Process Book

11.09.2018

About

Team Members:

- Archanasri Subramanian , u1141789@utah.edu
- Krithika Iyer , u1135255@utah.edu

Github Repository: https://github.com/archanasris/dataviscourse-pr-languagesoftheworld

Release :https://github.com/archanasris/dataviscourse-pr-languagesoftheworld/tree/1.0.0

Overview and Motivation

We looked at many datasets for visualization like effect of droughts on crops in USA, women in law dataset from World Bank repository etc. We felt that these datasets didn't provide us much scope for visualization. After some more search we found the WORLD ATLAS OF LANGUAGE STRUCTURES dataset which had a variety of features ranging from geographical prevalence to semantic structures which motivated us to come up with interesting and unique ways to visualise them.

People don't speak one universal language, or even a handful. Instead, today our species collectively speaks over 7,000 distinct languages.And these languages are not spread randomly across the planet. Hence we thought it would interesting to study the different properties and structures of languages spoken in the world.

Related Work

Our inspiration was the visualizations created by the Puff Puff team of the Dipartimento Di Design at Politecnico di Milano. During their Design Density course, they developed interesting visualizations using the same dataset to understand the popular families of languages, the number of people that spoke them and the loanword status of popular languages.

Link: http://www.puffpuffproject.com/languages.html

Questions

We are trying to answer the following questions through visualizations:

1. What are the different families of languages that has existed over time?
2. How many countries and people speak a particular family of language?
3. How many countries and people speak each language within a particular family of language?
4. What are the different gender systems used by different languages?
5. How are the word orders and grammar rules different between languages?

Data

The data set used for this project is the **"WORLD ATLAS OF LANGUAGE STRUCTURES".**

Link: https://wals.info/

The World Atlas of Language Structures (WALS) is a large database of structural (phonological, grammatical, lexical) properties of languages gathered from descriptive materials (such as reference grammars) by a team of 55 authors (many of them the leading authorities on the subject)

The data set is rich in features (~ 200 features) and has data for 2680 languages. We do not plan on using all the features for visualization. As we go around working on the visualizations we will create our data structures with the necessary features on fly using javascript and python (if required).

