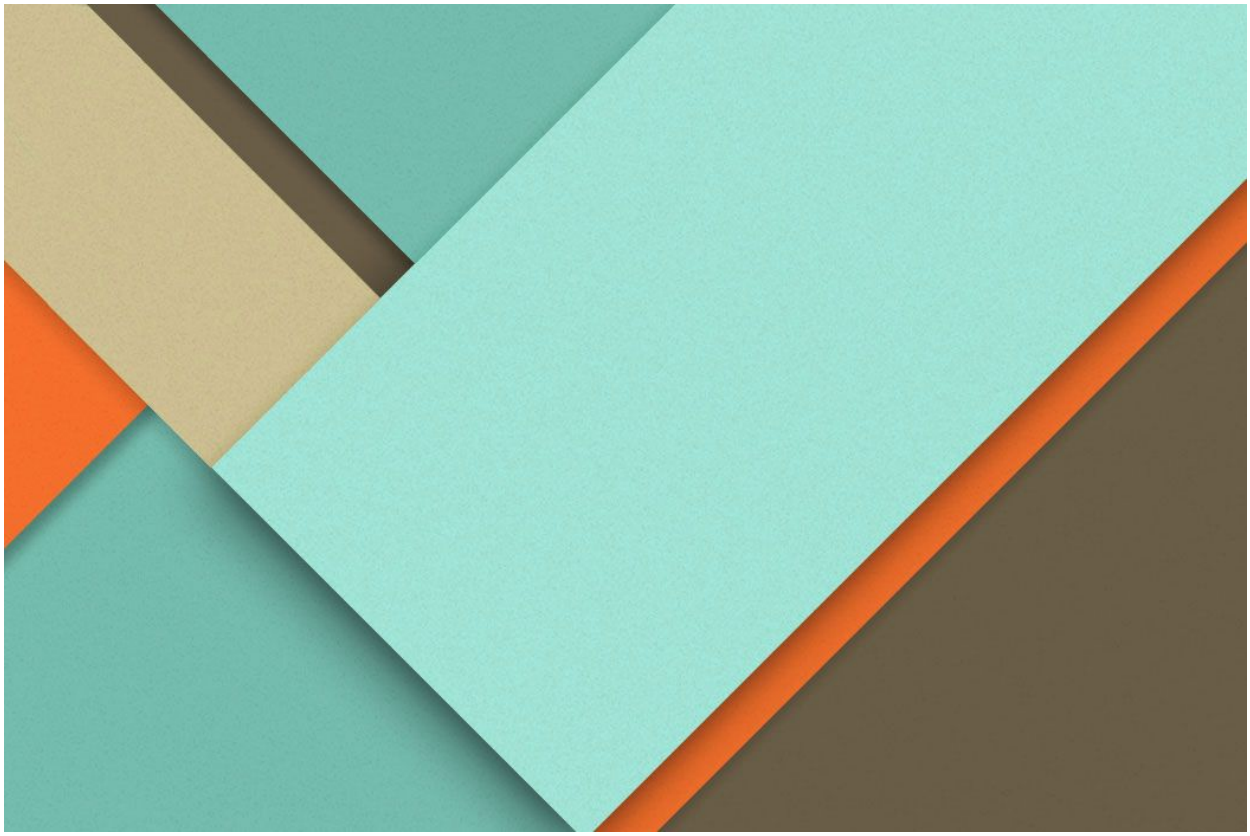


Languages of the World

28.10.2018

Archanasri Subramanian, U1141789

Krithika Iyer, U1135255

Overview

Team Members :

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

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

Background 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.

Project Objectives

Data we plan on visualizing:

1. Families of languages
2. Number of countries that speak a particular family of language
3. Language Structures and Nuances
    - Analysis of Number of Genders in Gender based Systems
    - Word Orders and Grammar Rules in languages

