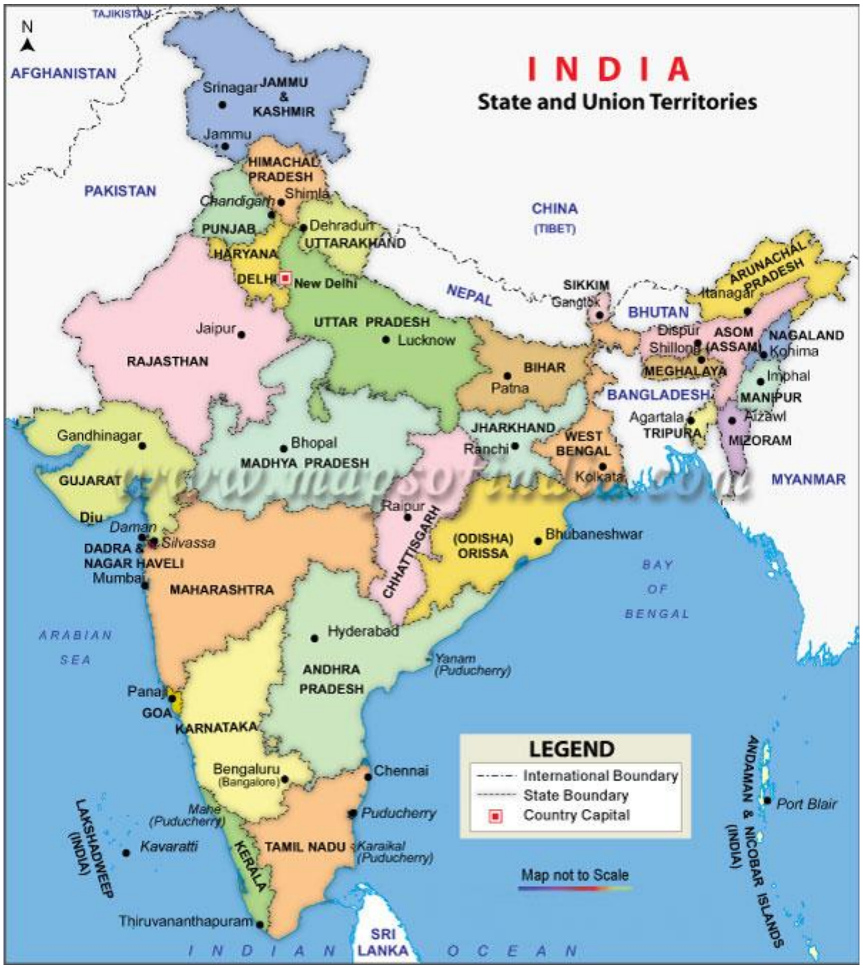
India - Language Diversity

1. Background

With a population of over 1.3 billion, it is a well known fact that India is a diverse country. It is often said that India is a nation with a vibrant history and boasts of a rich culture or rather India's culture is actually an amalgamation of different cultures co-exisiting in harmony. From the Himalayas of the North to the Deccan Plateau in the South, the country comprises of people from all walks of life with unique cultures, values and of course, different languages.

The Constitution of India specifies 22 official languages in the Eighth Schedule although an estimate according to the Census of 2011 resulted in there being about 121 languges, not counting the many various dialects.