Implementation

Landing Page

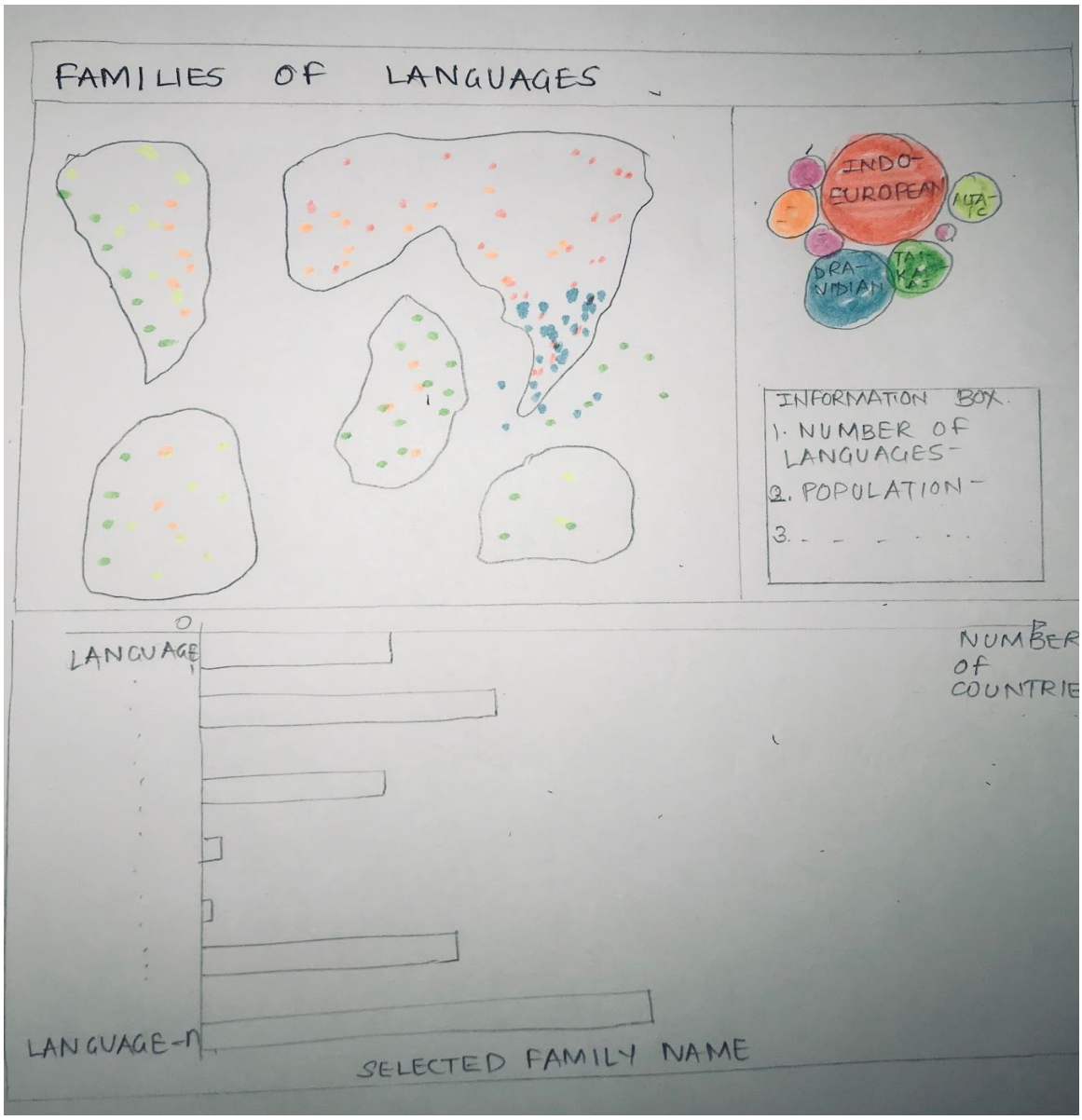
After the peer feedback session we decided to have a landing page.The landing page has options for viewing the two main visualizations that we have

1. Families of Languages
2. Language Structure and Nuances

Clicking on any of them takes you to the associated visualization pages.

In addition to these two pages, we have a ‘Project Details’ page which will have details such as Project Proposal, Process Book, Screencast etc.

Visualization: 1 and 2



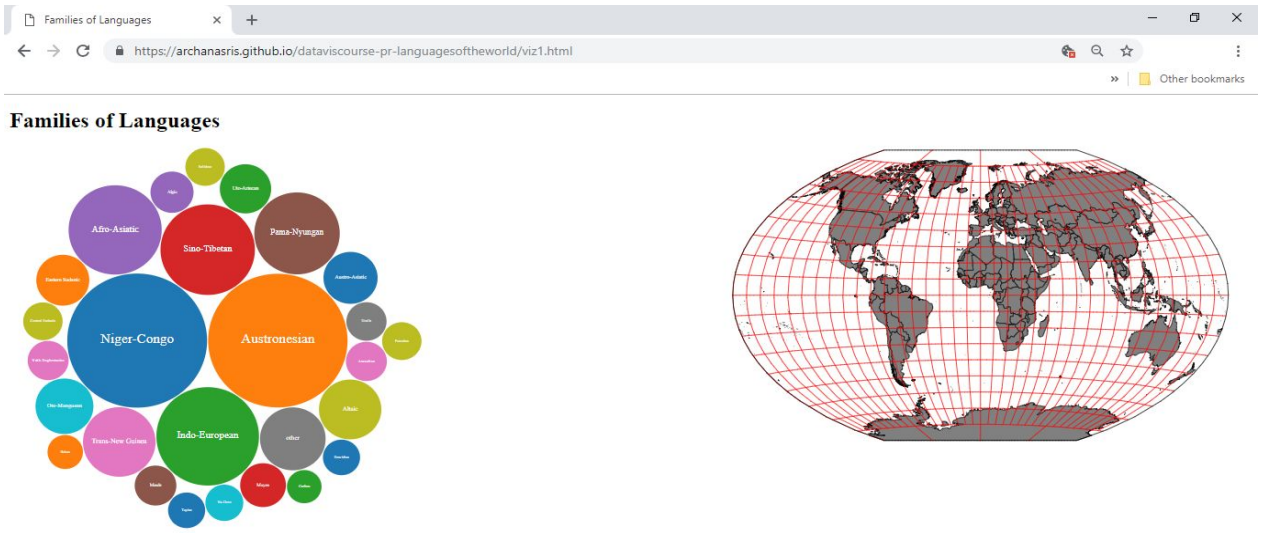
Data Processing:

