**Scientometric Mapping of Interpreting – Steps #1-10**

**Step 1:**

**Official project title**: Scientometric Mapping of Interpreting Globalization of the United States, 1789-1861

**Step 2:**

**Title of planned visualization**: Scientometric Mapping of Interpreting Evolution of topics, authors and collaboration’s over time and space

**Step 3:**

**Team**: Chandan Uppuluri, Vinay Kumar Ranganath Babu, Nisha Chadwani and Shruthi Ramakrishnan

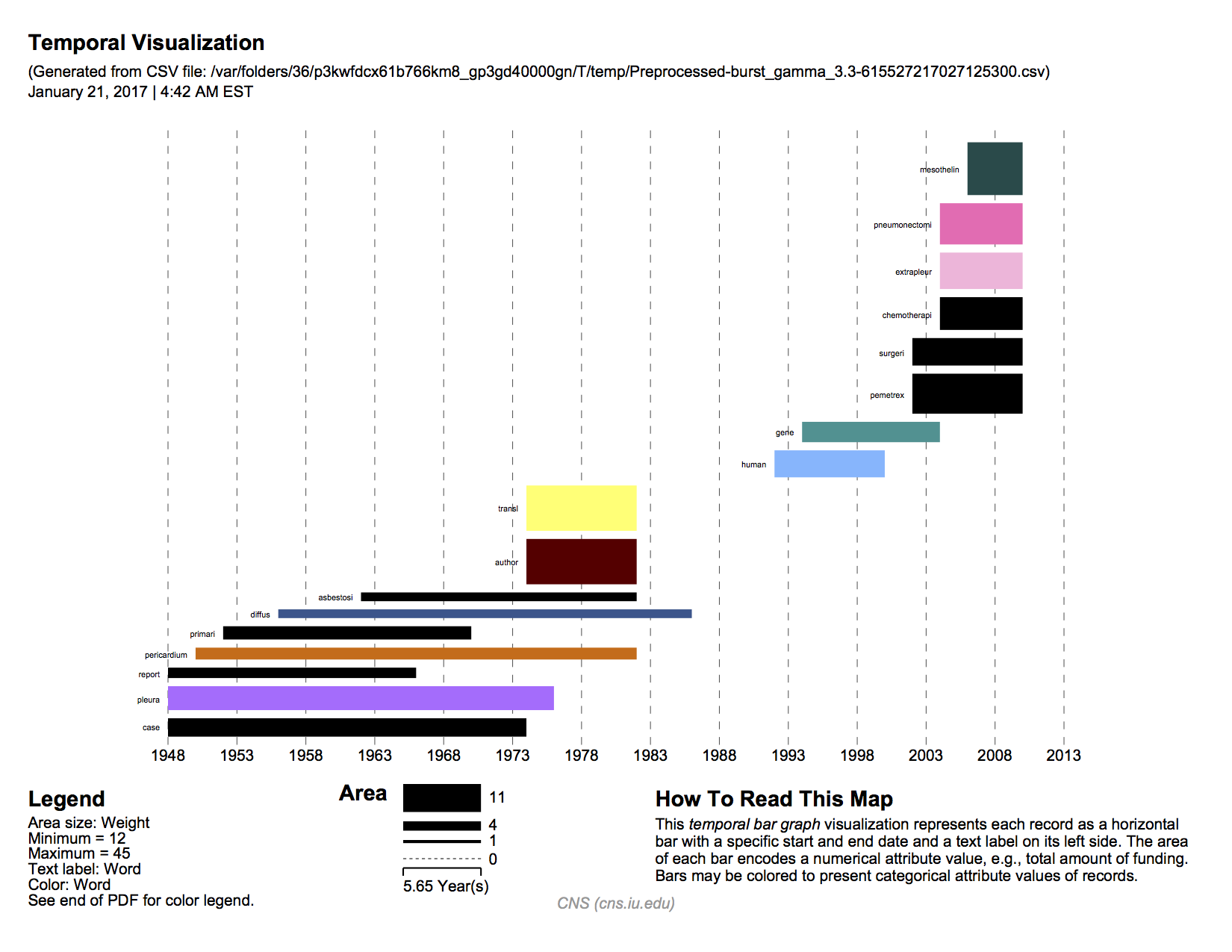
**Step 4:**

* **Description of visualization goal:** The client needs visualizations that gives insights about “domain of interpreting” by analyzing the articles that have been published related to this domain. We will be creating visualizations which gives insights about the evolution of topics, authors, collaborations and citations across time (when) and space (where: geospatial).
* **Planned Visualizations:** Below are the various visualization that we have planned to implement for this project:

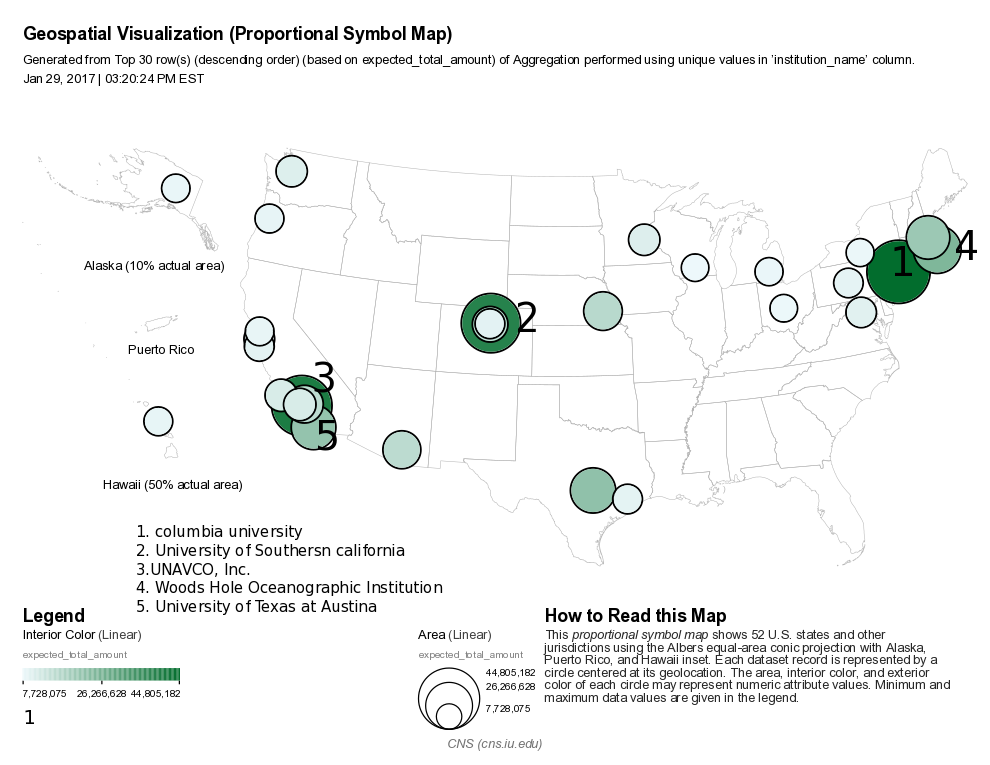
1. Topic Analysis – Word Cloud: The goal is to create a chart in the form of a world cloud. We will overlay topics data. Here is an example visualization:



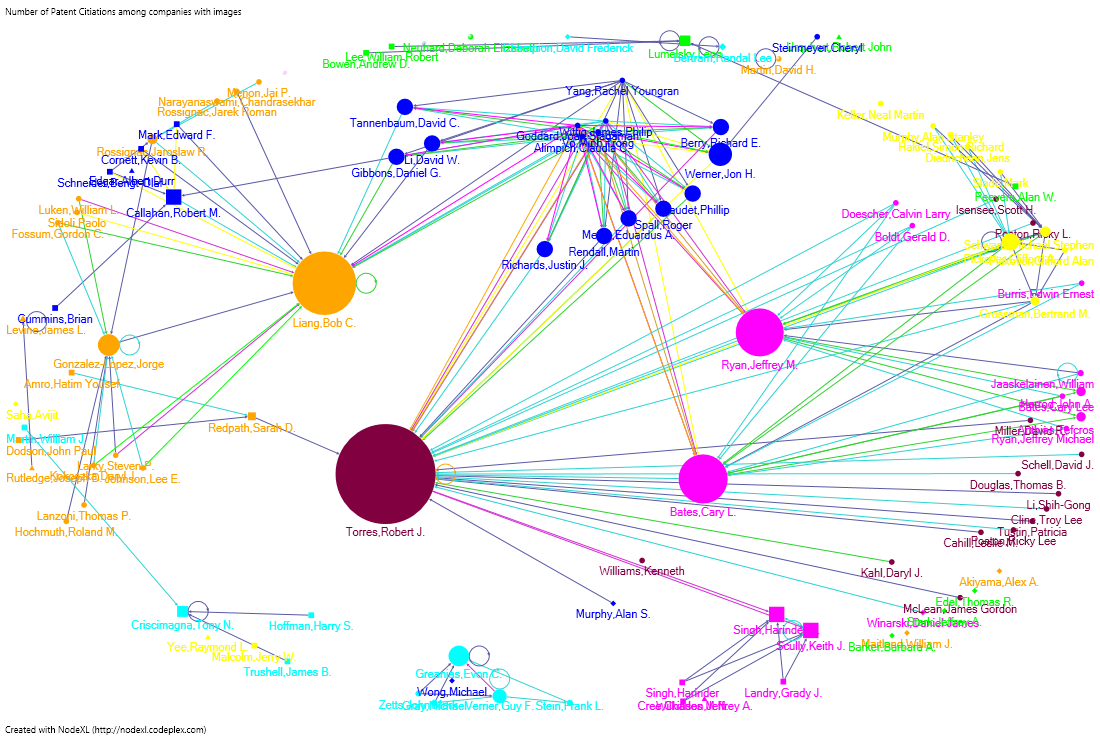
1. Topic Analysis: Burst of terms across time: In this visualization, we can see the burst of topics in the domain of interpreting across time. Below is an example visualization:



1. Geospatial Location of Lead Authors: The reference system will be a world map or any other map. We will be using proportional symbol map to overlay the data of number of authors that have contributed towards the domain of interpreting. An example proportional symbol map would look something like below (it shows the universities that receive the most amount of funding. And size of the symbol is proportional to the amount of funding. In our case this will be the number of authors that have contributed from that university or the location of lead authors, i.e., the authors who have contributed the most):



1. Paper Citation Network: The reference system that will be used here is a network graph that shows the citations network of the authors. This depends on the dataset that we have and how much information it holds as the dataset is still tentative and not yet finalized. But here is an example of Patent Citation network of Computer Graphics Processing Group [3]:



1. Geospatial location of topics across time: This uses world map or Country maps as a reference system. We will overlay lead topics that have evolved at various locations across the world. We can also include one more factor which is the time. We can produce series of maps for various time frames. The visualization would be similar to the one provided in (c) above though our attempt would be to get a more interactive visualization with respect to time.