Data Source and Processing

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

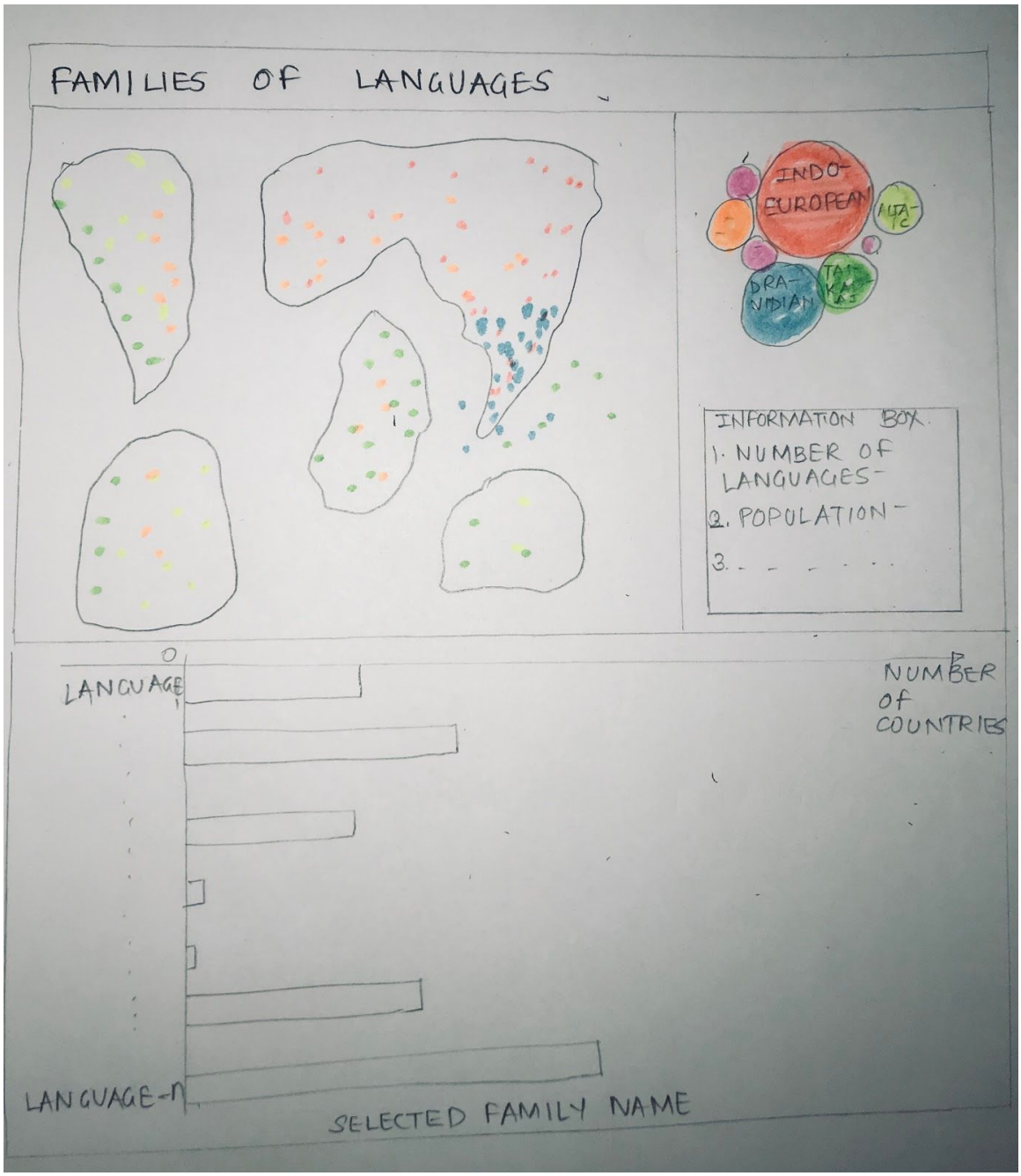
Link: https://wals.info/

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).

Visualization Design

We plan on using a combination of maps, bubble chart, bar chart and tree layout to visualize the relationships we have derived from the dataset.