1. In this visualization, all the families of languages with more than 20 countries speaking them were first extracted. They were later sorted and stored as an object.
2. A hierarchy structure (d3.hierarchy(myObject)) is used on the extracted dataset to identify the root nodes and children nodes.
3. The size of each bubble is dependent on the size of the family of language to show prominence of each of them and color coded with the help of d3.scaleOrdinal(d3.schemeCategory10).

Fig.Data used to plot the Bubble Chart

Visualization and Interactivity:

1. We plan on having interactivity by including selection of a particular family so the languages associated with that particular family are highlighted on the map based on the region where they are spoken. The data points on the map will be color coded.
2. The information box below the bubble chart will mention details such as the number of languages that belong to that family, the number of people who speak it etc. This would help in understanding the spread of macro languages over the world.
3. For each family of language, a bar chart will be created indicating all the languages in the family and the number of countries that speak each language. This bar chart will help in visualizing the popularity of each language within a family.



Visualization 3:

In this visualization, we plan to explore the languages and their gender based systems.

Data Processing:

1. Map - We use the **world.json** file for plotting the map. The projection used is **d3.geoWinkel3()**. As we go about with the development of the project we could look into other projections if required. The dataset has the location of the places where the languages are spoken (latitude and longitude) which we will use in order to plot those points on the map.
2. For the tree layout, the data is grouped using d3.nest using the feature gender based system in order to get out first grouping of “Sex-bases systems” and “Non-Sex based Systems”.
3. These groups are further separated based on the number of genders used in the languages to get our second level of data. These separated groups are stored in a object specified by us which we could append to the map or other elements to plot them.

