Visualizing topics in Indian Parliamentary Discussions

Salil Harsulkar| Indiana University| samihars@iu.edu| GitHub

Project abstract:

India is the largest democracy in the world and has a bicameral legislature consisting of an upper house called 'Rajya Sabha' which represent the states of the Indian federation and the lower house called 'Lok Sabha' which represent the people of India as whole. [1] These two houses are the heart of Indian legislature and all decisions pertaining to the present and future of India are made here. During its regular working sessions, each house discusses and debates on a variety of topics – some trivial while some are of critical national importance. At the start of each session, there is a questions hour where members can ask questions to cabinet ministers on topics related to national interest. Analyzing these questions and their answers can help extract key themes that shaped the country over the past few years.

Project Aim:

The main aim of this project is to perform exploratory data visualization on the vast archives of parliamentary publications publicly available on the Rajya Sabha website with the intent to uncover interesting insights about the composition, transformation and functioning of the Upper House. This project also aims at performing topical analysis of questions asked on the floor of the House and highlight any trends in discussion topics by State and by Ministry over the past decade.

Motivation:

A democratic country like India can only be successful if its people actively participate in the politics and decision-making process for their country. For the majority population, the news media is the primary source of political information. Though news media is effective in its reach, it may not be free from bias.

Rajya Sabha publishes records of its debates, discussions and questions on its website [2]. There is currently no platform available in India that can source data from the parliamentary publications, transform it and present it to the people in a way that is easy to comprehend.

The other constraint is that of language. India is a multi-lingual country with 22 official languages and close to 150 languages that have a sizable population. The legislative verbiage is mainly published in English and Hindi and is too complicated for a layman to understand. Crisp visualizations will help in making the archives more accessible to the general public.

With the power of machine learning, natural language processing and effective data visualization, this project will attempt to provide a solution to these problems.

Analysis of existing visualizations:

I will be referencing the work done by my fellow students for the visualizations [4] created for IVMOOC course as part of MS in Data Science program of Indiana University, Bloomington in spring 2018. They had similar objective of visualizing government meetings data from the city council for Bloomington, Indiana. Their project report has some good visualizations that I plan to replicate in this project – for instance the use of Sankey graph to plot the flow of applications through the legislative process can be reused. They have also used network visualization to plot interdependencies between words used in the applications. It may be applicable for this project as well.

On the downside, the dataset that was used for their project is considerably smaller and has less variance in features than what this project has.

Data Introduction:

Data Dimensions:

Data has been collected primarily from the Rajya Sabha archives [2]. Data used for this project can be categorized into below dimensions –

Dimension	Data Timeline	Features	Count of records
Members of Parliament	1952 - 2018	Name, Tenure, Political Party, State represented	3261 member records
Rajya Sabha Sessions	1999 - 2018	Session Number, Start and End dates	60 session records
Question and Answers	2009 - 2018	Question title, Question description, Answer, Question by, Answer date, Ministry question was asked to	89093 questions and answers

Data Collection:

Dimensions - Members of Parliament, Rajya Sabha Sessions:

The data supporting these two dimensions was scraped from the Rajya Sabha website into Comma Separated Value (CSV) formatted files.

Dimension – Question and Answers:

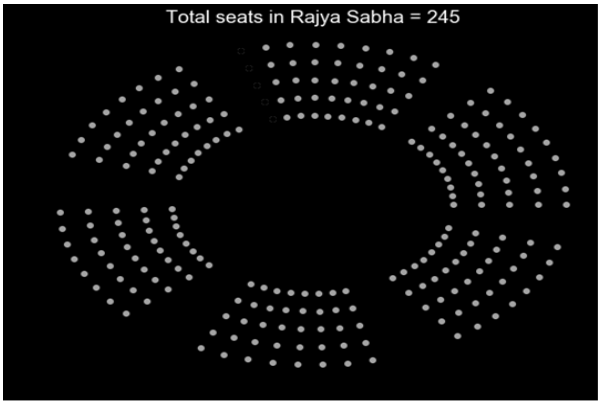
A dataset of all questions from 2009 till 2018 is available on Kaggle here [3]. This dataset has been web scraped from the website above using R code to generate nine year-wise CSV files.

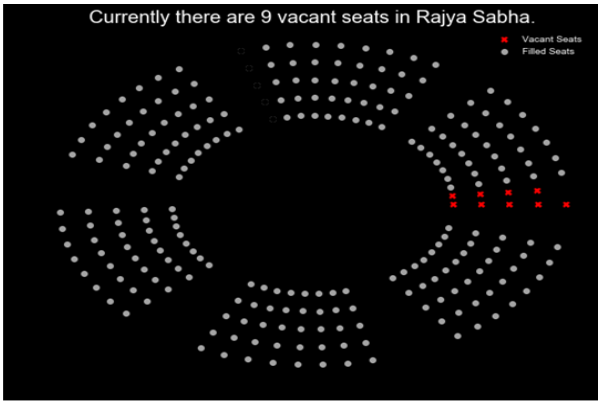
Visualizations and Insights:

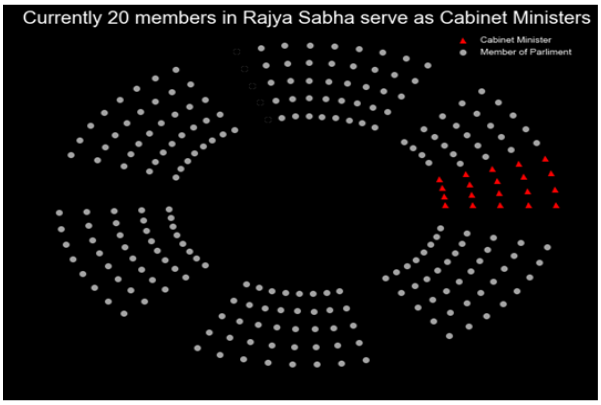
1. Visualizing the current (2018) composition of Rajya Sabha

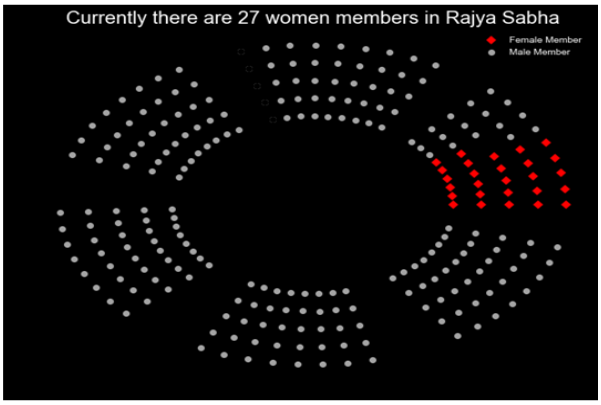
Description:

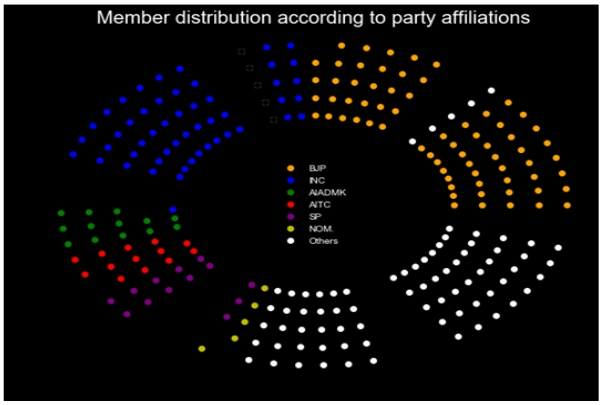
Rajya Sabha, also called the Council of States, has 245 seats for representative of the States and Union Territories of Republic of India. This custom designed visualization shows the composition of the 2018 Rajya Sabha and the distribution of seats along the dimensions of gender, special responsibilities and political affiliations. Each point represents a seat.











Why this visualization type:

Bar chart and donut chart were considered before finalizing on this visualization type. Both bar and donut charts looked cluttered and mundane and were unable to present the idea effectively. This custom scatter plot was designed to mimic the arrangement of a circular assembly with each point in the plot representing a member. Color encoding was used to highlight different topics. The size and position of the points was maintained so that all five plots can be animated as a GIF.

How was this created:

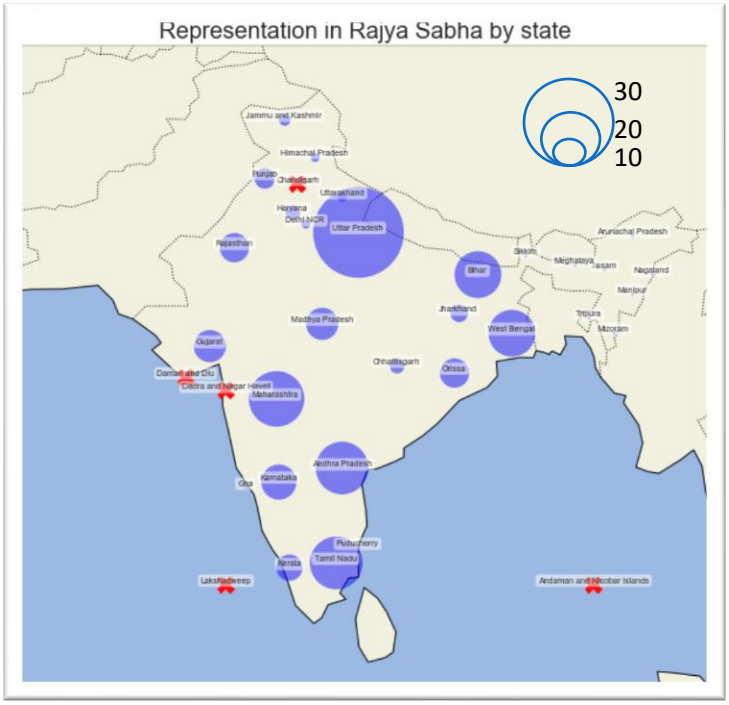
Data Source:

The data for this visualization has been scraped from the Rajya Sabha website [2] from the section 'Sitting members'.

Visualization tools used:

This is plotted entirely using Python's Matplotlib library. Each plot contains 245 points plotted using polar coordinates such that they form 5 concentric circles of increasing radii. The angles are managed by creating a sequence of 50 equidistant points between (0, 360). Please refer the accompanying Jupyter notebook on GitHub for code.

2. Visualizing number of seats by State



Description:

The Indian Constitution has allotted fixed number of seats per state and union territory in proportion to the population residing in the state.
Only the Union Territories of Pondicherry and NCR Delhi are allocated membership, while the other five are not (visualized by red 'X' in the map).

