Data Analytics and Cognitive Computing Group Project

Detecting Duplicate Questions in Quora Dataset

Victoria Sardelli Orens Xhagolli

1. INTRODUCTION

The goal of our project is to produce a model that will determine if two questions are duplicates of each other. Let's take two different questions for example, "How can I become a good geologist?" and "What should I do to become a great geologist?" These questions have the same intended meaning and should therefore lead to the same body of knowledge to get the answer. On the other hand, "Where was George Washington born?" and "When was George Washington's birthday?" are two different questions and they should not necessarily lead to the same body of knowledge. Is it possible for a computer to confidently determine that two seemingly distinct questions carry the same meaning? We will be using data provided by Quora to achieve a better understanding of this problem and attempt to find a solution.

Quora is a website in which a wide variety of questions are asked and answered by its online community. In any question-and-answer resource, the presence of duplicate questions can cause inefficiency in answer retrieval. If the same question is asked in multiple different ways and the system does not recognize them as logically-equivalent, then a distinct web page would exist for each way in which the question was asked. As a result, answers to that question would be divided across multiple locations, and knowledge seekers would find different answers depending on the page they visit. In the worst case, people may never see the most helpful and correct answers if they are unaware that those answers exist in a different location. Therefore, if all the answers to logically-equivalent questions were consolidated on the same page, Quora users would be able to access their desired answers in one location and writers would be able to increase their audience. We will be developing a model that will attempt to determine if two questions are duplicates. If successful, this would be useful in a real-world context because duplicates in question-and-answer resources such as Quora would be easily able to identify duplicates in their systems and take proper action to resolve such issues.

We will be using the "Question Pairs Dataset" provided by Quora on Kaggle.com.[1] This dataset has over 400,000 rows of question pairs that may or may not be duplicates. Each row contains an ID for each question in the pair, as well as the full text for each question and a binary value indicating if the pair of questions are duplicates. A value of 0 represents two questions that are not duplicates, and a value of 1 represents duplicates. For example, the questions "What is the step by step guide to invest in share market in india?" and "What is the step by step guide to invest in share market?" have a value of 0 for their binary attribute, while the questions "Astrology: I am a Capricorn Sun Cap moon and cap rising...what does that say about me?" and "I'm a triple Capricorn (Sun, Moon and ascendant in Capricorn) What does this say about me?" have a value of 1.

Using this dataset, we intend to build at least 3 different models and compare their performance and accuracy through IBM Watson. Most importantly, we will try to document every step and approach that we take and try to learn from our mistakes while correcting them along the way. So the final goal for us is to learn as much as we can from this project and try approaches that we haven't used/encountered before, while it would be nice to come up with an extremely accurate model as a by-product. Since this data is publicly available, others have been able to use it for similar purposes and have discussed their findings on Kaggle. As a result, we will be able to compare our results to those of previous researchers in the Kaggle community.

2. REFERENCES

[1] Quora. Question pairs dataset. January 2017.