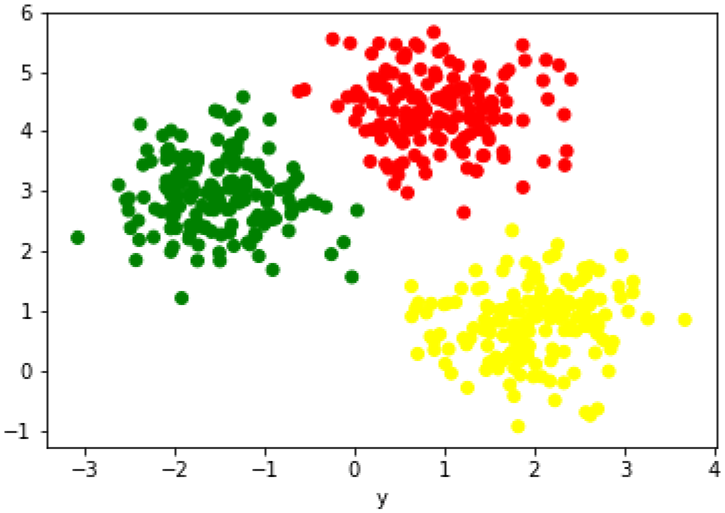
2. Introduction

Languages of India is a very fascinating topic to study. It is indeed very intersting to explore how the Indian languages are spoken throughout the country. Many states in India often have a main local language for example: The state of Gujrat has Gujrati as its most popular language.

Although a lot of states in Northern India have many Hindi speaking natives. One can even say that Hindi is to Northern India what English is to the world.

For this project I have decided to study how the languages of India are distributed and which states are similar linguistically speaking.

Each state has a population that speaks a dozen different languages. My idea is to collect the language data for each state and perform clustering on it to analyse which states are linguistically similar. Of course tools such as the Foursquare API, folium maps, and other libraries will be used for visualizing and analysing data but the main essence of the project will lie in clustering the data.



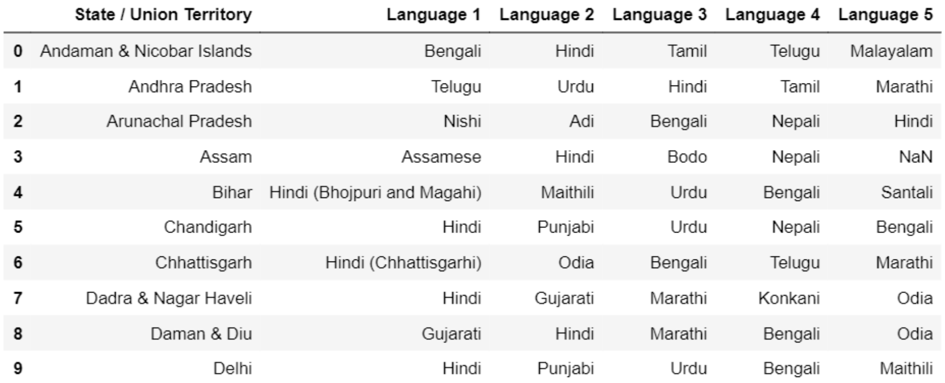
The idea behind the project is to extract data of about 5 or 6 spoken languages in each state and then use K-Means for clustering to assign each state to a cluster based on languages spoken in each state. This proejct, of course can not be limited to the country of India and can be used for studying any country's or even the world's language distrubution to discover any similarity hidden or otherwise that can be visualized on a map.

3. The Dataset

As specified in the problem description I intend to perform clustering based on the languages spoken in each state of India and analyze its results. For this I require reliable data which depicts the languages spoken in India for every state and Union Territory in detail.

After some research I was able to find what I needed on Wikipedia. Here's a link to the wikipedia page: https://en.wikipedia.org/wiki/Indian_States_by_most_popular_languages

Let us proceed to import the table and have a quick look at the dataset to see what we're dealing with.