Why this visualization type:

Proportional symbol map was chosen for this Visualization because it accurately encodes the number of seats in relation to the physical location of the state. In addition to showing location and magnitude, the visualization is also able to clearly show which Union Territories do not have representation. Even though the map only outlines India's international border, most of the audience can recognize states by their location.

How was this created:

Data Source:

The data for this visualization is sourced from the 2011 Census data available here [8].

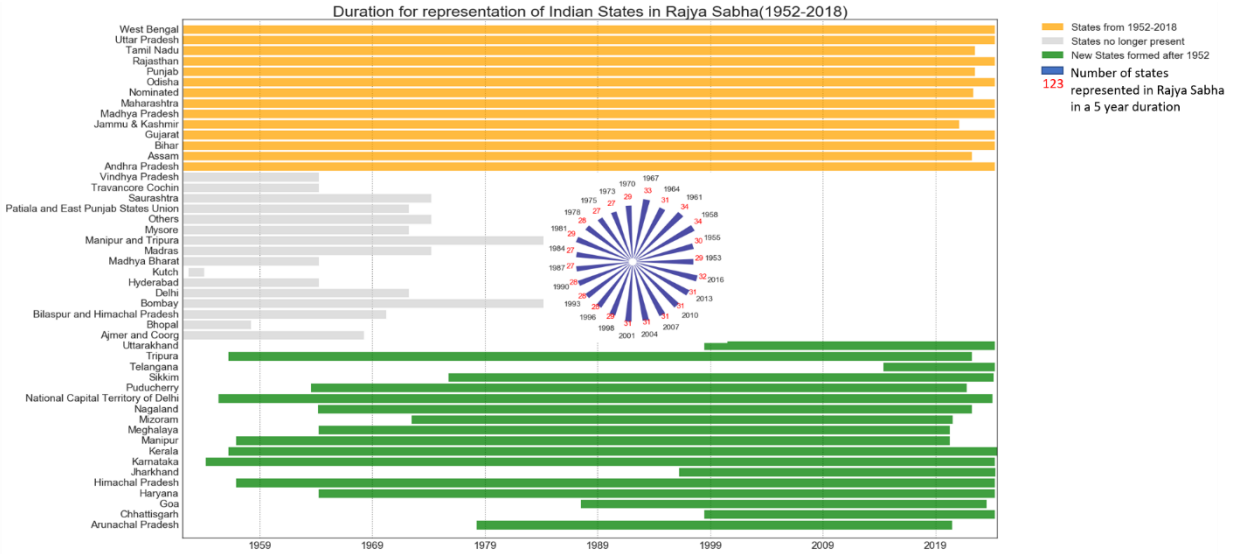
Visualization tools used:

This proportional symbol map is plotted in Python using Cartopy library. This map was focused on India by passing a range of Latitude and Longitude values. It also allows users to enable geographic features like land, ocean, coastline and borders. The proportional symbols are plotted using the latitudes and longitudes of the state capitals and are proportional to the number of seats allocated in Rajya Sabha.

3. Visualizing India's Geographical Evolution (1952-2018)

Description:

This visualization encoded to look like the Indian National Flag plots the representation of States and Union Territories in Rajya Sabha on the timeline from 1952 – 2018. The states encoded in orange color have a continuous representation across the entire timescale. The states in gray color no longer exist while the ones in green color were formed after 1952. The circular bar chart in the center shows the count of states represented in a 3-year time slots between 1952 – 2013.



The Rajya Sabha convened for its first session on May 13th 1952, five years after India's independence in 1947. India's journey post independence has been one of inclusion and consolidation. We see two distinct phases here – pre-1970 shows many provinces and princely states that existed in pre-independence India. The States Reorganization Act passed in 1956 and implemented over the next few decades saw a period of consolidation resulting in the birth of many new states and Union Territories that constitute today's India. This process is still in progress as can be seen by formation of Telangana – India's 29th state created in 2014.

Insights:

- When India became independent in 1947, it was divided into 500+ princely states and provinces. Rajya Sabha started operating from 1952 and has representation from 29 states and provinces. The States Reorganization Act of 1956 helped restructure and consolidate the states boundaries giving rise to new states and union territories. Post

1970, the states that exist today started to form slowly. The newest state of Telangana was formed in 2014 which can be seen as difference between 2013 and 2016.

- The circular bar chart shows a concise picture of the geographical evolution – the period of 1952 – 1970 sees a general increase in the number of represented states from 29 in 1953 to 34 in 1958 which then decreases to 26 in 1969. 1970 onwards, the number of states has increased steadily from 27 in 1975 to 32 in 2016.

Why this visualization type:

The Gantt chart is the most suitable chart to plot time series. This visualization plots the timelines of representation of states in Rajya Sabha with start and end years. This chart also enabled me to customize it according to the underlying context of India.

How was this created:

Data Source:

The data for this visualization has been scraped from the Rajya Sabha website [2] from the section 'Members->Former members'. Data had to be manually scraped out into a CSV file.

Data Processing:

- The source data has one record per individual member. To build a Gantt chart [10] by State, the minimum date and maximum dates per State were aggregated and inserted into a new dataframe. This data frame was then color coded to highlight three categories of data. Then the dataframe was sorted based on the color coding.
- To create the circular bar chart [11], polar coordinates were used.

