India - Language Diversity

Final Capstone Project by Yogya Tewari

Introduction

With a population of over 1.3 billion, it is a well-known fact that India is a diverse country.

The Indian Constitution recognizes 22 languages.

A census conducted in 2011 estimated around 121 languages and over a thousand dialects.

Problem Statement

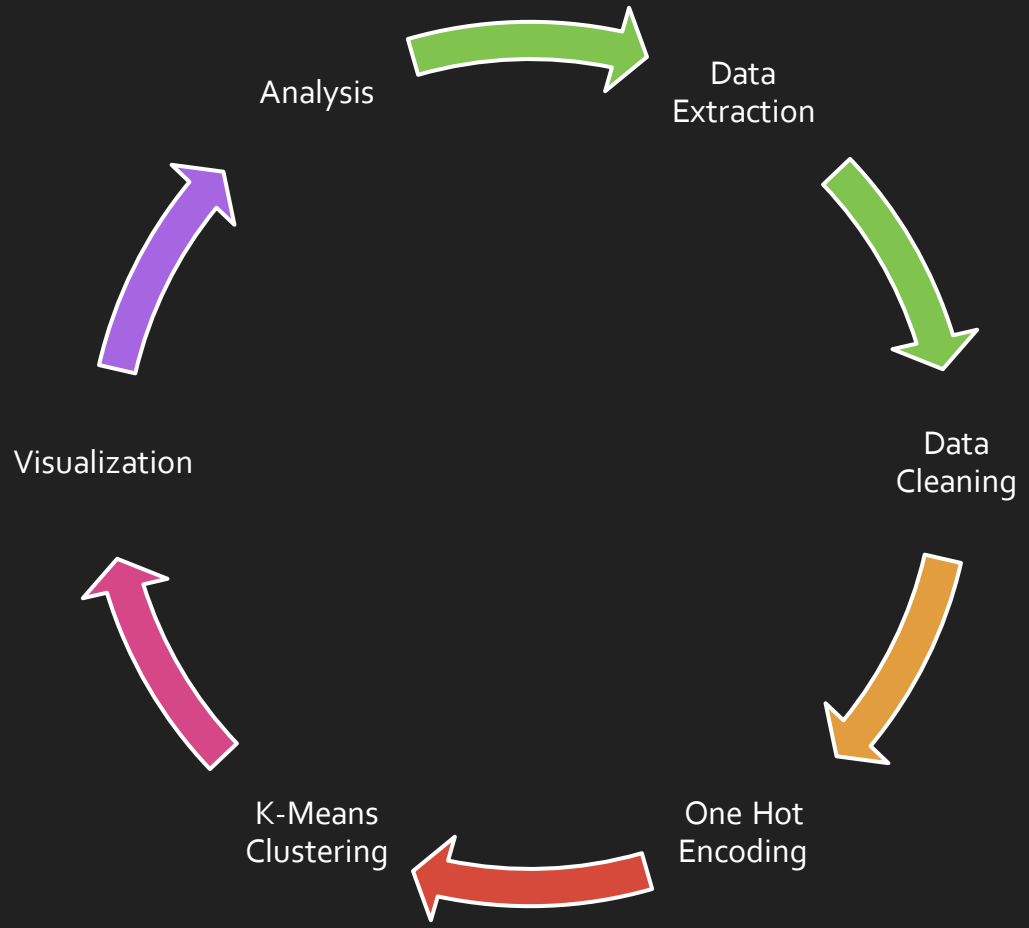
The goal of the project is to compare the languages spoken in each state of India and then form groups of clusters based on how similar each state is regarding the languages spoken by the people of the state.

