Project Proposal Report

PROJ 201 Project Proposal Report

Project Title: Searching most accurate and existing methods to detect the language of texts

Name, Surname & ID of group members: Osman Enes Erdoğdu & 29344, Sinem Gül Kaya & 30298

Supervised by: Yücel Saygın

Month, day, year: 04/11/2022

Abstract

The project is about to understand NLP which a way to teach computers our language. Also, we will use machine learning to learn natural language process better. In this project we will specifically work on texts to categorize if they are English or not. Also, we will go on some specific steps that starts with data collection and processing. Then finding some models to compare our aim which is understand a text if it is English or not. Then we will choose the most appropriate and correct methods from all of them. Finally, we will evaluate our work and clean it. We will separate unnecessary parts from our data’s. Also, at the bottom there is a part that shows specific steps of our work week by week and who’s responsibility of that specific step is.

Introduction

Natural language processing is about modelling of computational linguistics, which is basically a tool to understand some abilities that human can do such as communication, and to adapt it to develop artificial intelligence. Some NLP tools facilitate and help the computer to analyze and understand the human voice and text and make it clear to annotate them. Understanding the use of human language and adapting it to the virtual environment is very difficult and complex for computers to understand. There are many texts that written on different languages, this project is about to determine which language of used in the text. To able to do that, there are some methods and programs that identify languages of texts. In this project it will be considered to find clear method and algorithm.

Purpose

The aim of this project is to specifically search existing methods for detecting texts whether they are fully English or not, and to compare models with each other to find the most accurate and applicable one.

Description of Specific Steps

The first step of the project is collecting and choosing data from the online appropriate sources. Secondly, we will skim through, manually analyze and put some annotations to data. The third step is about preprocessing the data, in this step we will consider that if the data are suitable for our project. Then, we will remove the unnecessary parts and characters such as link, wrong punctuations, extra spaces. The other step is selecting the best model or models that we searched to achieve our aim. Finally, we will examine the methods that we choose and analyze how accurate are they.

Description of Responsibilities of Individual Members

We started with determining the project and getting ready. In the sixth week, Osman will collect data and then, he will select the most suitable and related data to our project. After the choosing data, Sinem will do data annotation and she will work on the data to make it clearer and more understandable. The tenth week Osman will identify and explain the models. And the next week, Sinem will skim through the models and compare them to determine which method is more suitable and accurate. Finally, we will decide our method evaluate our project.

Gantt Chart

