Phrase Matching Similarity: NLP Methods

Project Type: Real Data Analysis
MATH-678: Statistics in Data Science
Michael Woo & Veena Chaudhari

Abstract

This project aims to Identify Similar Phrases in U.S. Patents, thereby helping the patent community connect the dots between millions of patent documents. Data comes from the U.S. Patent and Trademark Office (USPTO) public archives. Archives are in CSV (Comma-Separated Values) 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).

Dataset

Dataset was obtained from Kaggle. Link is here. The training dataset contains the following attributes below.

Attribute Name	Data Type	Description
id	Numeric	Unique identifier for each pair of phrases
anchor	String	The first phrase (Predictor)
target	String	The second phrase (Predictor)
context	String	The CPC (Cooperative Patent Classification) Classification, which indicates the subject within which the similarity is to be scored (Predictor)
score	Numeric	A similarity scores that ranges from 0 to 1. (Response)

Table 1: Training Dataset Layout

The response is *score* that ranges from 0 to 1. Here is a breakdown of each similarity score means:

Score Value	Meaning
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	Unrelated

Table 2: Meaning of Scores

Implementation

In this project, we plan to use the following ML (Machine Learning) models: Random Forrest Regression and Logistic Regression. The dataset meant for training will be divided into two sets: the training set and the validation set to report the accuracy and fit of the model. Figure 1 is a layout of the step blocks we plan on doing in the project. As for model evaluation, we plan to use the following metrics: MAE (Mean Absolute Error), MSE (Mean Square Error)/RMSE (Root Mean Square Error), and R square/Adjusted R squared.

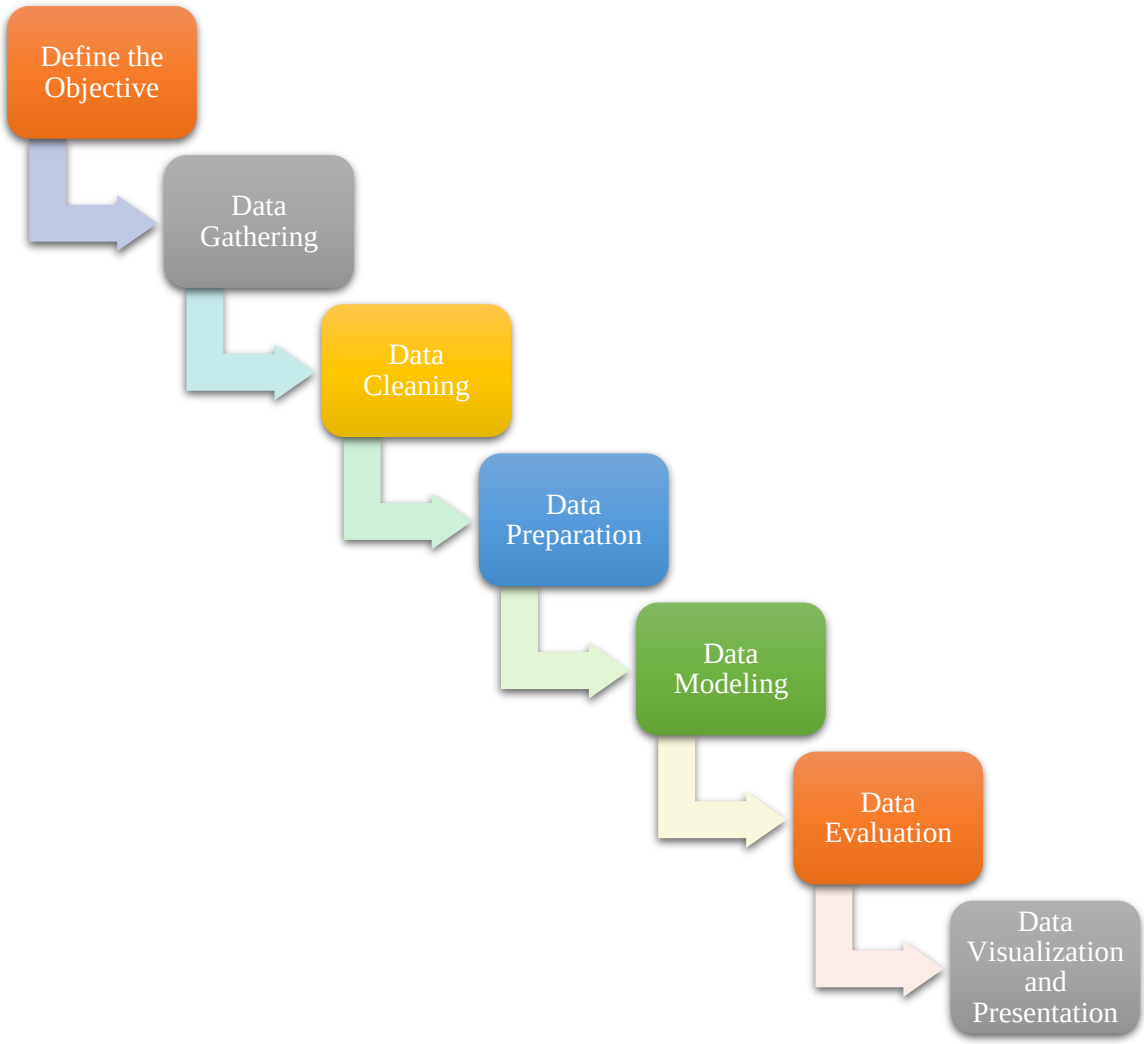

Figure 1: Project Implementation