Project Dataset

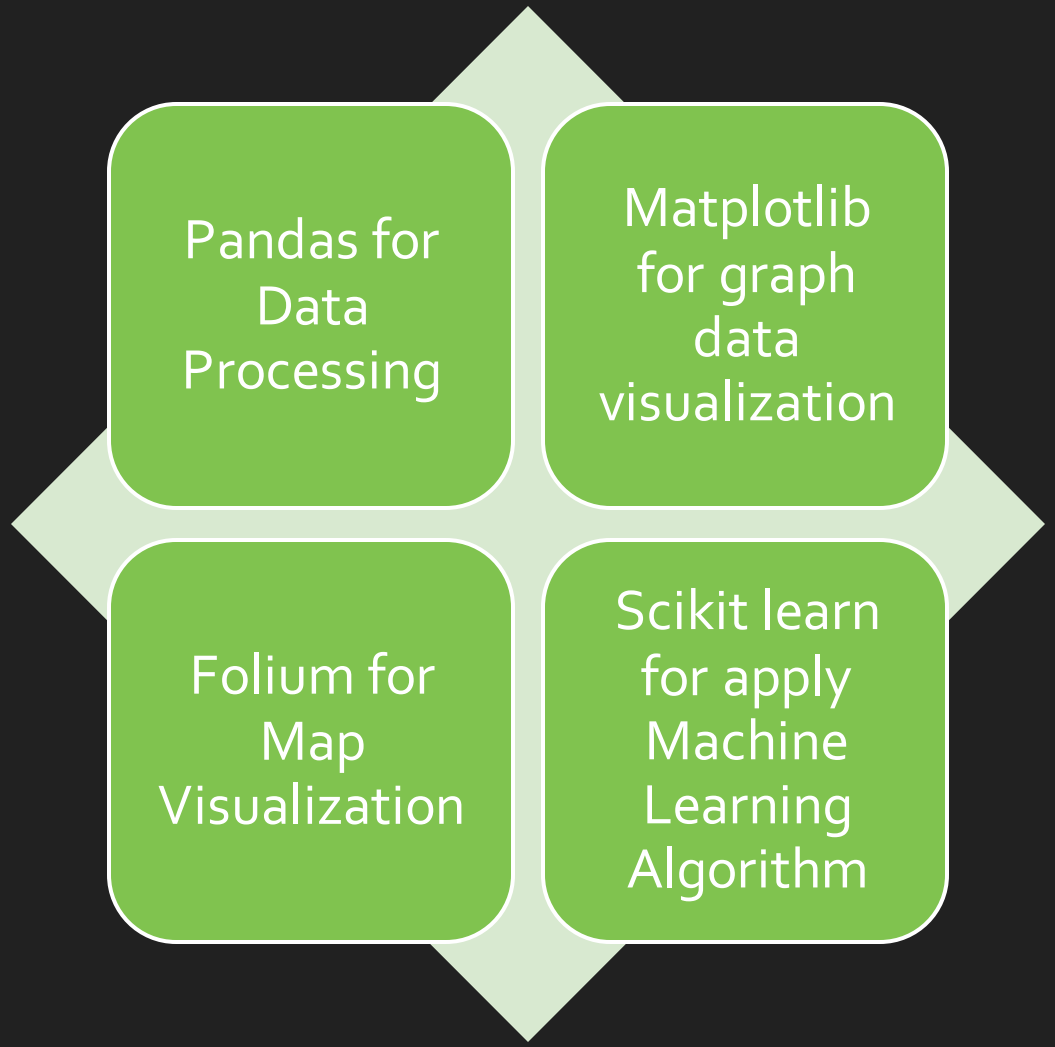
- A dataset was taken from wikipedia containing each state and union territory of India and 5 languages spoken in that state.

	State / Union Territory	Language 1	Language 2	Language 3	Language 4	Language 5
0	Andaman & Nicobar Islands	Bengali	Hindi	Tamil	Telugu	Malayalam
1	Andhra Pradesh	Telugu	Urdu	Hindi	Tamil	Marathi
2	Arunachal Pradesh	Nishi	Adi	Bengali	Nepali	Hindi
3	Assam	Assamese	Hindi	Bodo	Nepali	NaN
4	Bihar	Hindi (Bhojpuri and Magahi)	Maithili	Urdu	Bengali	Santali
5	Chandigarh	Hindi	Punjabi	Urdu	Nepali	Bengali
6	Chhattisgarh	Hindi (Chhattisgarhi)	Odia	Bengali	Telugu	Marathi
7	Dadra & Nagar Haveli	Hindi	Gujarati	Marathi	Konkani	Odia
8	Daman & Diu	Gujarati	Hindi	Marathi	Bengali	Odia
9	Delhi	Hindi	Punjabi	Urdu	Bengali	Maithili

Project Methodology



Libraries Used



Clustering

Clustering is one of the most common exploratory data analysis technique.

It is used to get an intuition about the structure of the data.

It is the task of identifying subgroups in the data such that data points in the same cluster are very similar while data points in different clusters are very different.