*\*Please note that the above visualizations are just sample visualizations and not derived from the actual dataset of this project. They are just to give an idea of the visualizations that we are planning to implement as part of this project.*

* **Why this project is important:** Interpreting is the domain of translating speech or signs from one language to another [1]. Although science mapping is relatively established in other domains, the field of “interpreting” has not looked at itself from a Scientometric perspective. The project is a first step toward gaining first insights into this academic field by analyzing the academic articles published in the main journal in the field, Interpreting, using Scientometric analysis methods and information visualization

**Step 5:**

**Related work:** There is not much of a work done in the Scientometric analysis of the interpreting studies domain. But we have found very few works related to this. In Xu and Pekelis\_Chinese Interpreting Studies A Data Driven Analysis\_Peerj 2015 paper [5] it discusses about how we can identify various trends in the interpreting studies using quantitative and qualitative analysis. But this study focusses on mainly the Chinese works. Doors and Gambier Measuring relationships in Translation Studies Perspectives 2015 [6], in this study they mainly focus on the geographical spread of the Translation and Interpreting Research using Affiliations.

**Step 6**:

**About the Data (Statistics and Attributes):** The required data for this project is obtained from the Scopus database. It has research papers related to various domains including the field of interpreting.

The search query is as follows: "interpreting studies" OR "conference interpreting" OR "court interpreting" OR "medical interpreting" OR "sign language interpreting" OR "community interpreting" OR "simultaneous interpretation" OR "consecutive interpretation".

The dataset has 2932 records in them. Which suggests that there are 2932 records of different papers/articles related to Scientometric Interpreting. But it has a lot of noise. It Should be filtered to exclude publications related to statistics, remove non-articles such as books etc. The main attributes in the dataset which are relevant and interesting for our analysis were the below ones:

• Authors: This column gives us the name(s) of the authors who were responsible for the publication

• Title: This column gives us the title of the paper/publication

• Year: This column gives us the year in which the paper was published

• Affiliations 1: This column gives us the information of the author's affiliation to an organization/University

• Abstract: This column gives us the brief abstract of the paper's content

• geo\_latitude: The latitude value of the author's affiliated organization meaning from which place the paper emerged

• geo\_longitude: The longitude value of the author's affiliated organization meaning from which place the paper emerged

• Cited by: This attribute gives us the information on how many times the article was cited by other authors in their work

For the geo-spatial mapping, we can use the Affiliation’s field. It has the university location address. This address can be used for geocoding to obtain the latitude and longitude which can then be used for spatial mapping.

**Step 7:**

**Data analysis/visualization (algorithms) applied and resulting visualizations:**

1. Burst Detection: Temporal Analysis

We decided to perform 'Burst Detection' on two different fields namely 'Title' and 'Abstract' to understand on which topics/areas, the publications concentrated on. And finally, to create a temporal bar graph to visualize the burst detection over time.

Following are the steps which were carried out in executing 'Burst Detection':

• Load the data into Sci2 in 'Standard csv format'. The dataset loads and appears in Sci2's data manager

• Right click on the data and click on 'view' to check the right data have been loaded. We can also explore different fields of our data by doing so

• Since we want to perform Temporal Analysis via 'Burst Detection' and to visualize it via Temporal bar graph, our main concentration is on the fields Title, Abstract and Year

• The first and foremost step in our analysis to perform 'Burst Detection' is to normalize the 'Title' field. This will first lowercase the words, break the words into tokens, stem the words and remove all the stop words. This process will make sure that the 'Burst Detection' algorithm runs effectively

• In Sci2, click on the dataset and go to Preprocessing -> Topical -> Lowercase, Tokenize, Stem, and Stop word Text. Since we are interested in normalizing the 'Title' field, we need to select that and click on 'Ok'. The algorithm runs successfully and the 'Title' field gets normalized. We can cross-check this by viewing the file again

• After this pre-processing, the next step is to run the 'Burst Detection' on the normalized 'Title'. To do this, click on the normalized file, go to Analysis -> Topical -> Burst Detection. We need to input the below parameters. Keep rest of them to its default.

Data Column - Year

Text Column - Title

Click on Ok.