Visualization: 1 and 2

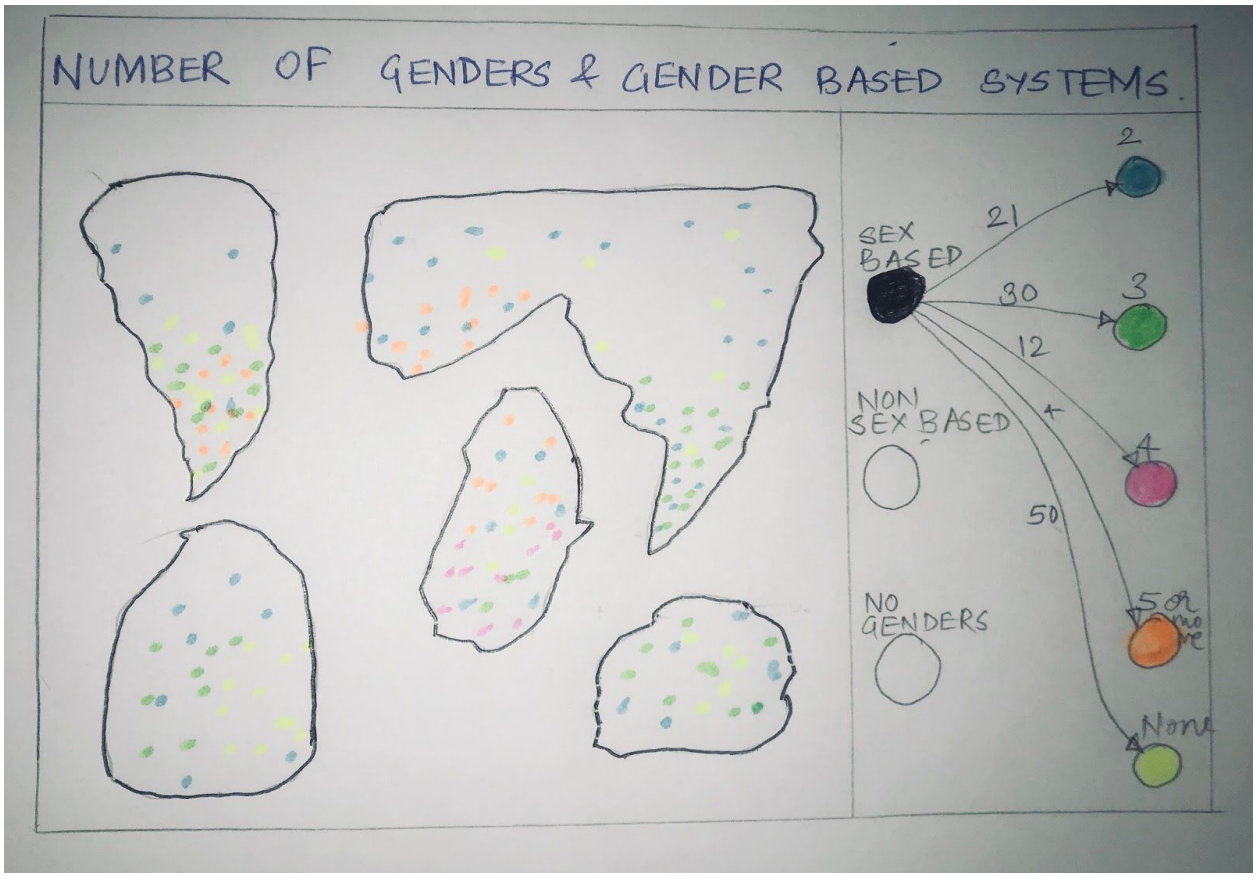


In this visualization, a bubble chart is designed which displays all the family of languages. The size of the bubble on the bubble chart will depend on the size of the family, thus showing prominence of the family language.

On selection of a particular family, 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. The information box below the bubble chart mentions details such as the number of languages that belong to that family, the number of people who speak it etc. This visualization helps in understanding the spread of macro languages over the world.

For each family of language, a bar chart is created indicating all the languages in the family and the number of countries who speak each language. This bar chart helps in visualizing the popularity of each language within a family.

Visualization 3:



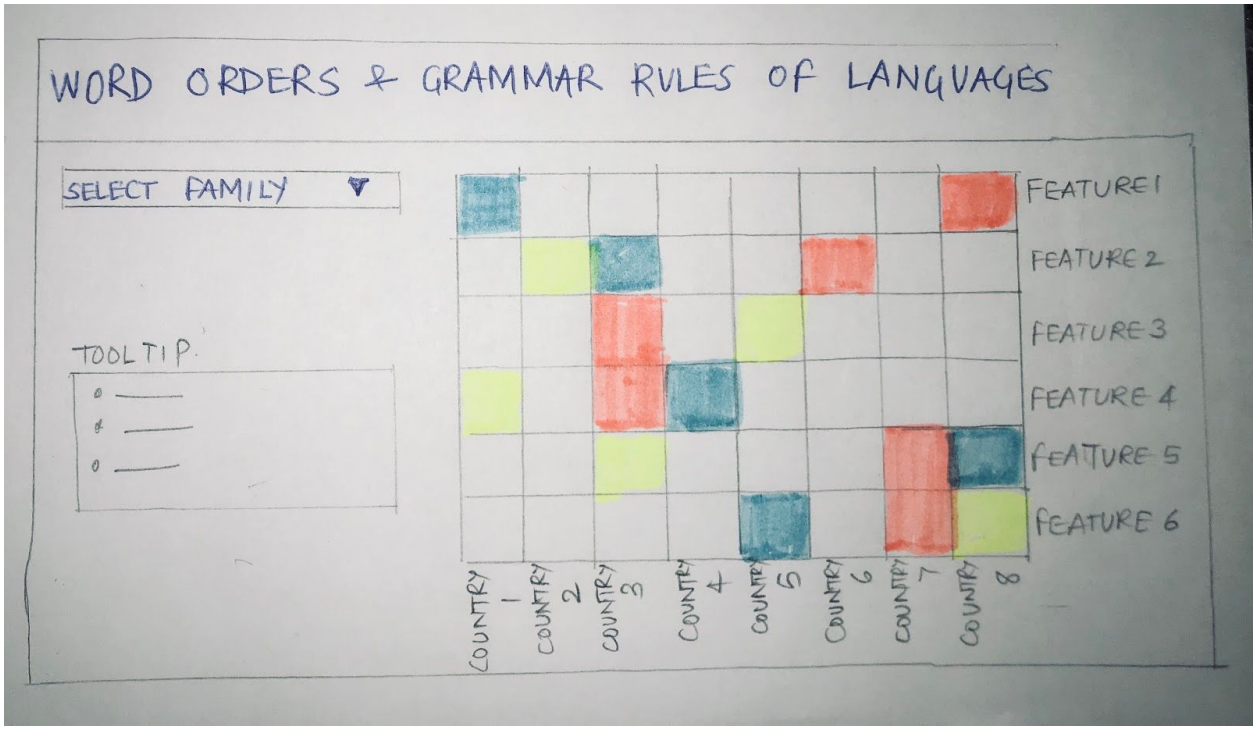
The visualization explores the the number of genders and gender based systems in different languages. So each language can fall into one of the 3 gender based systems and then there could be some fixed number of genders.

- Number of genders can take the following possible values : 2/3/4/5 or more/ None
- Gender Based Systems: Sex Based Genders / Non-Sex Based Genders / No Genders

A tree layout will represent the relationship between the above mentioned features. When the user clicks on a specific node of the gender based systems the links will show up. The number of languages which fall under the specific link will be mentioned in text on the link.

The number of gender nodes will be color coded and when the user clicks on specific node the languages which fall under that category will be shown on the map with the same color as the map.

Visualization 4:



This visualization explores the semantics of the different language families.

The user will be able to select a specific family which will be visualised. A heat map for the different languages of the selected family will be drawn with a few features. The features which we are hoping to visualise are:

- Order of Subject.Object and Verbs
- Order of Subject and Verbs
- Order of Object and Verbs
- Order of Adjectives and Nouns
- Order of Numeral and Noun
- Order of Relative Clause and Noun
- Position of Negative word w.r.t to Subject,Object and Verb
- Position of Tense-Aspect Affixes

When the user hovers over the heat map tool-tip information will be displayed showing the name and family of the language, value of the feature and some more relevant information if possible.

Project Features

I. Must Have:

The main objective of the project is to explore the different structures and properties of languages spoken in the world hence it will definitely include all the objectives listed above.

II. Optional:

If time permits, we would like to explore the some features like popular languages using wordmap, loanword status of words in each language.

Project Schedule

Week 1

1. Set up a GitHub website
2. Design outlines of visualizations

Week 2

1. Process data as required for each visualization
    a. Analyse the features selected and using javascript create data structures as required for the visualizations either on fly or store the data in separate csv files.
2. Start working on implementing the outline and have all visualizations in place.

Week 3

1. Add interactions and animations to enhance the visualizations

Week 4

1. Implement remaining visualizations with full functionality

Week 5

1. Wrap up work and if time permits, try to implement ideas mentioned under optional.