Automatic Language Identification

Using k-means

NAME	ID	GRADE	LEVEL	COMMENT
Amir Samir Fawzy Gaber (T.L)	202000160			
Abdelrahman Sameh Sayed	202000515			
Amr Abdelkhalek Abdellah Aly	202000616			
Ibrahim Saad Mohamed	201900008			
Fady Osama Ekram	202000621			
Mariam Eid Kamel Melika	202000882			
Youstina Gamal Aziz Iskander	202001068			

Team Members: Team Number -> 152

SUBMISSION AND DISCUSSION OF A.I

DOCUMENTATION:

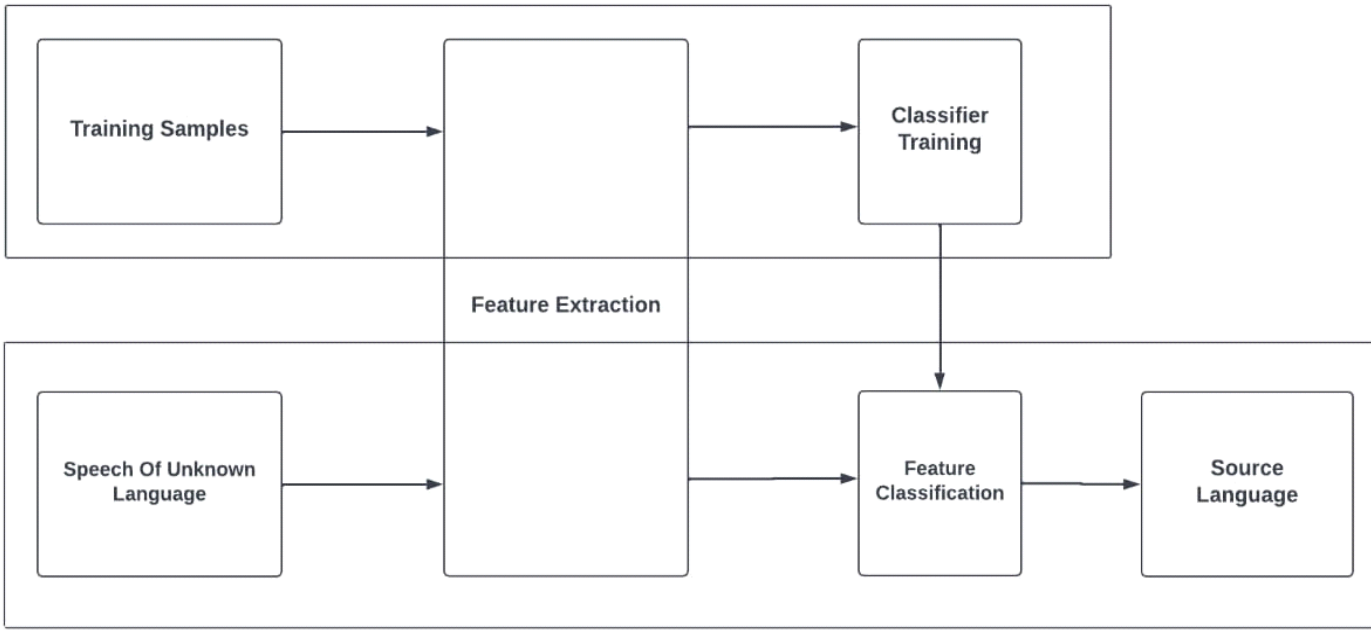
- **The Project idea in detail**

This provides a detailed introduction to past and present approaches to automatic language identification (LID) and discusses their strengths and weaknesses with respect to their practical applicability. LID is the task of automatically

recognizing language from a spoken utterance. LID plays an essential part in providing speech applications to a large, multilingual user community. These may include multilingual spoken dialog systems, spoken-document retrieval, and multimedia mining systems, as well as human-to-human communication systems. The core problem in solving the LID task is to find a way of reducing the complexity of human language such that an automatic algorithm can determine the language identity from a relatively brief audio sample. The ultimate role model for automatic LID systems is the human listener. Human listeners are capable of recognizing a language even from extremely short audio samples, provided they have a certain degree of familiarity with the language. Familiarity can vary from full knowledge of the lexicon, grammar, and pronunciation with native or near-native proficiency to a simple acoustic experience.