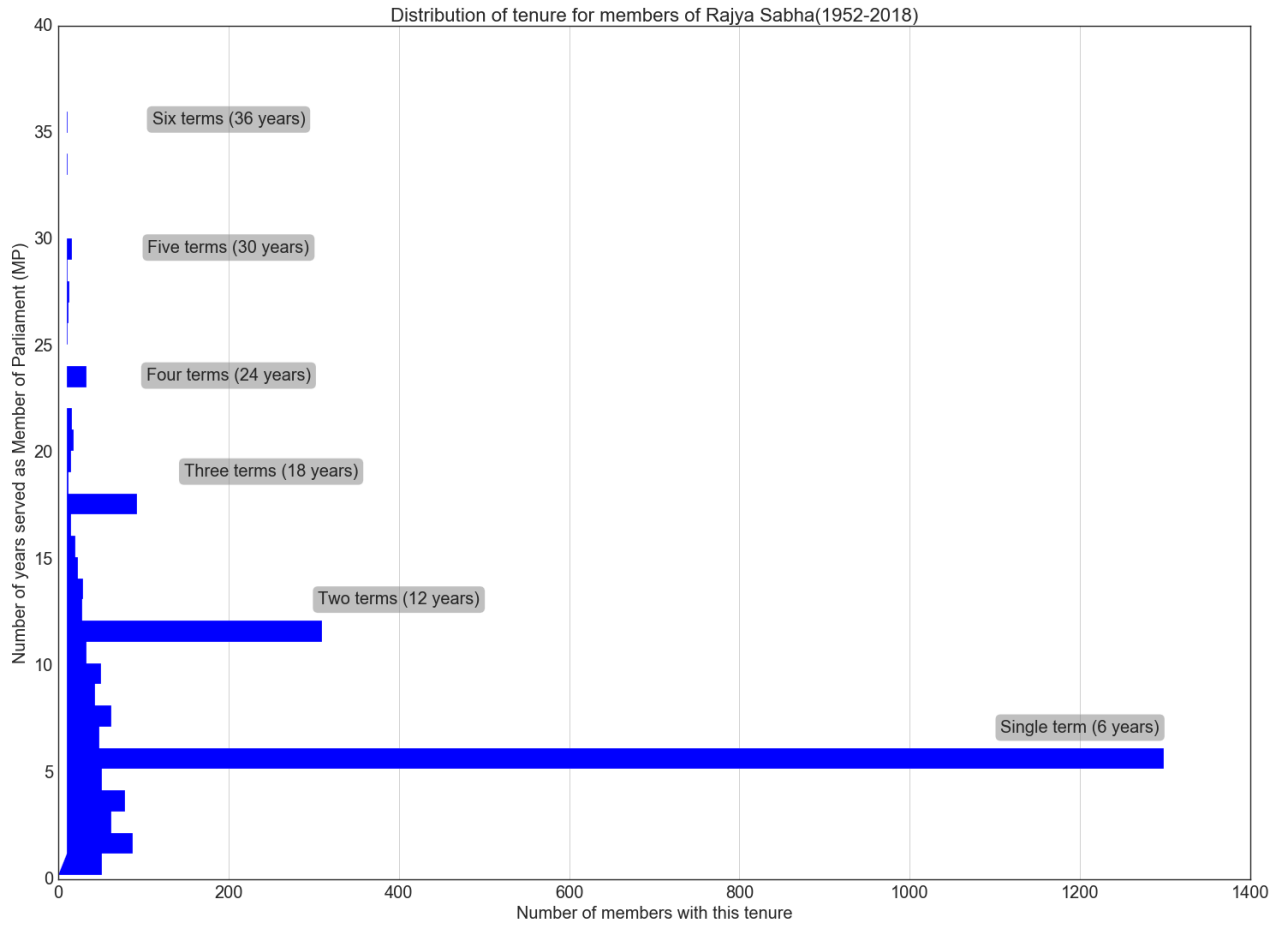
Visualization tools used:

This Gantt chart is plotted using Python's Matplotlib library. The Gantt chart and circular bar graph was later combined in Microsoft PowerPoint to create the final visualization.

4. Visualizing Distribution of Tenure (1952-2018)

Description:

This visualization plots the distribution of tenure of all members of Rajya Sabha highlighting the values for six terms.



Insights:

- The Indian Constitution has limited the number of years a member can serve on Rajya Sabha in one term to 6 years. There is no restriction on the number of times a member can be reelected to the House. This plot gives good insights on distribution and magnitude of tenures for all its 3000+ members since inception. Majority of the members have served just 1 term and did not get reelected. It is interesting to see a handful of members having served more than five terms which equate to 30+ years as Member of Parliament. It may be worth finding more details about their contributions and success stories. Member profile of one such member is visualized in here.

Why this visualization type:

The intent of this visualization was to plot the distribution of member tenure. The most obvious choice to use is the histogram. This plot is a histogram plotted horizontally with bin size equal to 36. The horizontal orientation was more visually effective in combination to the tags than vertical orientation.

How was this created:

Data Source:

The data for this visualization has been scraped from the Rajya Sabha website [2] from the section 'Members->Former members'. Data had to be manually scraped out into a CSV file.
The Term Start and Term End dates were used for this visualization.

Data Processing:

- The term start and end dates for each member were in DD/MM/YYYY format and had to be reformatted into YYYY-MM-DD format.
- The difference in years was then computed between the Term start and end dates.