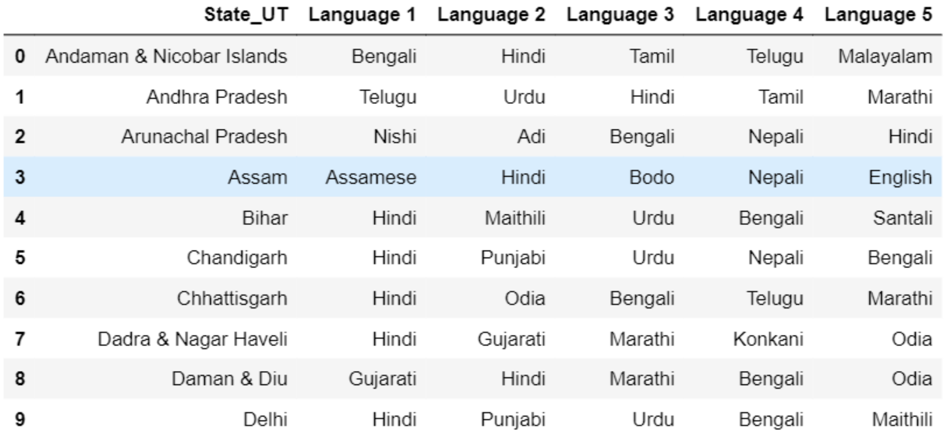


	State / Union Territory	Language 1	Language 2	Language 3	Language 4	Language 5
0	Andaman & Nicobar Islands	Bengali	Hindi	Tamil	Telugu	Malayalam
1	Andhra Pradesh	Telugu	Urdu	Hindi	Tamil	Marathi
2	Arunachal Pradesh	Nishi	Adi	Bengali	Nepali	Hindi
3	Assam	Assamese	Hindi	Bodo	Nepali	NaN
4	Bihar	Hindi (Bhojpuri 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

The table has 6 columns with the first column containing the name of the state or the union territory and the next 5 columns basically tell us which languages are spoken in that particular state. However, in the 'Language 1' column we do notice dialects of Hindi specified as well and that is some information that would not be relevant to the project here. Therefore we will have to perform some data cleaning for that particular column and replace it with only Hindi

We also notice some null values for the column 'Language 5'. Since our data consists of categorical data we cannot replace the NaN values with the mean of the values in that particular column. Therefore, it is safe to replace the absent data with 'English' and it is after all a common language most people of India know and speak, perhaps after Hindi.

Table after cleaning the data looks like:

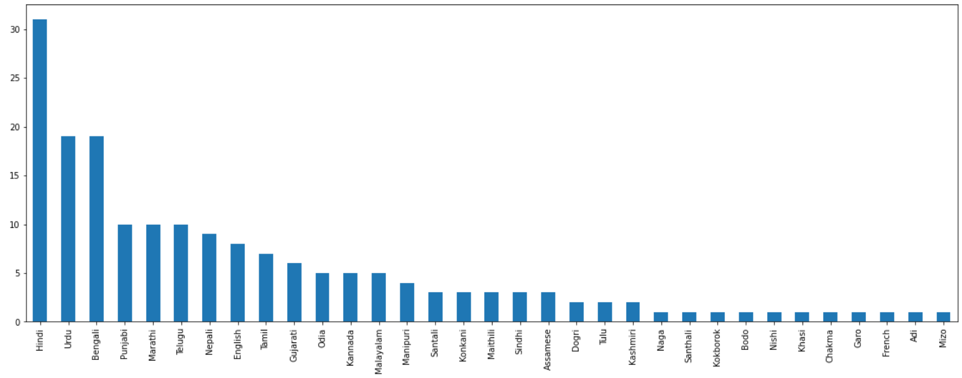


	State_UT	Language 1	Language 2	Language 3	Language 4	Language 5
0	Andaman & Nicobar Islands	Bengali	Hindi	Tamil	Telugu	Malayalam
1	Andhra Pradesh	Telugu	Urdu	Hindi	Tamil	Marathi
2	Arunachal Pradesh	Nishi	Adi	Bengali	Nepali	Hindi
3	Assam	Assamese	Hindi	Bodo	Nepali	English
4	Bihar	Hindi	Maithili	Urdu	Bengali	Santali
5	Chandigarh	Hindi	Punjabi	Urdu	Nepali	Bengali
6	Chhattisgarh	Hindi	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

4. Methodology

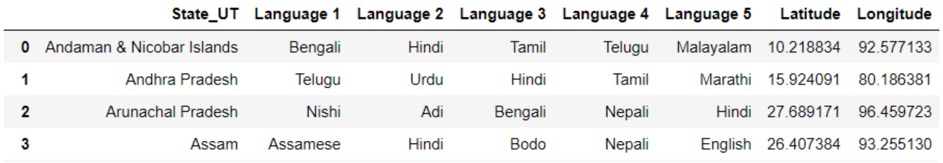
4.1 Data Visualization

In order to understand the data better I visualized a bar chart.