The core problem in solving the LID task is to find a way of reducing the complexity of human language such that an automatic algorithm can determine the language identity from a relatively brief audio sample. Differences between languages exist at all linguistic levels and vary from marked, easily identifiable distinctions (such as the use of entirely different words) to more subtle variations (e.g. the use of aspirated vs. unaspirated syllable-initial plosives in English vs. French). The latter end of the range is a challenge not only for automatic

A considerable increase in the amount of and access to data provided not only by experts but also by users all over the Internet has resulted in both the development of different approaches in the area of LID – so as to generate more efficient systems – as well as major challenges that are still in the eye of the storm of this field. Nowadays, LID systems are being used in connection with different fields; although the same basic approaches introduced and developed in the 1990s are still in use. Despite the fact that the current approaches have accomplished considerable success, future research concerning some issues – especially greater incorporation of semantic content in the different LID systems – remains on the table. The field of LID activity dates back to the 1970s, and a considerable number of methods have been developed in its furtherance. Due to the requirements that rule the following project, the goal shall not be to describe the historic background of this field of studies, but rather to provide an overview of the current state of LID systems, as well as to classify the approaches developed to accomplish them.



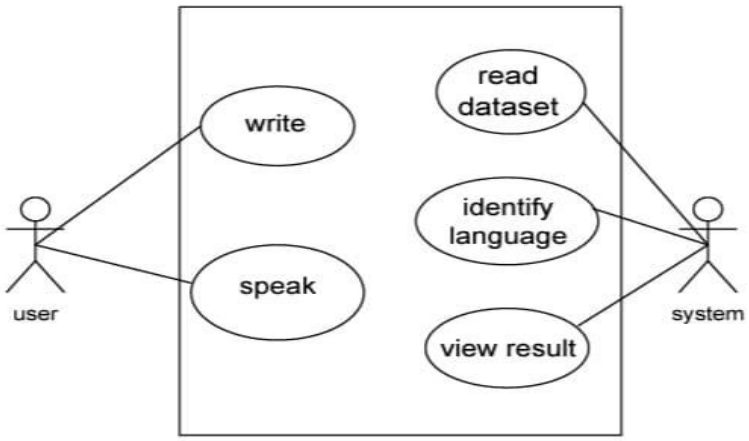
Block Digram Of Language Identification Using K-mean Clustering

---

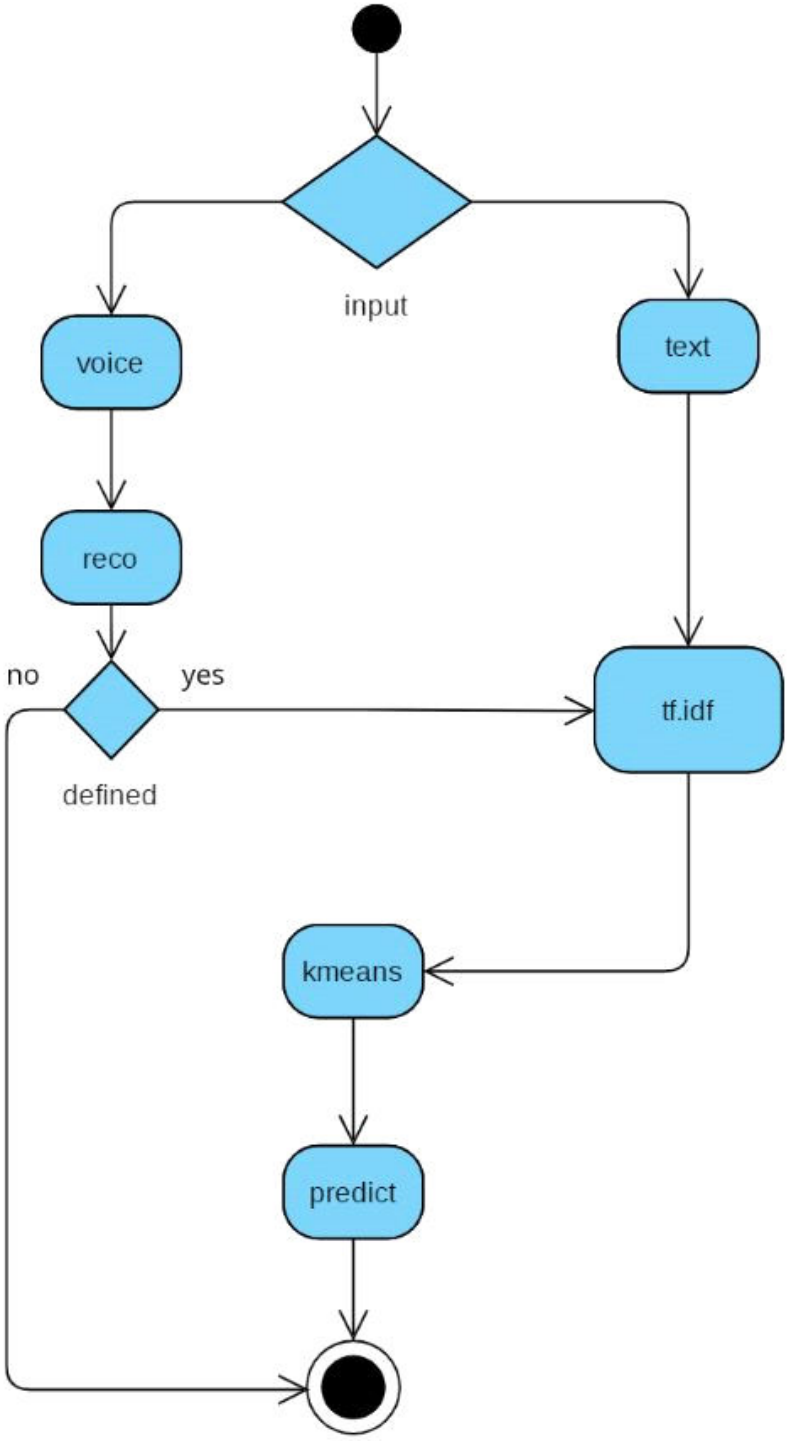
2-Main functionalities:

