Faculty of Engineering & Applied Science

SOFE 4620U: Machine Learning & Data Mining

Date: Feb 6, 2023

Language Identification
Project Proposal

Group 8

Name	Student Id
Preet Patel	100708239
Tiwaloluwa Ojo	100700622
Manreet Kaur	100766207
Aaditya Rajput	100622434

Introduction/Problem Statement

Natural Language Processing (NLP) is a branch of artificial intelligence associated with enabling computers or machines to understand human languages in different forms (text or speech) [1]. One of the problems that is faced with NLP is to give the computer the ability to first identify what language it is currently processing so that it can then use this information to appropriately understand the input and then use the input for various functions. Without accurate language identification, it is impossible to process and understand any language. Language identification is the first step in any problem within the NLP domain. For this project, our group will be creating a model that will be able to accurately identify languages from any text or documents on the web.

Project Background

NLP is a primary driver in language translation through machine learning. It is an important tool in translating and responding to spoken commands and textual input [1]. It is often observed in dictation softwares and GPS systems [1]. NLP has various challenges it must overcome in order to accurately translate its input. Some of these challenges are speech recognition and converting speech to a textual format, grammatical tagging and understanding the context of the words used, natural language generation, and extracting subjective qualities such as attitudes, emotions, sarcasm, etc. [1].

There are many use cases for NLP. These include but are not limited to spam detection/filtering, machine translation, chatbots, and linguistic blending [1][2]. The most famous use case of NLP is spam detection/filtering which identifies the language and its features such as poor grammar and threatening language to flag emails for spam detection [2].

Possible Solutions

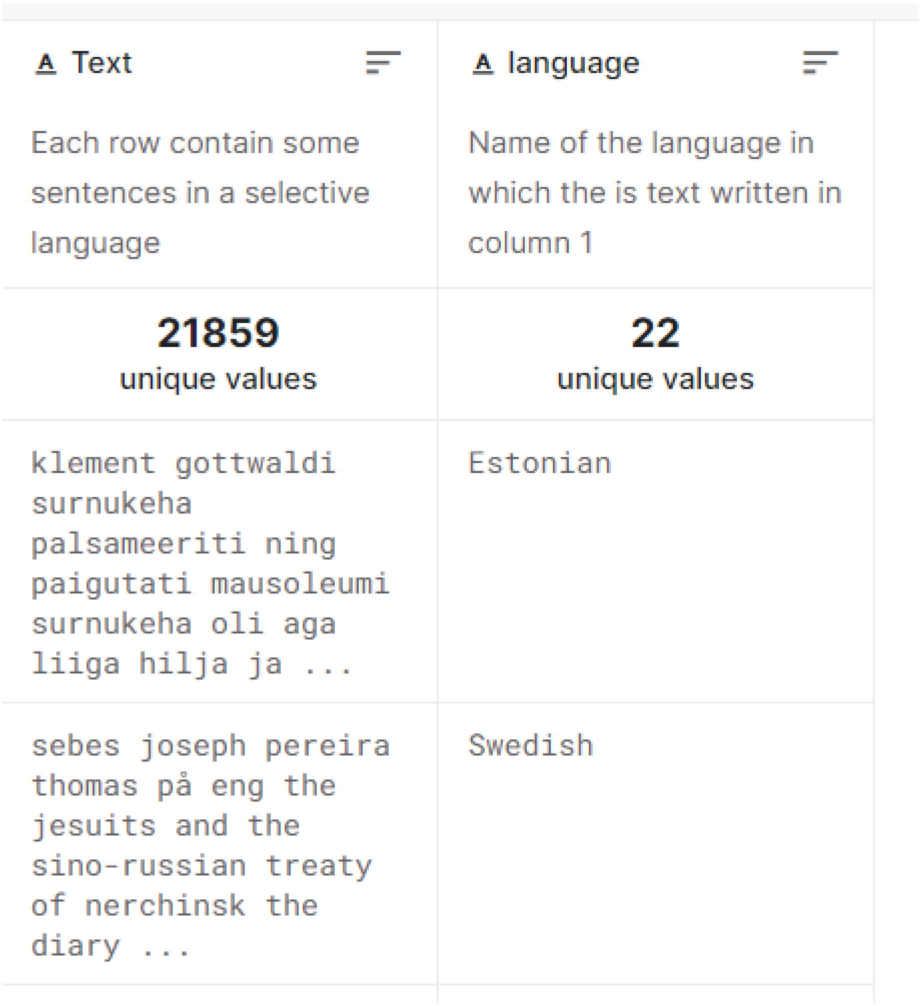
Some solutions used within the industry include the NLP Toolkit (NLPTK) and statistical NLP in the python language [1]. The NLPTK is a series of libraries for many NLP tasks such as sentence parsing and word segmentation, while the statistical method uses manually coded ML models. Moreover, the statistical NLP also uses various deep learning models to extract and classify elements within the provided input to produce a probability of a match to the proposed elements [1]. Today various statistical methods use RNNs and CNNs to learn and optimize themselves while digesting raw input and datasets [1]. Other possible solutions include using python frameworks discussed in the course such as Scikit-Learn to train a model which can distinguish between different human languages.

Dataset

We will be using the WiLi-2018 dataset which contains 235000 paragraphs of 235 languages. This dataset is known as the Wikipedia language identification benchmark . WiLi-2018 dataset have data splitted into two parts , one if training data and the other one is test data .[2]

This is the link to the dataset :
https://huggingface.co/datasets/wili_2018/viewer/WiLI-2018%20dataset/train

We will choose languages from this dataset which best matches our project needs and are most commonly used in the world . For example , from the above dataset , we can choose 20 selective languages and create a new dataset from it . The diagram below displays the first two rows of the selective dataset .



Text	language
Each row contain some sentences in a selective language	Name of the language in which the is text written in column 1
21859 unique values	**22** unique values
klement gottwaldi surnukeha palsameeriti ning paigutati mausoleumi surnukeha oli aga liiga hilja ja ...	Estonian
sebes joseph pereira thomas på eng the jesuits and the sino-russian treaty of nerchinsk the diary ...	Swedish

References

[1] "What is Natural Language Processing? | IBM,"   Ibm, https://www.ibm.com/topics/natural-language-processing (accessed Feb. 05, 2023).

[2] S. Ayodeji, "Using machine learning for language detection | by Shittu Olumide Ayodeji | Heartbeat,"   Heartbeat, Nov. 7, 2022. https://heartbeat.comet.ml/using-machine-learning-for-language-detection-517fa6e68f22 (accessed Feb. 05, 2023).

[3] M. Thoma, *Datacite Search*, 07-Jan-2018. [Online]. Available: https://search.datacite.org/works/10.5281/zenodo.841984 .(Accessed: 06-Feb-2023).