• When we right click, and view our modified dataset, we could see the word column which gives us all the bursting words from Titles. We also have burst Weight and, the Start date and End date for the burst. Many of the End date columns were empty which suggested that those words were still bursting. However, as we knew that the publications/papers till 2016, all the missing rows were imputed with '2016'

• Save the file with csv format

• Reload the saved file back to Sci2 with standard csv format. This new file would be used to create the temporal bar graph

• Select the file and go to Visualization -> temporal -> Temporal Bar Graph and input the below parameters:

Subtitle - Scientometric Interpreting

Label - Word

Start Date - Start

End Date - End

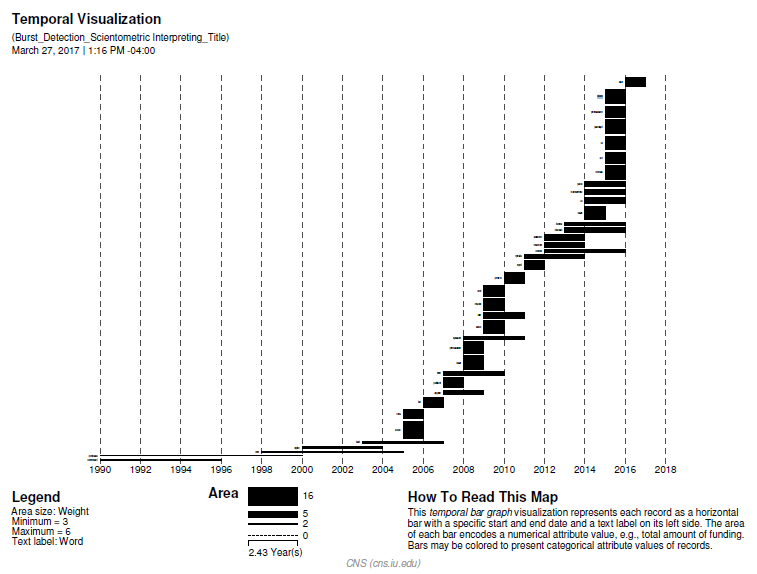
Size By - Weight

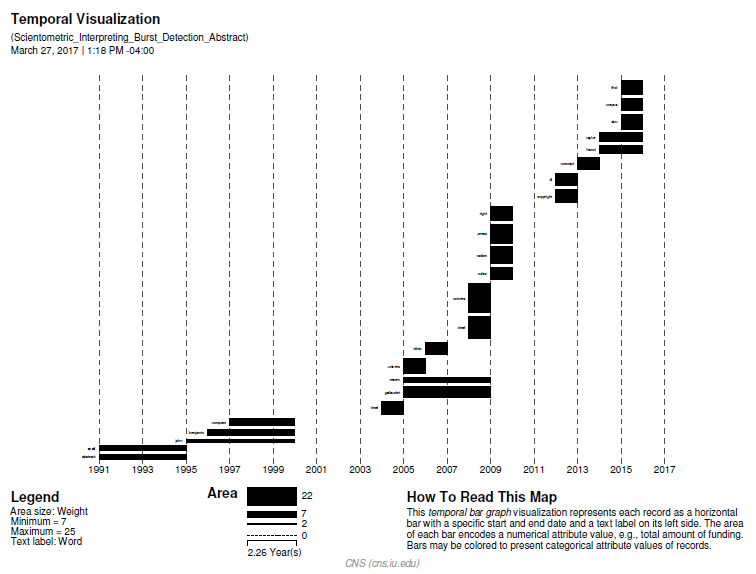
Scale Output - Checked

Click 'Ok'.

• The Temporal Bar Graph would be created in the Data Manager. The next step is to save this file as a 'PostScript' file and convert it into pdf

Perform the same steps for 'Burst Detection' on Abstract field. Below are the visualizations for the same.