We can clearly see that Hindi is the most widely spoken language followed by Urdu and Bengali. We will now use this data for clustering and further analysis for our project.

Furthermore for the project I will be using Folium to visualize maps. In order to do that I need to make use of location data. Therefore using geopy I added the columns of latitude and longitude to the dataset. Here's a peek at the dataset.

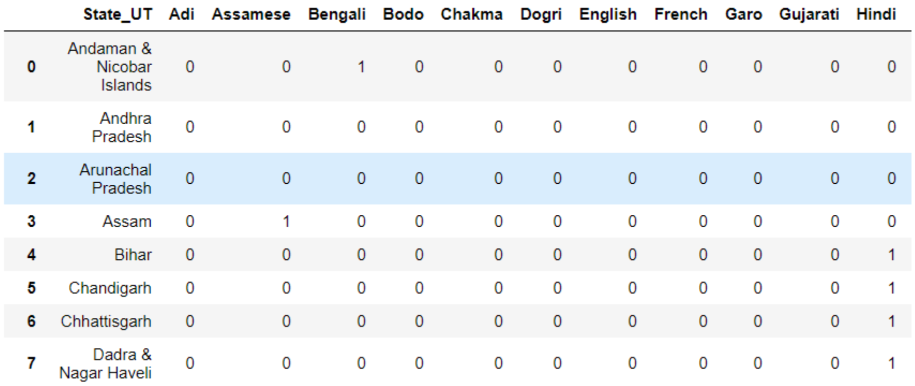


	State_UT	Language 1	Language 2	Language 3	Language 4	Language 5	Latitude	Longitude
0	Andaman & Nicobar Islands	Bengali	Hindi	Tamil	Telugu	Malayalam	10.218834	92.577133
1	Andhra Pradesh	Telugu	Urdu	Hindi	Tamil	Marathi	15.924091	80.186381
2	Arunachal Pradesh	Nishi	Adi	Bengali	Nepali	Hindi	27.689171	96.459723
3	Assam	Assamese	Hindi	Bodo	Nepali	English	26.407384	93.255130

4.1.2 One Hot Encoding

Many machine learning algorithms cannot operate on label data directly. They require all input variables and output variables to be numeric. As we can clearly see in this dataset, we are dealing with categorical features which are non numeric.

Therefore before we proceed any further and apply a machine learning algorithm we need to convert all categorical variables to binary and one way to do it is by One Hot Encoding. One Hot Encoding basically involves treating each categorical variable as a column and return a single value either 0 or 1 if that feature is present for a particular item. We can see a snapshot of the encoded dataframe below.



	State_UT	Adi	Assamese	Bengali	Bodo	Chakma	Dogri	English	French	Garo	Gujarati	Hindi
0	Andaman & Nicobar Islands	0	0	1	0	0	0	0	0	0	0	0
1	Andhra Pradesh	0	0	0	0	0	0	0	0	0	0	0
2	Arunachal Pradesh	0	0	0	0	0	0	0	0	0	0	0
3	Assam	0	1	0	0	0	0	0	0	0	0	0
4	Bihar	0	0	0	0	0	0	0	0	0	0	1
5	Chandigarh	0	0	0	0	0	0	0	0	0	0	1
6	Chhattisgarh	0	0	0	0	0	0	0	0	0	0	1
7	Dadra & Nagar Haveli	0	0	0	0	0	0	0	0	0	0	1

