Abstract

Quora is a place to gain and share knowledge—about anything. It's a platform to ask questions and connect with people who contribute unique insights and quality answers. This empowers people to learn from each other and to better understand the world.

Design

This project is one of the T5-Batch 3 Data Science BootCamp requirements. dataset created from dataset already on Kaggle '**Quora Question Pairs'**

In the Quora Question Pairs competition, we were challenged to tackle the natural language processing (NLP) problem, given the question pairs, classify whether question pairs are duplicates or not

Data

The dataset is provided in .csv format. It is a dataset of 404290 question pairs along with the corresponding label (duplicate or not) , each qoura has 6 features. The goal of this competition is to predict which of the provided pairs of questions contain two questions with the same meaning. The ground truth is the set of labels that have been supplied by human experts. The ground truth labels are inherently subjective, as the true meaning of sentences can never be known with certainty. Human labeling is also a 'noisy' process, and reasonable people will disagree. As a result, the ground truth labels on this dataset should be taken to be 'informed' but not 100% accurate, and may include incorrect labeling. We believe the labels, on the whole, to represent a reasonable consensus, but this may often not be true on a case by case basis for individual items

Algorithms

1. **Read data**
   From kaggle website, which contains more than 400,000 rows and 6 random randomized plots consisting of 10,000 points.
2. **We looked at the data**

- Visualization data

- Delete rows that contain null values

- Text preprocessing :

- Removing multiple spaces ,single characters from the start,Punctuations, capital letters,stopwords

- And apply lemmatize for text

3. **Features Engineering**
   - new columns that contain the length of the questions,
   - new columns that contain the number of words and
   - new columns that contain the percentage of convergence between each question and another by use jaccard simalrity and cosine simalrity
4. **Modeling**

- Convert text to numbers using TFIDF

- Four types of classification models were used for machine learning (LogisticRegression, KNeighborsClassifier, RandomForestClassifier, DecisionTreeClassifier)

- Choosing the best scoring in validation

Tools

- Python and Jupyter Notebook
- NumPy and Pandas for data manipulation
- NLTK, tashaphyne, and Farasa for NLP preprocessing
- Sklearn for ML algorithms
- Matplotlib, pyLDAvis and wordcloud for visuializations