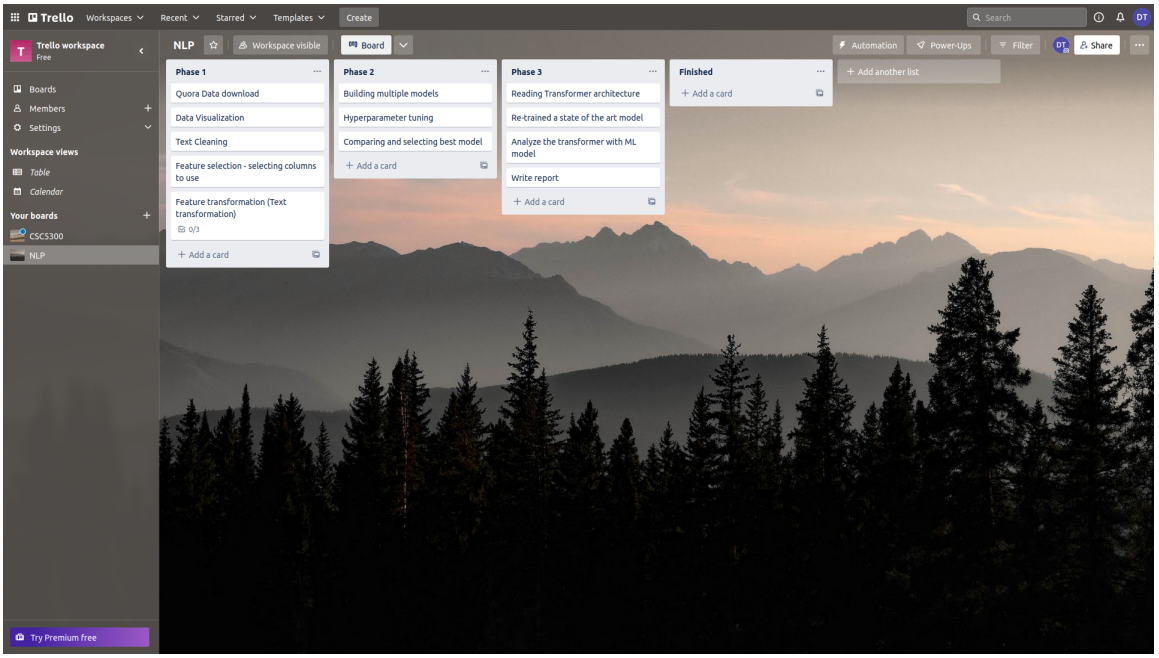
is_duplicate - the target variable, set to 1 if question1 and question2 have essentially the same meaning, and 0 otherwise." [1]

Project flows visualization:



We are thinking of dividing the project into three parts. The first part is about getting familiar with the Quora Question Similarity Dataset. In this phase, we are thinking about visualizing the data using Spacy, Scikit-learn. Then, we are going to do some text-cleaning and transforming techniques: removing stopwords, stemming, lemmatizing, and doing tf-idf by using Spacy, nltk, Scikit-Learn. The second part is about training a based line machine learning model: Tree-based model or Simple Linear Regression in Scikit-Learn. In the third phase, we are going to use a pre-trained deep learning model like GPT3, Transformers (BERT) and do an analysis in comparing performances between them.

Task table: (the image below is the table showing how many sub-tasks that we think we are going to do in this project.)



There are three phases which present three goals that we are going to progress. Each progress is going to have some sub tasks that we should finish; The sub tasks can be changed in the future. At the moment, we are just brainstorming what sub tasks are likely to happen when we begin the project. Every subtask in this table have been discussed in the project flow visualization.

3. References

1. https://www.kaggle.com/competitions/quora-question-pairs/data
2. https://arxiv.org/pdf/1907.01041.pdf
3. https://medium.com/analytics-vidhya/quora-question-pairs-similarity-problem-8e3ae90441f0

Image links:

- https://www.google.com/url?sa=i&url=https%3A%2F%2Fpython.plainenglish.io%2Fintroduction-to-nltk-library-in-python-6fa729b54ad&psig=AOvVaw0UI1bZ1nafWmImdgjkAVfB&ust=1665237885702000&source=images&cd=vfe&ved=0CAwQjRxqFwoTCMi1nOmkzvoCFQAAAAAdAAAAABAE
- https://www.google.com/url?sa=i&url=https%3A%2F%2Fen.wikipedia.org%2Fwiki%2FSpaCy&psig=AOvVaw0PxxdrkA8jwnCduIVf0j82&ust=1665237919850000&source=images&cd=vfe&ved=0CAwQjRxqFwoTCMixo_WkzvoCFQAAAAAdAAAAABAE
- https://www.google.com/url?sa=i&url=https%3A%2F%2Fen.wikipedia.org%2Fwiki%2Fscikit-learn&psig=AOvVaw0mRroBY9wV5OHhWaPilNVS&ust=1665237976555000&source=images&cd=vfe&ved=0CAwQjRxqFwoTCKDGsYKlzvoCFQAAAAAdAAAAABAE