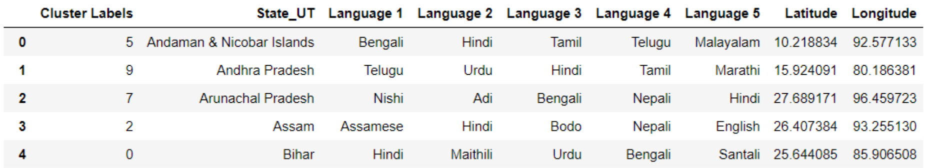
From the above dataframe we can clearly observe that Hindi is spoken in Bihar, Chandigarh, Chattisgarh and Dadra & Nagar Haveli as those rows show a 1 marked for the column labeled 'Hindi'

4.1.3 K-Means Clustering

Clustering can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different. Clustering is generally regarded as an unsupervised learning algorithm because here we are not training the model. We are just sending an input and the algorithm processes it and returns an output.

K-Means algorithm is an iterative algorithm that tries to partition the dataset into 'K' pre-defined distinct non-overlapping cluster where each data point belongs to only one group. Here, we need to specify 'K' or the number of clusters we want to divide our data into. For the purposes of this project we will take **K=10**

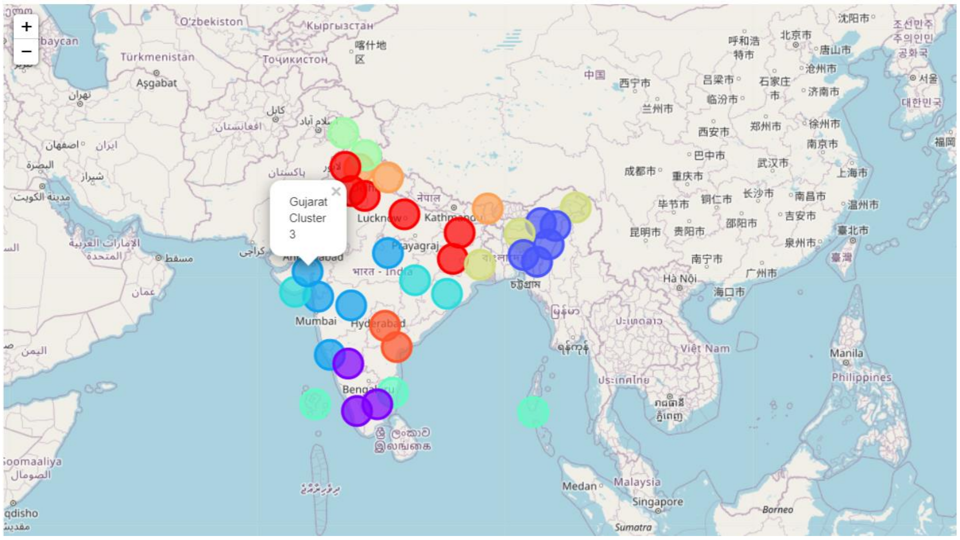
Once the K-Means algorithm has generated the clusters we then add the cluster labels to our dataframe so that we can move on to our final cluster visualization. Here is how the dataset looks like with the cluster labels:



	Cluster Labels	State_UT	Language 1	Language 2	Language 3	Language 4	Language 5	Latitude	Longitude
0	5	Andaman & Nicobar Islands	Bengali	Hindi	Tamil	Telugu	Malayalam	10.218834	92.577133
1	9	Andhra Pradesh	Telugu	Urdu	Hindi	Tamil	Marathi	15.924091	80.186381
2	7	Arunachal Pradesh	Nishi	Adi	Bengali	Nepali	Hindi	27.689171	96.459723
3	2	Assam	Assamese	Hindi	Bodo	Nepali	English	26.407384	93.255130
4	0	Bihar	Hindi	Maithili	Urdu	Bengali	Santali	25.644085	85.906508