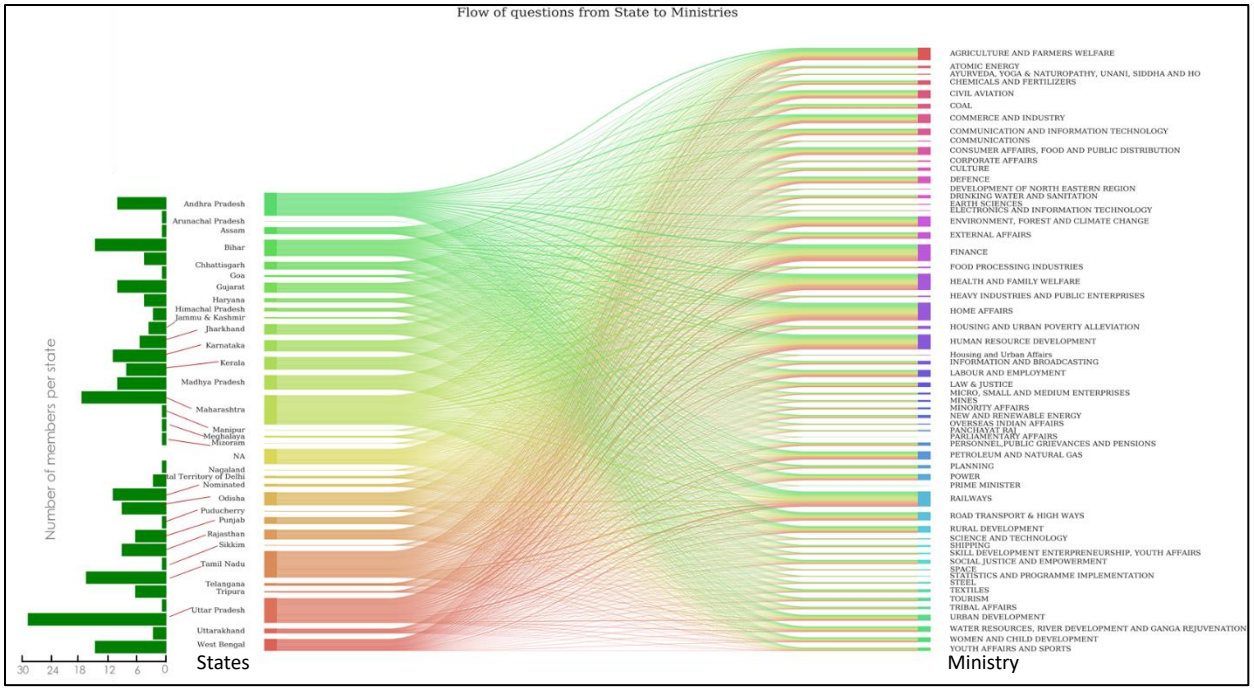
Visualization tools used:

This horizontal histogram plots the count of members vs the tenure in years. The bin size is set to the maximum tenure from the data such that each tick corresponds to one year on the Y-axis.

5. Visualizing flow of questions between State and Ministries

Description:

This visualization shows the flow of all the 89,000 questions from State to Ministry that were asked in this decade. It is complemented by a bar chart that plots the proportion of members representing each state and offers a good visual comparison to the proportion of question volume generated by that state.



Insights:

- The visualization confirms the logical hypothesis that the volume of questions is roughly proportional to the number of members representing a state. Some exceptions include

the state of Uttar Pradesh which has 31 members but has approximately same question volume as the next 2 states with highest representation namely – Maharashtra (19) and Tamil Nadu (18).

- The data has members whose states are unavailable and are represented under the state name 'NA'.
- The top five ministries that received questions are the Ministries of Finance, Railways, Home Affairs, Human Resource Development and Health and Family Welfare.
  The Ministries of Defense, Space, Atomic Energy have received less than expected questions.

Why this visualization type:

Sankey chart was chosen for this visualization because it perfectly encodes the magnitude and direction of flow from source to target. In addition to this, it is visually pleasing and does not overwhelm the audience with the scale of underlying data. This visualization plots the flow of 89K questions from 32 states to over 50 ministries. For such a scale, the chart type does a very good job. To complement the insights provided by Sankey chart, a simple bar chart showing the number of representatives allocated per state is provided for visual correlation between the number of representatives and the question volume.

Enhancements:

- The top five ministries that were the focus of majority of questions in this decade show that the country's focus is still on building essential infrastructure and may indicate the 'Developing' phase of India's journey. How would this change when India becomes a 'Developed' country? Further analysis is needed to find if there is a correlation between the top 5 ministries and the development index of the country.
- Geographically, India can be divided into 6 regions – North, East ,West, South, Central and North East. Plotting the above graph by encoding these geographical regions with color can lead to further insights on regional priorities.

How was this created:

Data Source:

This visualization uses the Q&A data for years 2009 – 2018 that is uploaded on Kaggle [4].
The States information is sourced from the Members dataset.

Data Processing:

- The member names are in the 'Last name Title. First name' format and are standardized to 'Title.First name Last name' format.
- The State information is sourced from the Members dataset by looking up on the standardized name.
- A separate dataframe is created with the state and ministry information for all 89000 questions.