1. Take a voice record then the application tries to identify the language.

2. Take a plain text then the application tries to identify the language.



Usecase diagram of a language identifier functionalty



Activity diagram of language identifier functionality

3- Similar applications in the market:

K-mean is very popular algorithm in the market for example k-mean has been used in document clustering, market segmentation, face recognition, Clustering treatment options within a cohort to make data-driven decisions, identifying similar patients based on their attributes to explore costs, treatments, or outcomes, etc. All applications undergo a cluster analysis so we cluster data then predict where different models will be built for different subgroups.

4- An initial literature review of Academic publications relevant to the idea:

1)paper 1: Automatic Detection and Language Identification of Multilingual Documents Marco Lui, Jey Han Lau and Timothy Baldwin Department of Computing and Information Systems The University of Melbourne NICTA Victoria Research Laboratory Department of Philosophy King's College London mhlui@unimelb.edu.au, jeyhan.lau@gmail.com, tb@ldwin.net

The conclusion:

We have presented a system for language identification in multilingual documents using a generative mixture model inspired by supervised topic modeling algorithms, combined with a document representation based on previous research in language identification for monolingual documents. We showed that the system outperforms alternative approaches from the literature on synthetic data, as well as on real-world data from related research on linguistic corpus creation for low-density languages using the web as a resource. We also showed that our system is able to accurately estimate the proportion of the document written in each of the languages identified. We have made a full reference implementation of our system freely available,8 as well as the synthetic dataset prepared for this paper (Section 5), in order to facilitate the adoption of this technology and further research in this area. 8https://github.com/saffsd/polyglot

- Paper 2: Automatic Language Identification in Texts: A Survey Tommi Jauhiainen tommi.jauhiainen@helsinki.fi Department of Digital Humanities The University of Helsinki Marco Lui saffsd@gmail.com School of Computing and Information Systems The University of Melbourne Marcos Zampieri m.zampieri@wlv.ac.uk Research Institute in Information and Language Processing

University of Wolverhampton Timothy Baldwin tb@ldwin.net School of Computing and Information Systems The University of Melbourne Krister Linden krister.linden@helsinki.fi Department of Digital Humanities The University of Helsinki

