



UNIVERSITAT_{DE}
BARCELONA

NATURAL LANGUAGE PROCESSING

MASTER IN FUNDAMENTAL PRINCIPLES OF DATA SCIENCE

ASSIGNMENT 1

QUORA CHALLENGE



DATA SCIENCE @ UNIVERSITAT DE BARCELONA

Authors

Irene Bonafonte Pardàs

Otis Carpay

Vladislav Nikolov Vasilev

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

ACADEMIC YEAR 2021-2022

Contents

1	Introduction	2
2	Simple solution	2
3	Improved feature vectors and distances	2
3.1	Irene’s features	2
3.2	Otis’ features	2
3.3	Vladislav’s features	2
4	Final results	2
	References	3

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 first 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

3 Improved feature vectors and distances

3.1 Irene's features

3.2 Otis' features

3.3 Vladislav's features

4 Final results

References

- [1] Texto referencia
<https://url.referencia.com>