It is an unsupervised learning algorithm.

K-Means

Kmeans algorithm is an iterative algorithm that tries to partition the dataset into *K* pre-defined distinct non-overlapping subgroups.

Each data point belongs to **only one group**.

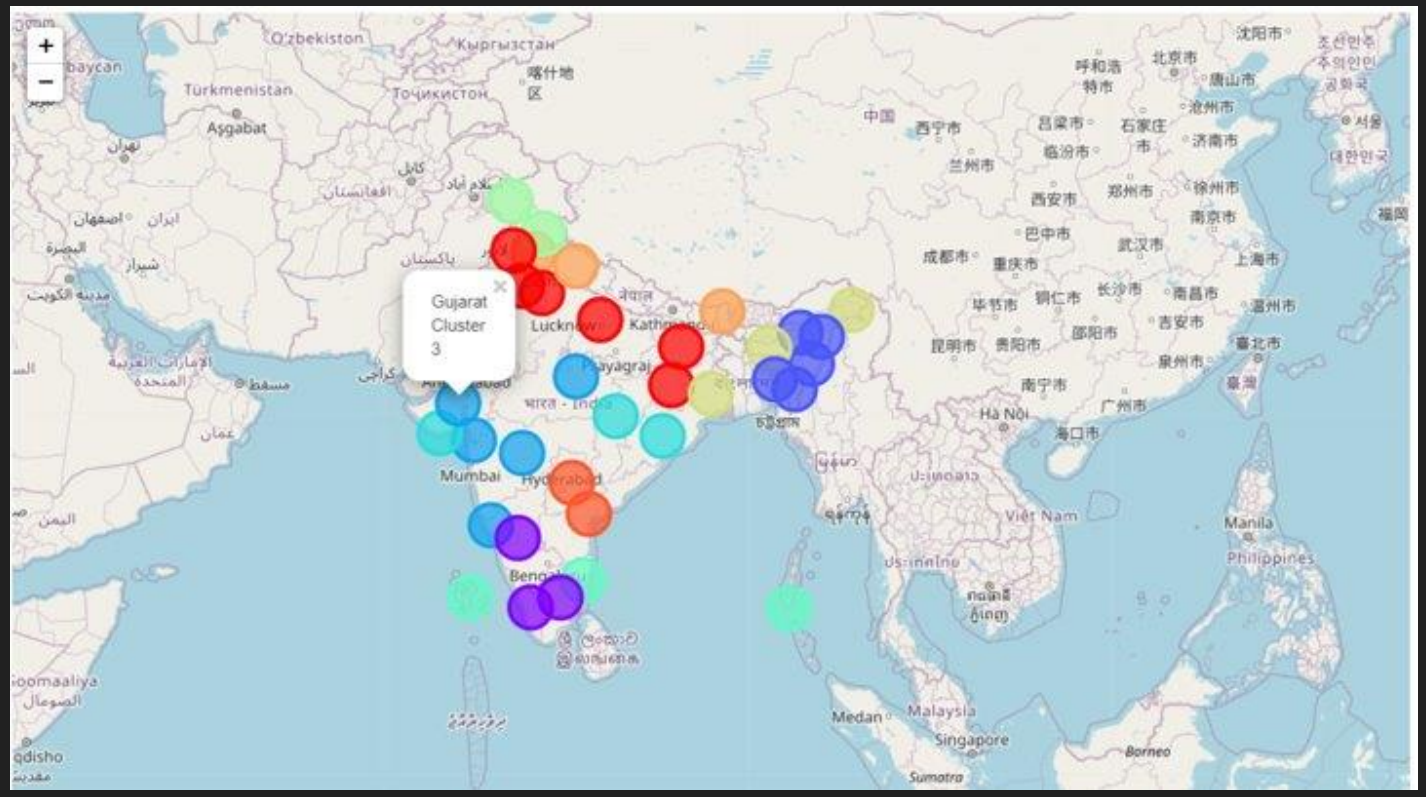
For the purposes of the project we have set K=10

Clustering Results

Each state has been assigned a colour.

Each cluster has a different colour.

Clicking on a circle will reveal the state name and the cluster it belongs to.



References

- https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a
- https://en.wikipedia.org/wiki/Indian_States_by_most_popular_languages
- https://en.wikipedia.org/wiki/K-means_clustering

Thank you