Language Identification using Deep Learning

Ahmet Tarık KAYA
Computer Engineering
Hacettepe University
Çankaya, Ankara
ahmet.tarik.kaya@gmail.com

Kaan Mersin
Computer Engineering
Hacettepe University
Çankaya, Ankara
kaanmersin07@gmail.com

1. Introduction

The topic of this project is Language Identification by using Natural language Processing methods. There are a lot of data in this topic becuase most of the platforms in internet, users are using different languages. By this different datasets are forming that contains different languages that used in same platforms.

There are many models that have been developed in this subject. But mostly these models are using same methods. Forexample some of them only using words that has at most four letters. But they are not using deep learning. In this project we are planing on starting with a small group on languages to develop a better model that is using deep learning.

2. Dataset

This is the dataset that we are planning to work on

- https://www.kaggle.com/alvations/old-newspapers

3. Related Papers

Here are some related papers on our project topic

- Language Identification: a Neural Network Approach
  - https://pdfs.semanticscholar.org/17b2/c9ae27d2ef1b6b901a0afd5aa8649f7795bf.pdf
- Text Language Identification Using Attention-Based Recurrent Neural Networks
  - https://www.researchgate.net/publication/333392357_Text_Language_Identification_Using_Attention-Based_Recurrent_Neural_Networks
- Language Identification from Text Documents
  - http://cs229.stanford.edu/proj2015/324_report.pdf

4. Schedule

Deadlines	Phase
26 April Sunday	Proposal
8 May Friday	Trying the related projects and learning the algorithms and testing our eliminations on the dataset
15 May Friday	Working on the dataset and creating the training and testing data(eliminating some part of the dataset after researches)
21 May Thursday	Progress Report
5 June Friday	Creating our model
15 June Monday	Testing and finishing the model
18 June Thursday	Final Report