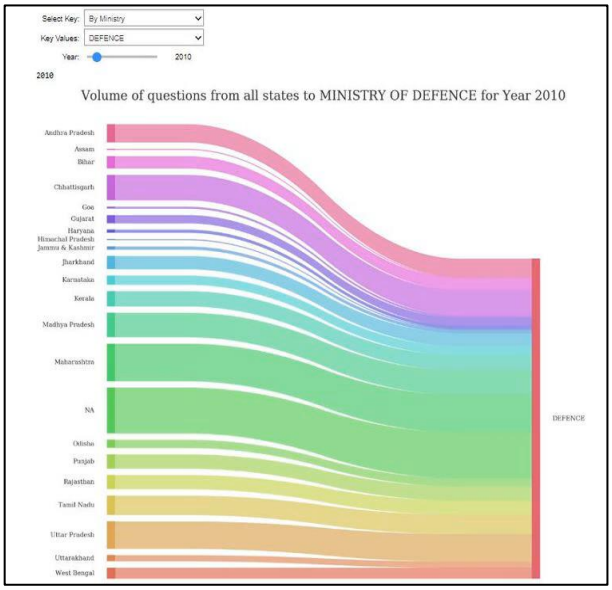
Visualization tools used:

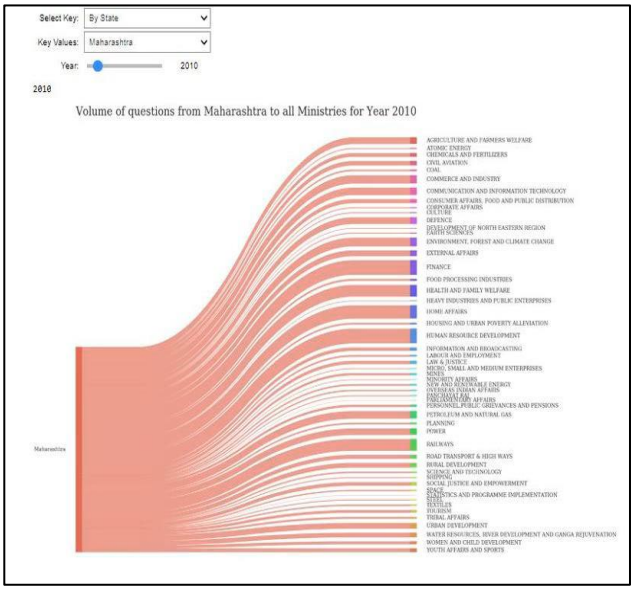
This Sankey chart is plotted using the PySankey library in Python.

6. Interactively visualizing volume of questions by Ministry or State

Description:

The above visualization gives a summarized view of how the questions were distributed by State and Ministry. To dig deeper, this interactive visualization allows users to select one Ministry or State and view how the volume of questions changes by year





Why this visualization type:

Jupyter Notebook provides useful widgets to add interactivity to plots. This visualization was created to enable further analysis on the question volume between states and ministries by allowing users to select specific states and ministries. Additional ability to filter by year was provided as a slider bar.

How was this created:

Data Source:

The same Pandas dataframe used for visualization 5 is used.

Data Processing:

- Year is extracted from the answer_date field using date library.
- To filter on the State/Ministry and Year, Pandas query is used.

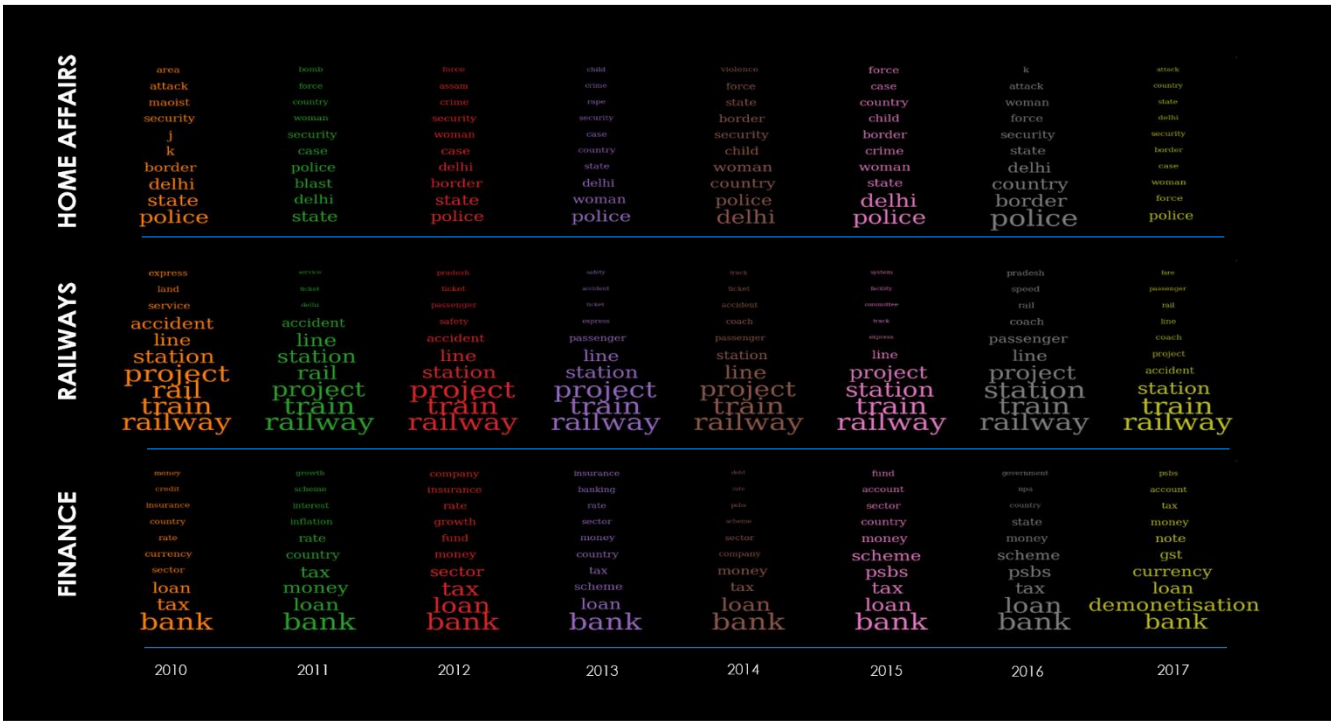
Visualization tools used:

This Sankey chart is plotted using the PySankey library in Python. Interactivity is built using ipywidgets which supports drop down boxes and slider bars to interact with the users.

7. Visualizing top 10 question topics by year

Description:

This custom built stacked word chart plots the word frequency of top 10 question topics for three Ministries from 2010 to 2017. The word size is proportional to the word frequency with the top-ranking words at the bottom and low-ranking words at the top. Python NLTK library is used to perform Natural Language Processing on the questions to extract tag words for question topics.



Insights:

- The top three topics don't change much over the years for any of the ministries. Overall, the topics visualized are reflective of reality and conform to our expectations as seen by the presence of the words '*demonitization*' and '*gst'* for Finance ministry in 2017 as both were hot topics in 2017.

Why this visualization type:

Word cloud was the method of choice for this visualization. However, the outcome was cluttered and did not allow comparing the trend of topics over multiple dimensions. This stacked word chart was custom built using the concepts of bar chart and scatter plot by managing horizontal and vertical offsets between words.

Enhancements:

- The NLP transformation done to extract topics from question titles can be enhanced to extract more meaningful tag words that are more reflective of the topic.

- Sentiment Analysis can be coupled with this chart to plot the sentiment of the question using color encoding.

How was this created:

Data Source:

This visualization uses the Q&A data for years 2009 – 2018 that is uploaded on Kaggle [4].

Data Processing:

This visualization needs the most data processing to extract topics from the question title and calculating word frequencies. Broadly, the data processing follows below steps –

1. *Import NLTK package and define Regex -*

- Import NLTK library for Natural Language Processing, define stop words
- Define the Regex pattern to use to extract the topics. After extensive analysis of the question title patterns, the below Regex was finalized to pull out nouns after Part of Speech tagging.
  NP: {<NN.*|VB.*|CC|JJ.*|RP>*<IN|TO>}
  {<NN.*|VB.*|CC|JJ.*|RP>*<IN|TO>*<NN.*|VB.*|CC|JJ.*|RP>*}

2. *Chunking –*

- Each question title is tokenized and lemmatized.
- The lemma are passed though POS tagger to tag Parts of Speech
- The POS tagged words are chunked to pull out Noun Phrases by applying Regex Parser

3. *Extract Topics –*

- Extract Nouns from the Noun Phrases and insert them into a python dictionary.

4. *Generate word frequencies –*

- Count the number of words in the dictionary to generate a bag of words for tags.

Visualization tools used:

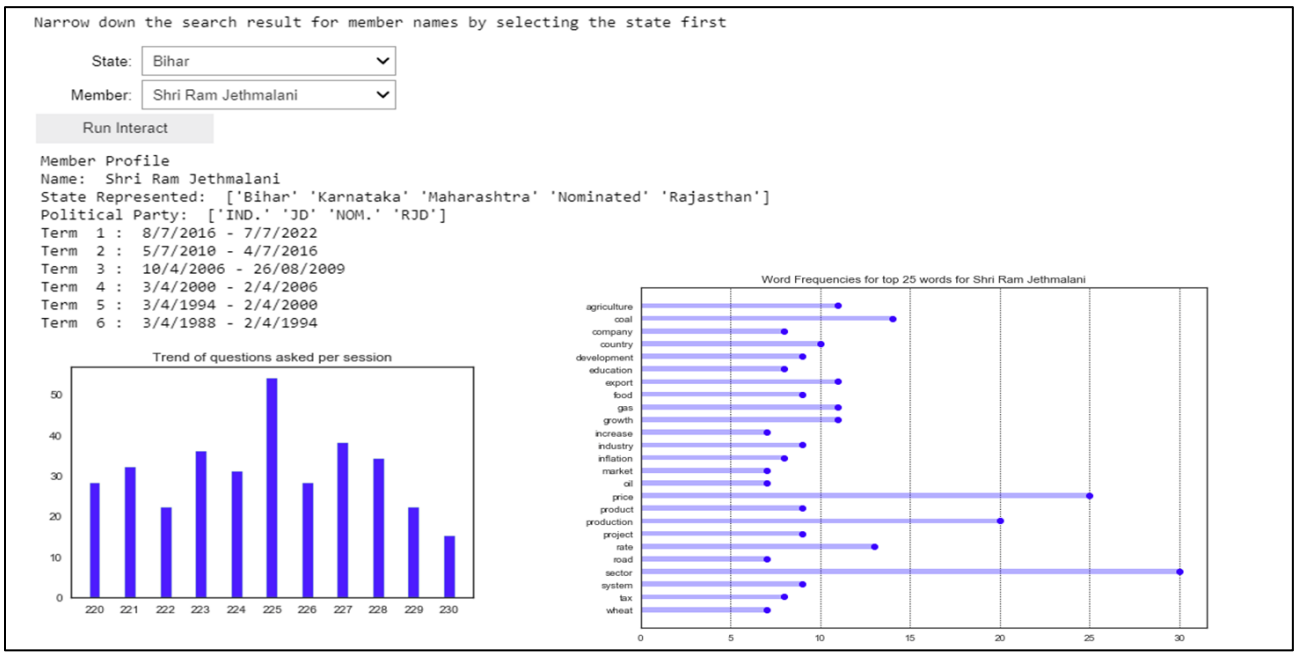
This is a custom-built visualization using Matplotlib. The text is arranged carefully by managing horizontal and vertical offsets such that the top 10 words are arranged according to rank and year.

