PROJECT ABSTRACT

Project for MATH 678- Statistics in Data science
Members: Micheal Woo(mw47), Veena Chaudhari(vac38)

Type of Project: Real Data Analysis

Title: **U.S. Patent Phrase to Phrase Matching**

Dataset:
https://www.kaggle.com/competitions/us-patent-phrase-to-phrase-matching/data?select=train.csv

This project aims to Identify Similar Phrases in the U.S. Patents thereby helping the patent community connect the dots between millions of patent documents. The data for this project is derived from the public archives of the U.S. Patent and Trademark Office (USPTO). These archives, offered in machine-readable formats.

In this dataset, there are pairs of phrases (an anchor and a target phrase) and asked to rate how similar they are on a scale from 0 (not at all similar) to 1 (identical in meaning).

The Training dataset contains the attributes:

Table 1: Table for training dataset

Attribute name	id	anchor	target	context
Data type	numeric	string	String	string
Description	a unique identifier for a pair of phrases	the first phrase	the second phrase	the CPC classification (version 2021.05), which indicates the subject within which the similarity is to be scored

Label is the score value which is numeric value ranging from 0 to 1 with :

- 1.0 - Very close match.(exact match except possibly for differences in conjugation, quantity)
- 0.75 - Close synonym
- 0.5 - Synonyms which don’t have the same meaning
- 0.25 - Somewhat related,
- 0.0 - Unrelated.

The model determines how the different phrases in the test csv are related.

In this project the training dataset is divided into a training set and validation set to report the accuracy of the models.

Implementation:

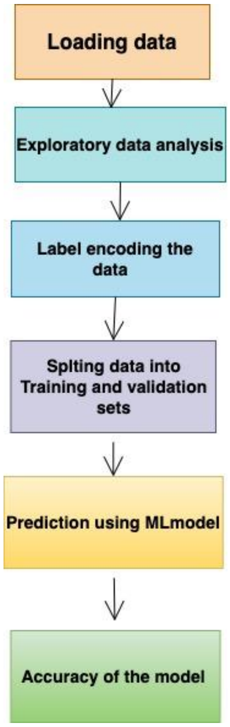

Figure 1: implementation of project

Candidate ML models:

1) Random forest
2) Linear regression

Accuracy: We plan to use mean absolute error for predicting the accuracy of the model.