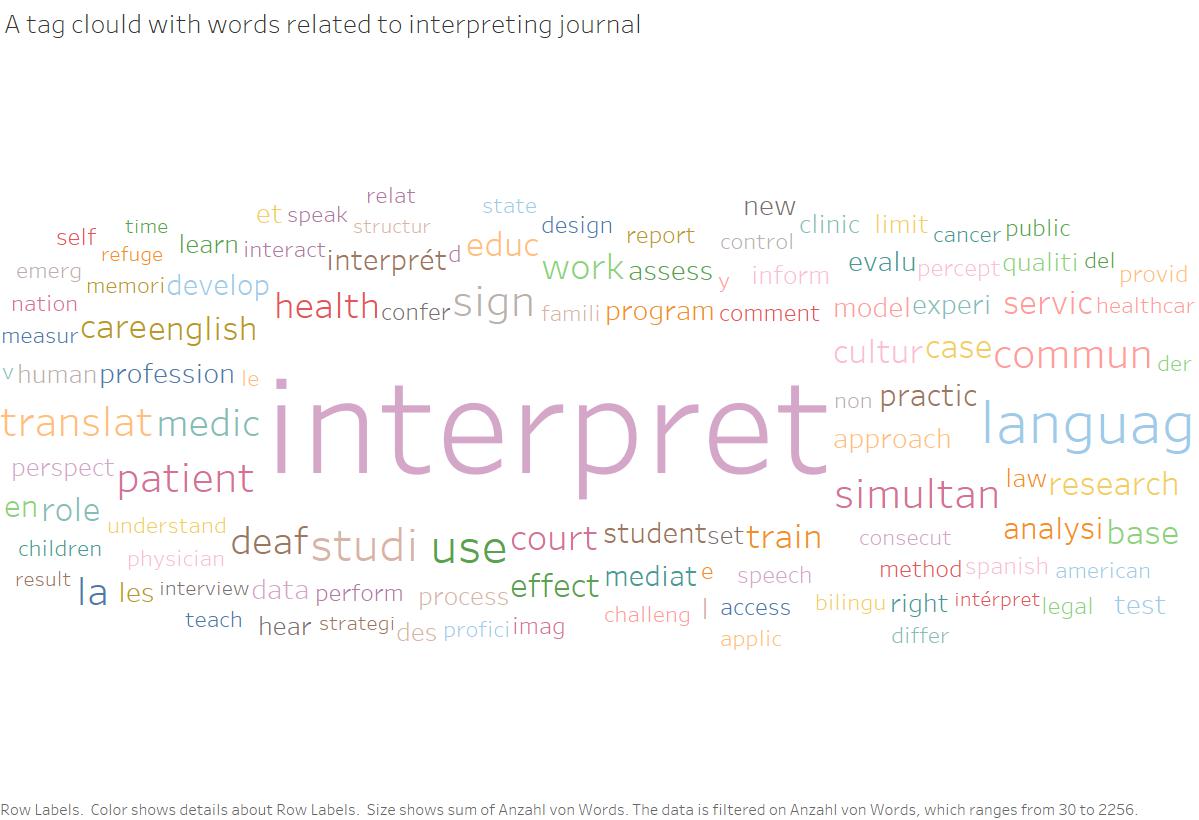


2. Word Cloud:

* Load the csv file onto Sci2 and perform preprocessing on that:

preprocessing -->topical --> lowercase, tokenize, stem and stopword text based on title

* Load the resulting file to excel. select the title column and split the words based on spaces. Find the frequency of each word.
* Load the words and frequencies into Tableau.
* Drag the words to Text and Frequencies to Size
* Select words with frequency above 30
* Drag the words to color
* The word cloud visualization is created
* Repeat the same process on the abstract column.



Above created from the field Title



Above created from the field Abstract

**Step 8**:

**Insights:**

* We preprocessed the original dataset by filtering out the records specific to the domain interpreting, cleaned the data set by removing nulls etc. But we still observe some noise in the results found as per the visualizations.
* Words like healthcar, servic, patient may not be related to our domain. We still need to look into the dataset to and try removing the unrelated terms.
* In the Temporal Visualization, we can see many bars. Each of the bars are associated with one of the bursting words. We can see these bursting words from 1990 to 2016. We can also see the difference in the weights for each burst term and can check how the topics/areas in Scientometric Interpreting evolved over time.
* Initially the algorithm did not filter away few things like years and single alphabets and these appeared as noise in our processed data as bursting terms. These had to be removed before creating the temporal graph.
* In the Title Temporal graph, we can see terms bursting like 'educ', 'studi', 'mediat', 'percept' and so on which emerged over time.
* In the Abstract Temporal graph, we can see terms bursting like 'gallaudet', 'univers', 'deaf' etc. which emerged over time.
* top 10 words for title: (from word cloud of title)
  + interpret 2,256
  + languag 368
  + use 224
  + studi 208
  + sign 183
  + patient 167
  + commun 163
  + simultan 157
  + translat 156
  + la 152
* top 10 for abstract: (from word cloud of abstract)
  + interpret 9,386
  + use 2,830
  + languag 2,610
  + studi 2,103
  + commun 1,547
  + patient 1,524
  + result 1,259
  + provid 1,180
  + court 1,088
  + differ 1,079
* Both word cloud and topic burst over time, shows which topics/words have more prominence in the domain of interpreting. It’s natural to have interpret word more weightage. But there are other interesting terms as well such as deaf, universe, court etc.

**Step 9**:

**What problems surfaced during validation and how does your redesign resolve them?**

On further analysis of the word cloud visualization which we had created initially, we realized that there were certain words which have a very high frequency of occurrence in almost all publications. Words such as Interpret, which occur in most of the publications were shown with significant importance in the word cloud. However, such words are not very helpful in deriving useful insights, but are rather words which are expected to occur more in general (since the domain itself is ‘Scientometric Mapping of **Interpreting**).

For creating the word cloud, we thus performed below pre-processing on the field, ‘Abstract’ to overcome this issue:

* Used Sci2 for pre-processing the Abstract field by applying, ‘Lowercase, Tokenize, Stem, and Stopword Text’.
* Used TF-IDF algorithm on the normalized Abstract field. This is to identify certain words that appear more frequently in general, such as Interpret in this case. The TF-IDF (Term Frequency – Inverse Document Frequency) algorithm [7] checks for how relevant a term is for a given document in a collection of corpus by considering the below-
  + - Compute TF (Term frequency): Measure of how frequently a word appears in the document, i.e., raw count of a term in a given document.
    - Compute IDF (Inverse Document Frequency): Measure of how much information that term provides, i.e., whether the term is common or rare across all documents
    - Take the product of TF and IDF to get the final score for each term in each document
* We then use this rank generated by the above algorithm to filter only the top 5% words from each publication.
* The words found using above are then mapped to their original count to generate the word cloud.