8. Interactive Member Profile

Description:

This interactive visualization allows users to view demographic information and key performance metrics for any Member of Parliament. The details include Name, States represented, Political affiliations, Tenure and Term dates, Question count per session (2010 onwards) and top 25 topics

for questions asked (2010 onwards). Members are categorized by most recent state represented so selecting the state filter first narrows down the Member list significantly.



Insights:

- Members can be elected from multiple states in different terms over their tenure in Rajya Sabha.
- Members often change political party affiliations to get elected or can be nominated by the President.
- One term in Rajya Sabha is for 6 years and members can serve more than 1 term. This member in the example is the highest tenured member of Rajya Sabha and is serving his 6th term which equates to 36 years as a Member of Parliament.

Enhancements:

- Though this visualization provides good insights, it need enhancement to beautify presentation. Further details about the member like educational background, profile photo, Ministry handled etc. can also be added to enhance this visualization.

How was this created:

Data Source:

This interactive visualization aggregates information from the members and Q&A datasets by members. It uses processed data frames already created for previous visualizations

Data Processing:

- Aggregations for the question counts are done per member.
- The State, Party and term information is source directly from the member data.
- The bag of words is used to identify top 25 question topics asked per member

Visualization tools used:

Interactivity is built using ipywidgets which supports drop down boxes and slider bars to interact with the users.

The member details are printed using print command. The bar charts are plotted using Matplotlib.

Conclusion:

The Rajya Sabha archives present tremendous opportunity for data visualization. The visualization and analysis performed as part of this project generated interesting insights into the functioning, evolution and history of Rajya Sabha. The topical analysis confirmed our assumptions about the correlation between question volume and member representation and proved that the top question topics are reflective of the most important subjects that concern the nation. Exploratory visualizations brought forward interesting stories about the formation of states and tenure of members.

Future Work:

This project has just scratched the surface and will need further refinement in improving the accuracy as well as scale of data. Specifically, future work can be planned in the below areas -

Improving accuracy of Topic Extraction algorithm -

The algorithm used by this project is very basic and needs to be enhanced to capture the question topics accurately. Advanced NLP and NLU libraries can be used to identify entities and relations in the question titles.

Sentiment Analysis and Topic Categorization -

Broadly, the questions can be categorized into three categories - to get more information, to hold minister accountable and to bring ministry's notice to some topic. Being able to identify the sentiment and category of question can lead to new insights.

Scaling up to accommodate entire Q&A archives -

This project only focuses on the question and answers data from 2009-2018. It can be scaled up to source all the available data from 1952 till current.

Interactive visualizations -

Adding more interactive capabilities like ability to get details by hovering over the visualization, filtering and aggregations will help create more engaging visualizations.

References:

[1] Wikipedia - https://en.wikipedia.org/wiki/Politics_of_India

[2] Rajya Sabha official website - https://rajyasabha.nic.in/

[3] Kaggle datasets - https://www.kaggle.com/rajanand/rajyasabha

[4] Jamie Israel et all., Visualizing government meetings (IVMOOC Spring 2018) -

https://iu.app.box.com/s/tlsqff22gscuyniw6byhdtyrulh2rm07

[5] Python interactive automated plots-

https://github.com/bloomberg/bqplot/blob/master/examples/Applications/Wealth%20of%20Nations.ipynb

[6] Jupyter widgets – https://ipywidgets.readthedocs.io/en/latest/examples/Using%20Interact.html

[7] NLTK documentation - https://www.nltk.org/book/ch07.html

[8] India 2011 census - https://www.census2011.co.in/states.php

[9] Factly https://factly.in/parliament-lessons-understanding-question-hour-lok-sabha/

[10] Plotting Gantt Charts in matplotlib http://www.clowersresearch.com/main/gantt-charts-in-matplotlib/

[11] Plotting circular bar charts https://stackoverflow.com/questions/46874689/getting-labels-on-top-of-bar-in-polar-radial-bar-chart-in-matplotlib-python3

Acknowledgements –

I am thankful to the Kaggle user Rajanand Ilangovan for his effort on scrapping this dataset from the Rajya Sabha official website and uploading it on Kaggle, thus making it available for easy use in this project.

I am also thankful to Prof. Y. Y. Ahn for his guidance and motivation.