UNIVERSITAT DE BARCELONA

NATURAL LANGUAGE PROCESSING

MASTER IN FUNDAMENTAL PRINCIPLES OF DATA SCIENCE

ASSIGNMENT 1

QUORA CHALLENGE

Authors
Irene Bonafonte Pardàs
Otis Carpay
Vladislav Nikolov Vasilev

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

ACADEMIC YEAR 2021-2022

Contents

1 Introduction

In this assignment we are going to try to solve the Quora Question Pairs challenge. Given a pair of questions, we have to automatically determine whether they are semantically equivalent or not. The goal of this is to reduce the number of duplicate questions and improve the overall user experience.

In order to solve this challenge, we are fist going to try a simple solution which will allow us to get a better understanding of the problem and identify possible flaws. After that, we are going to refine this initial solution in hopes of obtaining a model that is more robust and better able to identify duplicate questions.

2 Simple solution

As a simple solution, we are going to train a logistic regression classifier in order to detect duplicate questions. Since we cannot feed text data directly into the model, we have to use some other kind of numerical representation. In this case, we can use the bag-of-words representation in order to encode the questions.

When following this approach, we have run into some technical problems. One of them is that not every question is represented as a string. This has forced us to encode each one of the questions properly before trying to get the bag-of-words representation.

Because this approach is quite simple, it has some inherent limitations:

- No text preprocessing is applied, apart from the basic preprocessing that the `CountVectorizer` class from `scikit-learn` performs. This means that the corpus is filled with misspelled words, words spelled in different ways (for example, *e-mail* and *email*) and stop words that are not quite relevant.

- Even though the bag-of-words representation allows us to encode the questions using numerical values, it ends up falling short because it only considers the word frequency inside the document. For instance, it could also consider how many time a word appears in the whole corpus, which might be important when trying to identify relevant words. Also, there might be better ways to encode words, like using embeddings and combining them in some kind of fashion in order to create sentence embeddings.

- Using only the bag-of-words representation may not be enough. We can try to create custom metrics that allow us to capture the distance between the questions and use them in the training process, whether it is a custom metric or some kind of metric that allows us to capture the semantic difference between the sentences. Then, we can either use this metrics on their own or combine them with some other kind of representation.

3 Improved solution

3.1 Creating better feature vectors and distances

3.1.1 Irene's features

3.1.2 Otis' features

3.1.3 Vladislav's features

4 Final results

References

[1] Texto referencia
`https://url.referencia.com`