**Step 10:**

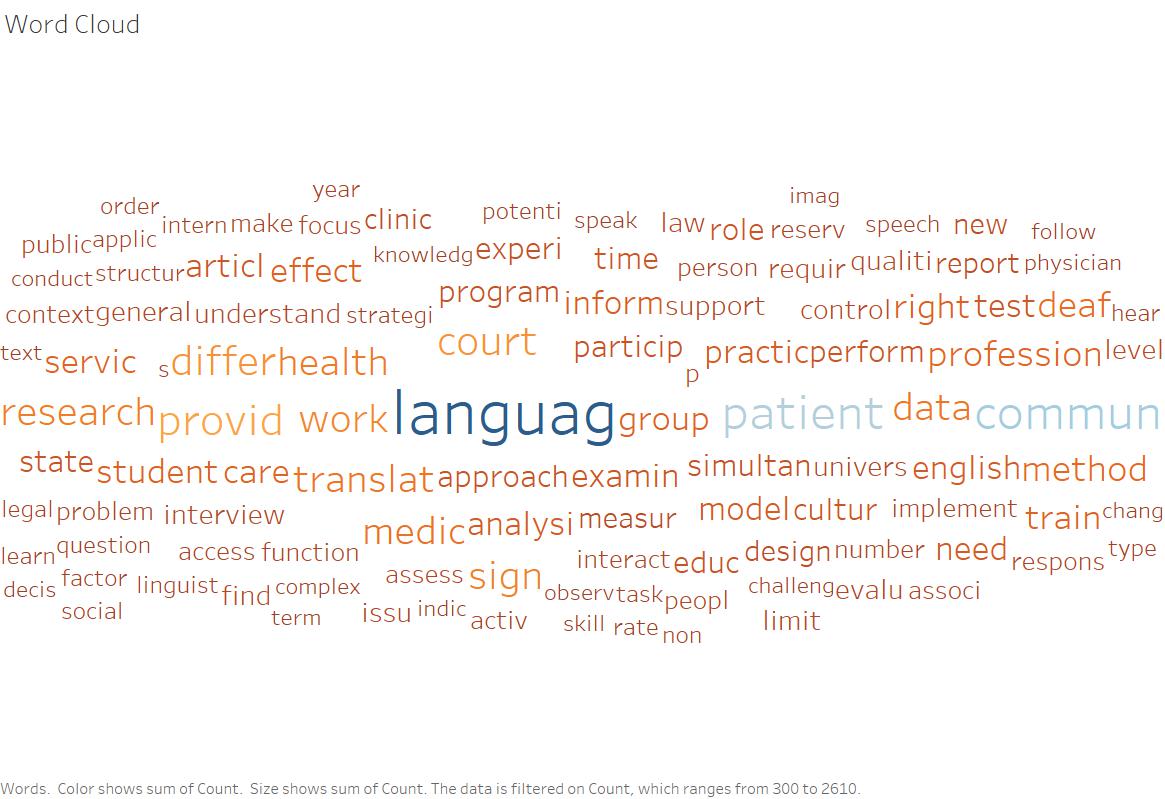
**Discussion of challenges and opportunities. Address complexity and scaling issues, desirable modifications and extensions of your work. You have only four weeks for this project restricting the amount of work that can be done considerably. I would like to know what promising avenues you see for future work.**

Discussion of challenges and opportunities:

1.) The major challenge is to obtain a clean data set as there is a lot of possible noise. But by refining query terms in Scopus database we can obtain a better dataset.

2.) Another challenge is to implement an interactive visualization’s. Using appropriate tools such as shiny web app/plotly/D3 etc should help us build interactive visualizations easily.

Extension to our previous work:

1. Word Cloud: After extracting the top relevant words from each publication (using the TF-IDF algorithm), we created the below word cloud visualization:

2. Co-authorship network:

**Steps for creating the visualization:**

* Read the csv file from Canvas into Sci2
* Extracted co-occurrence network based on properties file ‘project.properties’.
* Performed below steps for both the co-occurrence networks extracted:

1. Extracted nodes based on the attribute value, timescited > 3
2. Extracted edges based on the attribute value, numberofcoauthoredworks > 2
3. Removed isolates
4. Ran Network Analysis Toolkit to get the details of the network
5. Visualized network using GUESS and Kamada-Kawai layout
6. Resized nodes based on the number of times cited
7. Colorized nodes based on the number of works
8. Colorized edges based on the number of co-authored works
9. Labelled only the nodes with number of times cited >= 300