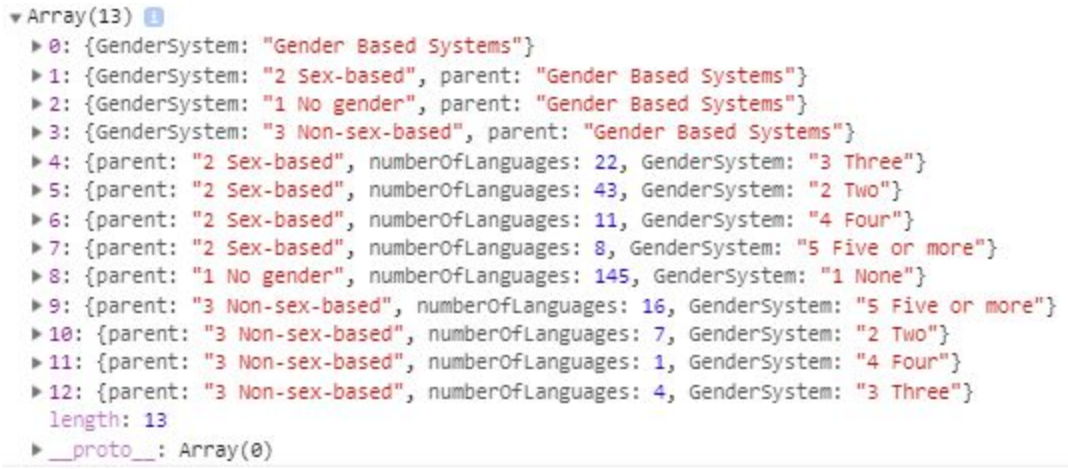


Fig.Data used to plot the tree

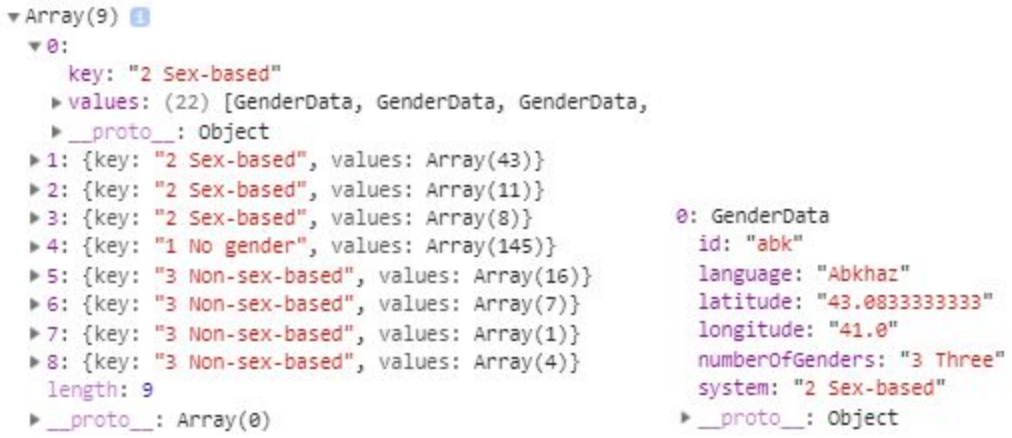
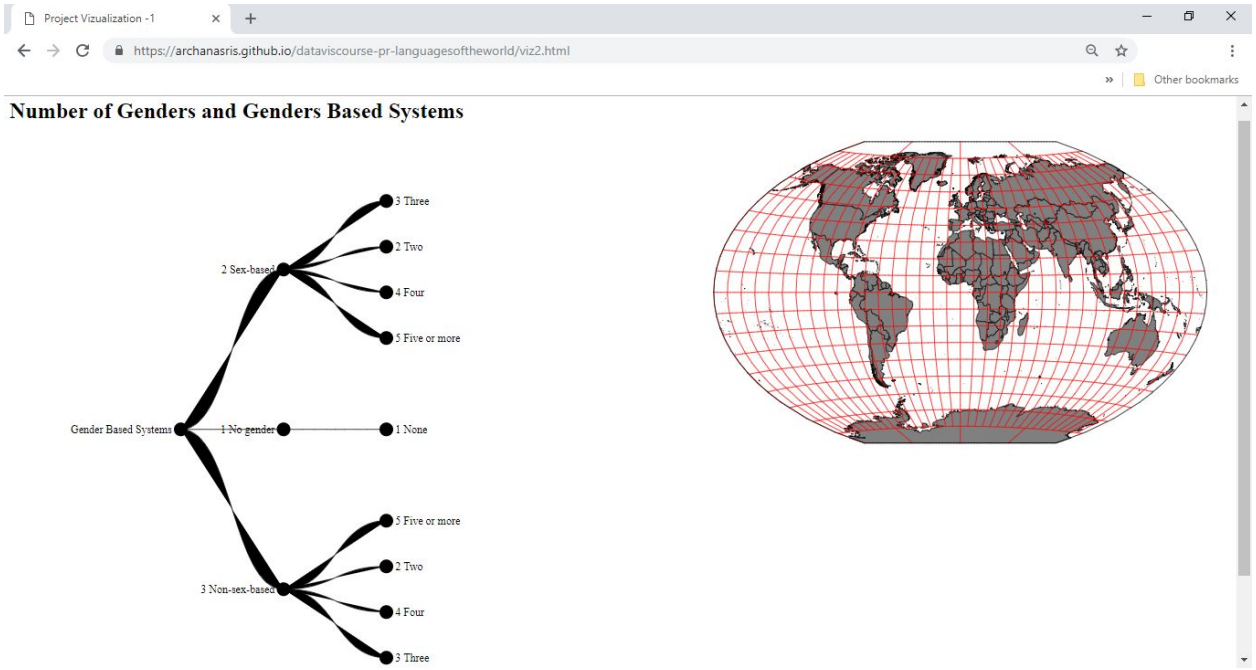


Fig.Dataset grouped by the gender systems and the GenderData object

4. Using the d3.stratify() and d3.hierarchy() a tree structure is rendered for this data.

Visualization and Interactivity:

1. We want to make the tree graph and the map interactive, such that when a node in the tree is clicked all the languages corresponding to that category are displayed on the map.
2. The categories will be color coded on the tree and on the map.



