Data Mining Proposal

Luis Duque
Joa Jin
Ethan Leeman
Yingnan Liu
Sophia Zheng

March 24, 2017

1 Introduction

For our project, we propose competing in the most recent Kaggle competition, "Quora Question Pairs." Quora is a question and answer website where users can both submit and answer questions in a collaborative manner. While handling a large volume of queries, duplicate questions are asked often, making it difficult for popular, accepted answers to be found. Currently Quora uses a Random Forest model to identify identical questions, and proposed this Kaggle competition to tackle this machine learning problem in natural language processing.

2 Data

Kaggle provides a training dataset and a test dataset. In the training set, there are approximately 400,000 question pairs and the target variable, whether Kaggle feels the questions are duplicates. Some questions appear multiple times in different pairs. For example, the questions "How can you determine the first ionization energy of lithium?" and "How is the ionization energy of silicon determined?" are distinct while the questions "Will Donald Trump shut down the internet?" and "If Donald Trump becomes president will we lose the internet?" are duplicate.

One issue with the data is the subjectivity of the target variable. For example, the question pair ”I am 24. Is it too late to get into medicine?” and ”Is it too late to study medicine at 23?” could be reasonably interpreted as duplicate or as distinct (note that it is marked as distinct). Additionally, there is some natural human error and noise in the data. The question pair ”What is the cultural shock?” and ”What is Culture Shock?” is marked as distinct, while these are clearly the same question. Another issue is that the characters are in unicode, and sometimes questions with non-english words appear.

The test data is about 2,000,000 question pairs and the competitors are asked to submit a probability of being duplicate, and the submissions are scored on Logarithimic Loss.

3 Possible Approaches

Kaggle provides a discussion board, and rewards competitors for providing helpful guides and comments to other commenters. In particular, there is a ”beginner’s guide” by ”shubh24” which provides many techniques in natural language processing.

- One first idea is to take every sentence, create a set of every k strings of words, after stripping the stopwords, and compare the overlap by some metric. For $k = 1$ we would measure how distinct the words are in each question, $k = 2$ would be word-pairs, and so forth. One key issue is that a single word, even if the rest of the sentence is identical, can drastically change the question, as in the lithium vs. silicon example above.

- Wordnet is an English dictionary which also contains a directed graph showing relationships between words. One possible feature we could make would be to find the rare words and see how distant they are. This would try to find questions that are asked with different synonyms or similar phrases.

- –Ethan stopped here–

Outline The remainder of this article is organized as follows. Section 4 gives account of previous work. Our new and exciting results are described in Section 5. Finally, Section 6 gives the conclusions.

4 Previous work

A much longer LaTeX 2_ε example was written by Gil [1].

5 Results

In this section we describe the results.

6 Conclusions

We worked hard, and achieved very little.

References

[1] J. Y. Gil. LaTeX 2_ε for graduate students. manuscript, Haifa, Israel, 2002.