* Included legend and title using Inkscape

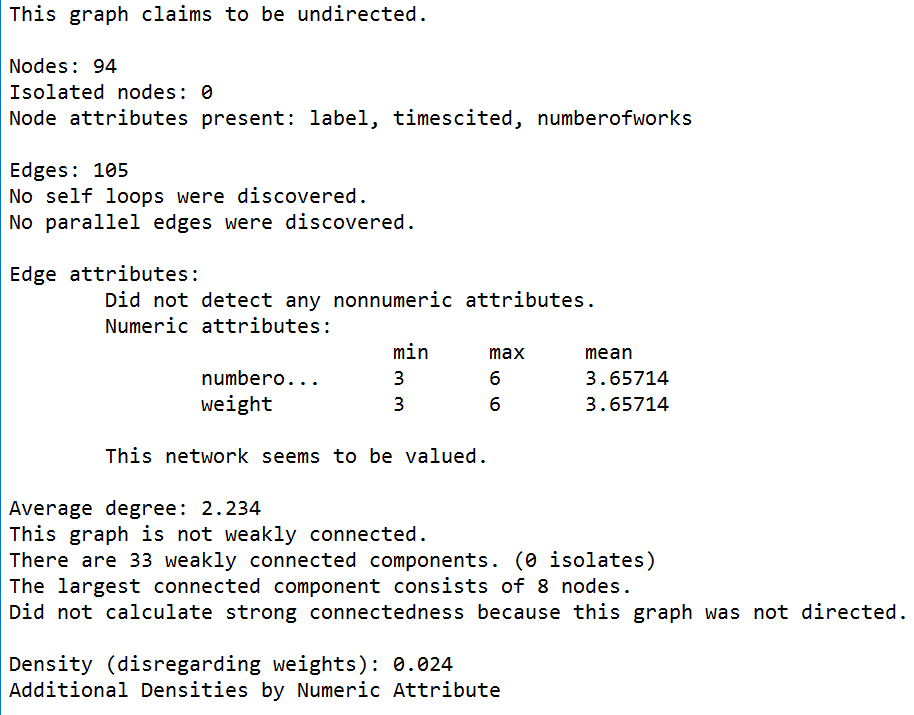
**Network Analysis Toolkit**

After filtering out the nodes based on ‘timescited’ and edges based on ‘numberofcoauthoredworks’, we have 94 nodes and 105 edges. Average Degree is 2.234.

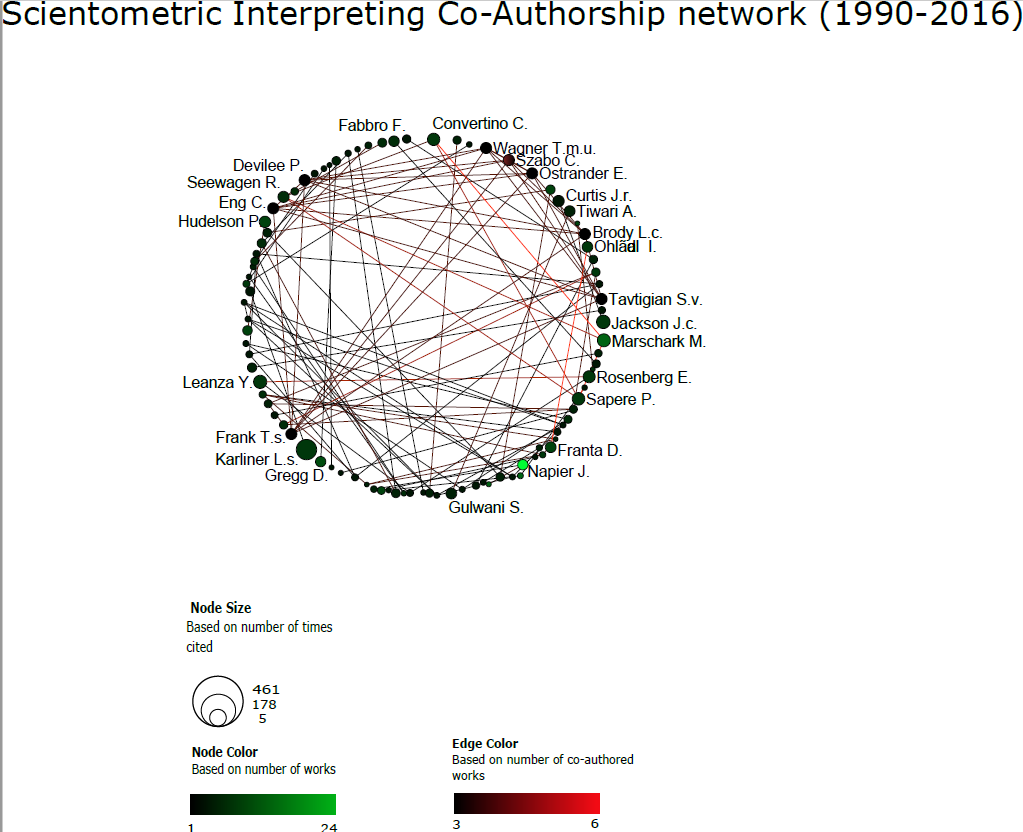
**Insights gained:**

From the visualization, we can see that there were many authors who collaborated with each other on this topic of ‘Scientometric Interpreting’ between the periods 1990-2016. We can see that many authors made a reference of others work and its visible in our visualization via ‘timescited’ attribute. The highest times a paper was cited was 461 and the lowest was 5.