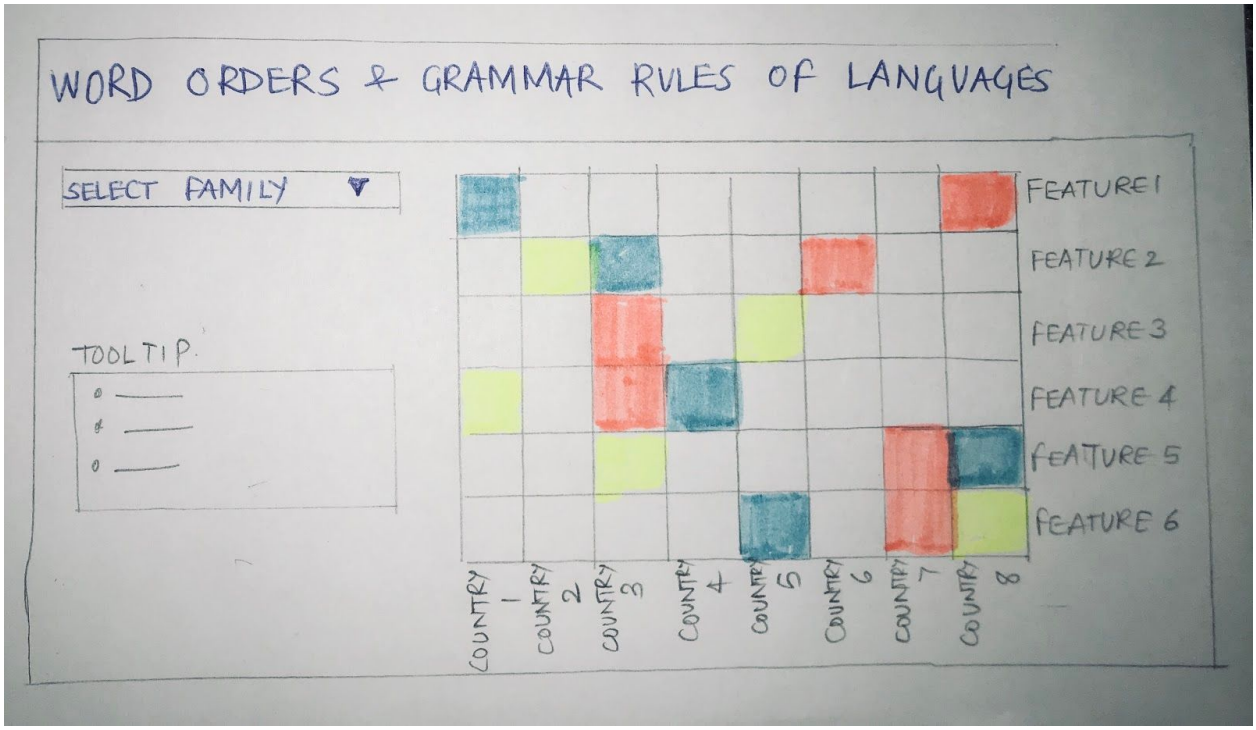
Visualization 4:

Data Processing: (To be implemented)

As per the feedback received from our peers, for the heatmap we plan to explore the grammar rules of two categories - countries and languages. Which ever presents intuitive patterns we will go ahead with the that category.

Visualization and Implementation:

Depending on the category of data selected during the preprocessing step we will integrate the interactivity with either visualization 1 to select family of languages or create a drop down for country.


WORD ORDERS & GRAMMAR RULES OF LANGUAGES
SELECT FAMILY
TOOLTIP
FEATURE 1
FEATURE 2
FEATURE 3
FEATURE 4
FEATURE 5
FEATURE 6
COUNTRY 1
COUNTRY 2
COUNTRY 3
COUNTRY 4
COUNTRY 5
COUNTRY 6
COUNTRY 7
COUNTRY 8