The conclusion

This article has presented a comprehensive survey on language identification of digitally-encoded text. We have shown that LI is a rich, complex, and multi-faceted problem that has engaged a wide variety of research communities. LI accuracy is critical as it is often the first step in longer text processing pipelines, so errors made in LI will propagate and degrade the performance of later stages. Under controlled conditions, such as limiting the number of languages to a small set of Western European languages and using long, grammatical, and structured text such as government documents as training data, it is possible to achieve near-perfect accuracy. This led many researchers to consider LI a solved problem, as argued by McNamee (2005). However, LI becomes much harder when taking into account the peculiarities of real-world data, such as very short documents (e.g. search engine queries), non-linguistic “noise” (e.g. HTML markup), non-standard use of language (e.g. as seen in social

media data), and mixed-language documents (e.g. forum posts in multilingual web forums). Modern approaches to LI are generally data-driven and are based on comparing new documents with models of each target language learned from data. The types of models as well as the sources of training data used in the literature are diverse, and work to date has not compared and evaluated these in a systematic manner, making it difficult to draw broader conclusions about what the “best” method for LI actually is. We have attempted to synthesize results to date to identify a set of LI “best practices”, but these should be

---

5-DataSet Employed:

We are using a dataset from Kaggle that contains 235000 paragraphs of 235 languages. Each language in this dataset contains 1000 rows/paragraphs. Our dataset contains 4 selective language which includes

1- English

2- Arabic

3- French

4- Dutch

we have 4 languages to train our model within form of excel sheet; So, we used pandas library to read it and k-means library to train these languages. For speech recognition we used speech recognition, for vectorizing the strings taken from the user we used TfidfVectorizer library for clustering we used KMeans library from sklearn. Cluster.

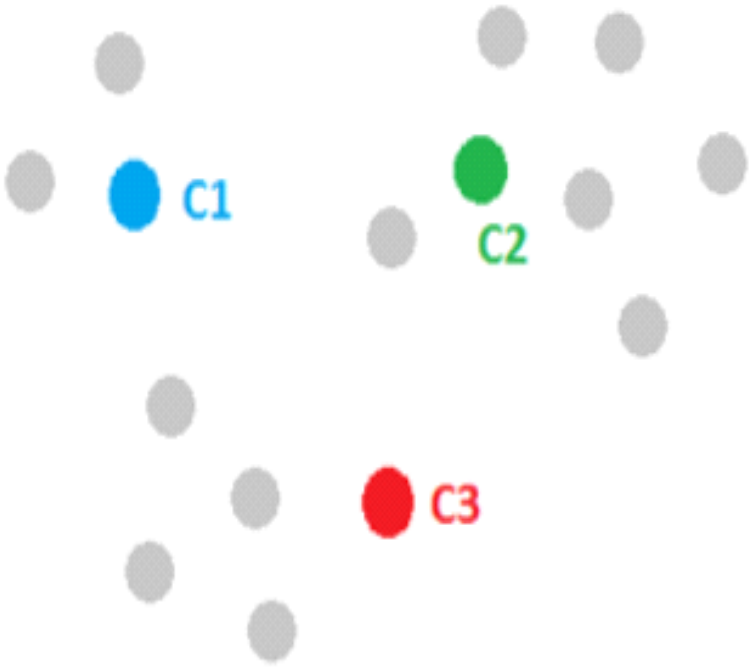
Dataset link:

https://www.kaggle.com/zarajamshaid/languageidentification-datasst

---

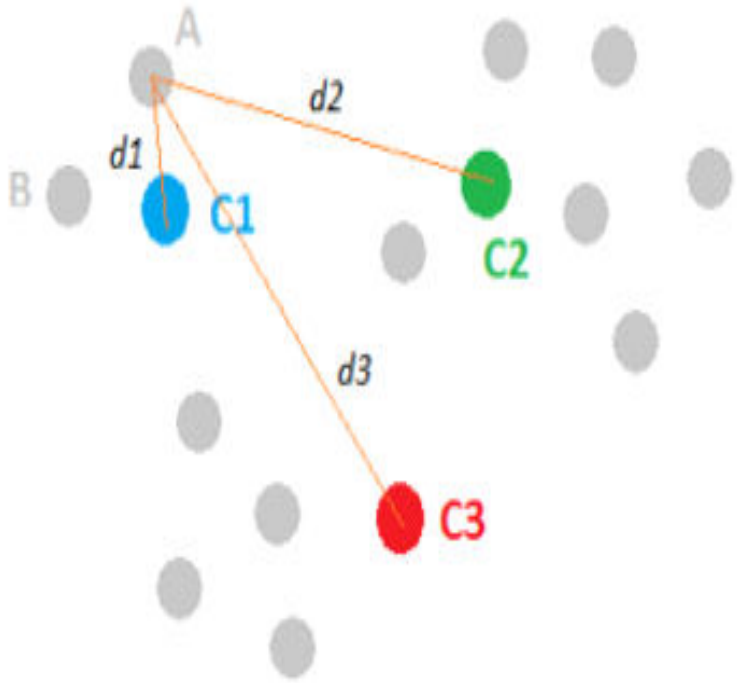
6- Details of the algorithm(s)/approach(es) that will be used:

