# Automatic Language Identification

# Using k-means

| Name | ID |
|---|---|
| فارينو الفريد فهمي | 201900555 |
| مينا طارق نجيب | 201900878 |
| ماريو منصور صلاح | 20180450 |
| ساندي ابراهيم برسوم | 201900336 |
| انجي عادل فكري | 201900186 |
| احمد مصطفي كامل | 201900102 |
| ماركو ماجد فؤاد | 201900598 |

Drive link:
https://drive.google.com/drive/folders/1IDtAkaPzFvN8ORs1Qnr7Yqk28JNomUiQ?usp=sharing
Shorten URL :
shorturl.at/efhqN

# Main idea of the project (problem definition):

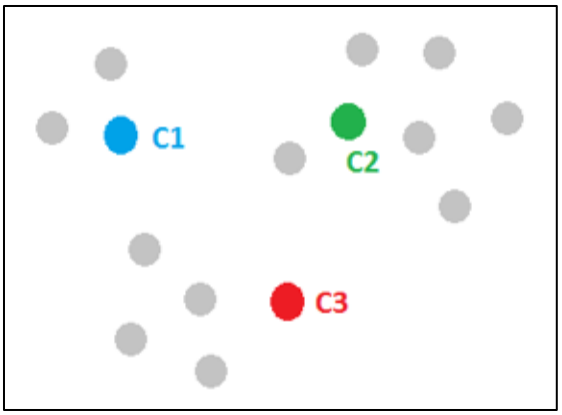
This project aims to identify language ether from a spoken utterance or a plain text. So, the main idea behind this project is to take a voice record from user and convert it to an ordinary text then the program we are creating convert this text to numbers so it can understand it, Here comes k-means clustering. K-means clustering is an AI technique that aims to group N-samples to k-clusters. In other words, every alike samples lie in the same group (cluster).

# More about k-mean cluster:

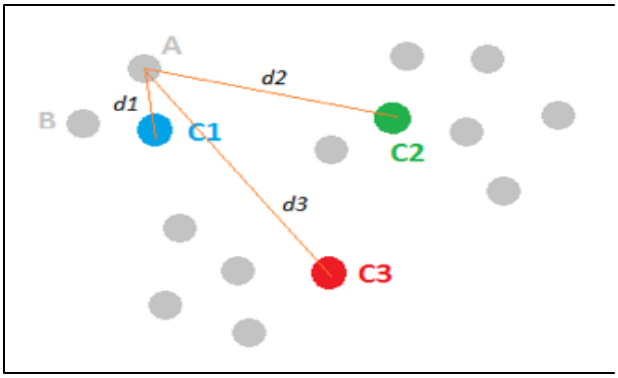
## Step one: Initialize cluster centers

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

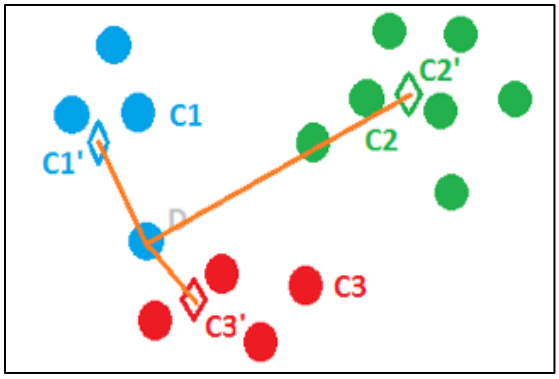


Once we have these cluster centers, we can assign each point to the clusters based on the minimum distance to the cluster center. For the gray point A, compute its distance to C1, C2 and C3, respectively. And after comparing the lengths of d1, d2 and d3, we figure out that d1 is the smallest, therefore, 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.

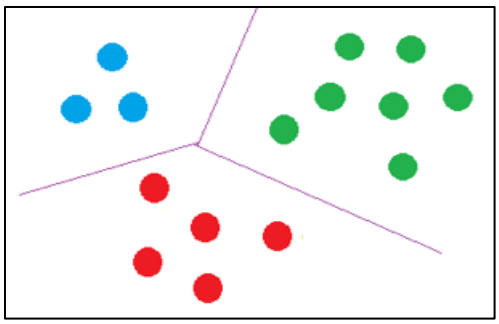



## Step three: Revise cluster centers as mean of assigned observations

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. Finally, we may get a solution like the following figure. Well done!

# Main functionalities:

This project contains two 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.

## Similar applications in the market:

K-mean is very popular algorithm in the market for example k-mean is 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

## Development platform:

In our project, we had worked with Pycharm cross platform. We used pandas library, pandas is one of the important library in machine learning, it used to read from dataset. And we also import nltk and re libraries to make a clean text like remove the stop words from the inputted text and give us the text without extras. We used TfidVectorizer that read the important and special words to count them and discover where these words were repeated. According of that we used KMeans and sr libraries. Sr library is used for taking record from user and transform it to a text.

## 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

## 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

- English
- Arabic
- French
- 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/language-identification-datasst

## 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 language 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

Our references:


1. An Overview of **Partitioning Algorithms in Clustering Techniques** Swarndeep Saket J, Dr. Sharnil Pandya
2. **A Comparative Agglomerative Hierarchical Clustering Method to Cluster Implemented Course,** Rahmat Widia Sembiring, Jasni Mohamad Zain, Abdullah Embong
3. **A study of various Fuzzy Clustering Algorithms,** Nidhi Grover
4. **A Survey of Some Density Based Clustering Techniques,** Rupanka Bhuyan1, Samarjeet Borah2
5. **Foundations of Data Science,** by Avrim Blum
6. **-Practical Guide to Cluster Analysis in R,** by A. Kassambara (Datanovia)


THANK YOU