4.1.4 Cluster Map Visualization

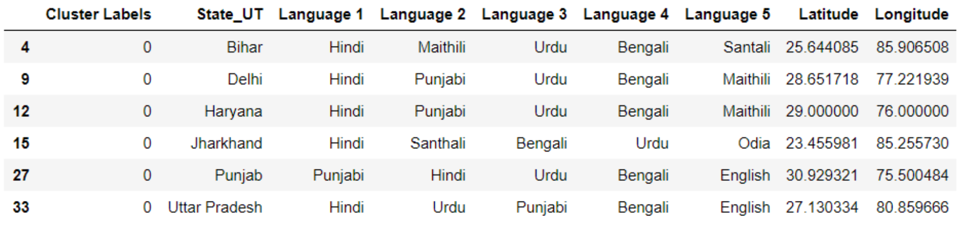
Folium library was used to visualize the clusters on the map generated below:



5. Results

As we can observe, states with similar spoken languages appear as clusters of 1 colour. On clikcing on any circle we can see the name of the state or the union territory and also the cluster group it belongs to. It is common knowledge that language distribution has a large impact on geography i.e. regions close to each other are known to speak the same or similar languages. Therefore we can conclude that clustering has been carried our correctly. In or der to verify this we can explicitly study each cluster. Let us proceed to study a couple of clusters to see our results properly.

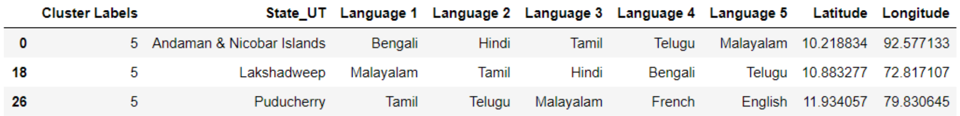
For cluster label 0



	Cluster Labels	State_UT	Language 1	Language 2	Language 3	Language 4	Language 5	Latitude	Longitude
4	0	Bihar	Hindi	Maithili	Urdu	Bengali	Santali	25.644085	85.906508
9	0	Delhi	Hindi	Punjabi	Urdu	Bengali	Maithili	28.651718	77.221939
12	0	Haryana	Hindi	Punjabi	Urdu	Bengali	Maithili	29.000000	76.000000
15	0	Jharkhand	Hindi	Santhali	Bengali	Urdu	Odia	23.455981	85.255730
27	0	Punjab	Punjabi	Hindi	Urdu	Bengali	English	30.929321	75.500484
33	0	Uttar Pradesh	Hindi	Urdu	Punjabi	Bengali	English	27.130334	80.859666

As we can see for the cluster label 0, the common languages are Hindi, Urdu, Bengali and Punjabi. These Clusters are mostly towards the north of India.

For cluster label 5



	Cluster Labels	State_UT	Language 1	Language 2	Language 3	Language 4	Language 5	Latitude	Longitude
0	5	Andaman & Nicobar Islands	Bengali	Hindi	Tamil	Telugu	Malayalam	10.218834	92.577133
18	5	Lakshadweep	Malayalam	Tamil	Hindi	Bengali	Telugu	10.883277	72.817107
26	5	Puducherry	Tamil	Telugu	Malayalam	French	English	11.934057	79.830645

As we can see for the cluster label 5, the common languages are Tamil, Telugu and Malayalam. These Clusters are situated towards the south of India.

5. Discussion

Based on the results we achieved k-means clustering with good accuracy. However, this project was carried out for K=10, i.e. 10 clusters were formed. Varying the number of clusters will greatly change our results. One way to improve the results further could be to incorporate the Elbow method which calculates an optimal cluster number K for a given set of observations.

Since India has a total of 36 states and union territories the dataset was highly coontrolled. In order to fit this algorithm for a country with more number states like the USA or less like Nepal, we will have to adjust the number of clusters, K, accordingly in order to obtain accurate results.

6. Conclusion

In conclusion, clustering using K-Means was carried out successfully. Pandas was the main library used to data cleaning and storing. Matplotlib and Folium were used in the project for visualization purposes. We used a clustering algorithm of K-Means by setting k=10, however other clutering approaches can also be applied. The results obtained were analysed by examining clusters.