Step one: Initialize cluster centers:



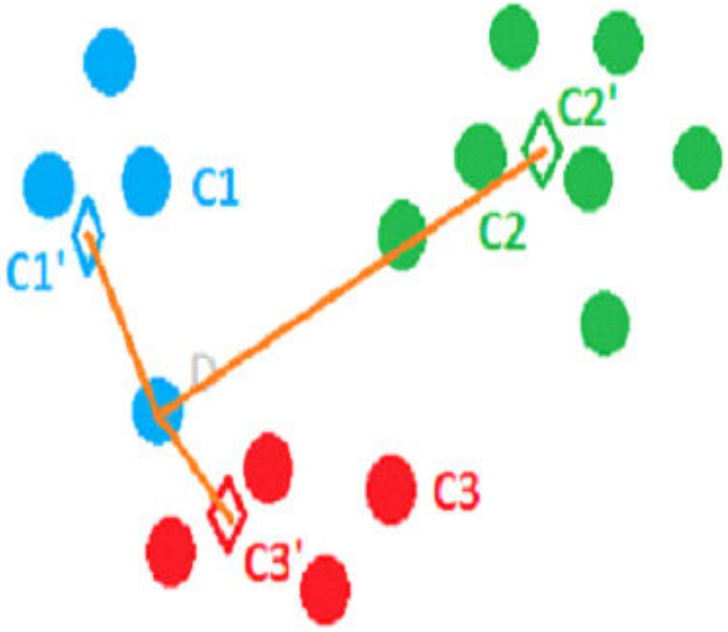
Initialize cluster centers We randomly pick k points and label them with separately to represent the cluster centers.

Let k=3 then we will pick three-point c1, c2, c3 and label them with blue, green, and red



Step two: Assign observations to the closest cluster center:

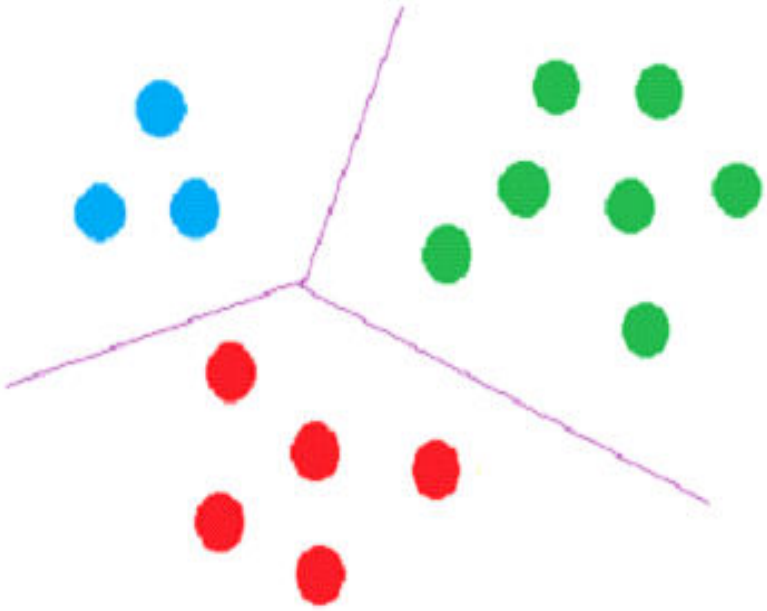
Once we have these cluster centers, we can assign each point to the clusters based on the minimum distance to the cluster center. For the gray point A, compute its distance to C1, C2 and C3, respectively. And after comparing the lengths of d1, d2 and d3, we figure out that d1 is the smallest, therefore, we assign point A to the blue cluster and label it with blue. We then move to point B and follow the same procedure. This process can assign all the points and leads to the following figure.



Step three:

Revise cluster centers as mean of assigned observations:

Now we have assigned all the points based on which cluster center they were closest to. Next, we need to update the cluster centers based on the points assigned to them. For instance, we can find the center mass of the blue cluster by summing over all the blue points and dividing by the total number of points, which is four here. And the resulted center mass C1', represented by a blue diamond, is our new center for the blue cluster. Similarly, we can find the new centers C2' and C3' for the green and red clusters.



Step four: Repeat step 2 and step 3 until convergence:

The last step of k-means is just to repeat the above two steps. For example, in this case, once C1', C2' and C3' are assigned as the new cluster centers, point D becomes closer to C3' and thus can be assigned to the red cluster. We keep on iterating between assigning points to cluster centers and updating the cluster centers until convergence.

INPUT explanation:

First User should input a voice of his own so the program identifies his language; so, we have a voice record as an input, then the program transforms if to a text so it can compare the with a data set that is included in the program; so, we a dataset employed and used by kmeans algorithm.

OUTPUT explanation*:*

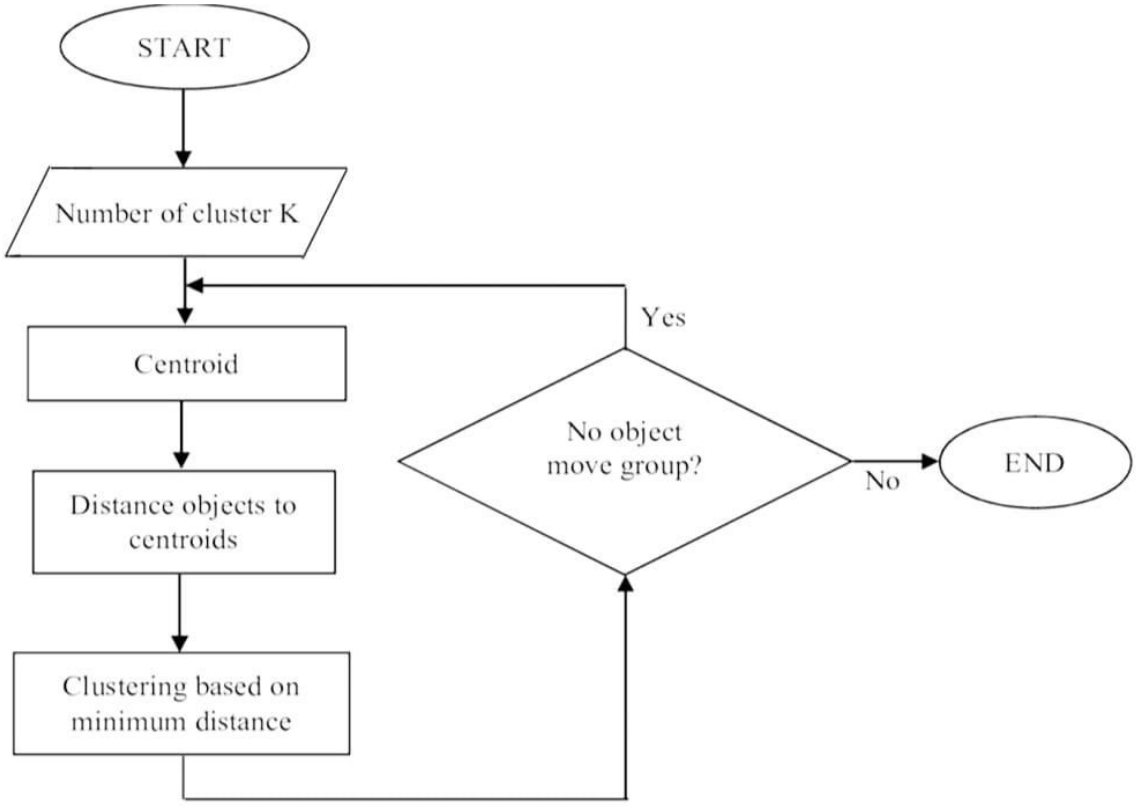
k-means algorithm is used to cluster the dataset so that the program can learn and predict the output correctly we have 4 languages in the dataset. Each language in this dataset contains 1000 rows/paragraphs. Total 4000 learning case. output of kmeans is a number so to obtain the right prediction we had to sort the languages alphabetically and mutably the result by 1000 because every language has 1000 row.

To simplify the output :

Suppose that a language X is sorted to be second language of the list and every language has 1000 row starting from 1001 to 1999 and the result of kmeans algorithm is 2 so it will predict the first language in the list if we did not multiply it by 1000.

Output sequence:

- List of all languages in the system
- Samples of learned languages
- Voice or string input
- Predicted output



GitHub Link:

https://github.com/ASamX/AI-Project-Language-Identification.git