Below is the snapshot of the statistics for this visualization:



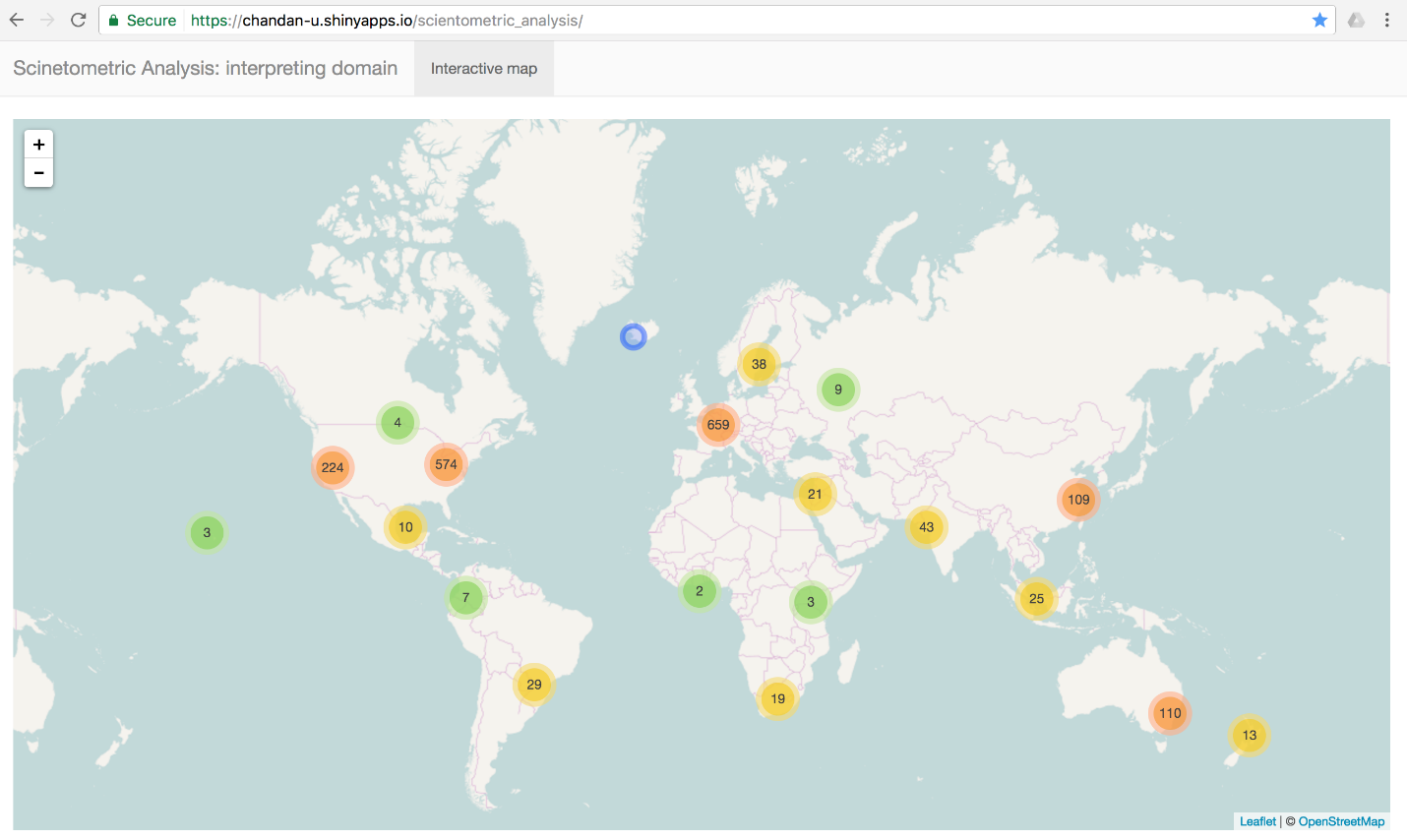
Below is the visualization:



3. **Geospatial Location of Lead Authors:** Please check the below URL for a more interactive version of this visualization:

**(**<https://chandan-u.shinyapps.io/scientometric_analysis/>**)**

Here is a snapshot of the visualization:



Challenges faced while creating the above visualization:

* The data given is partially geocoded. This wasn’t enough. We had to use the affiliations column to geocode as many addresses possible again related to lead authors. We used Google geocoding API to obtain the longitude and latitude of the lead authors.
* We wanted an interactive visualization where we can cluster the geo points as we zoom in and zoom out. For this we used R, leaflet, shinywebapp etc. The service is already hosted at the above-mentioned URL. This can be made better, with additional data filters which we are planning to do in the next iteration.

Future work:

As part of the remaining visualizations and modifications, we are planning the below:

1. Create a co-word occurrence network to highlight key words that co-occurred in most of the publications.
2. We are planning to use the information provided with respect to the University of the authors to identify the Universities across globe that contributed the most in this area.
3. Also, we are planning to use the field Language, which gives the original language of the publications across time. This will help us understand which languages dominated which timeframe.

**Acknowledgements:**

Lluís Baixauli-Olmos, Assist. Prof., Dept. of Classical and Modern Languages at the university of Louisville has helped with us in understanding the scope of the project and any other queries that we had. We have worked with LLuis closely to obtain the necessary dataset from the Scopus database.

**References:**

**[1]** Interpretation or Interpreting Wikipedia: [https://en.wikipedia.org/wiki/Language\_interpretation (Links to an external site.)](https://en.wikipedia.org/wiki/Language_interpretation)

**[2]** World Cloud example image: [http://www.jsquaredanalytics.com/word-clouds/ (Links to an external site.)](http://www.jsquaredanalytics.com/word-clouds/)

**[3]** Citation Network wiki: [https://wiki.cs.umd.edu/cmsc734\_11/index.php?title=File:Citation\_IBm.png (Links to an external site.)](https://wiki.cs.umd.edu/cmsc734_11/index.php?title=File:Citation_IBm.png)

**[4]** Sci2 Team. (2009). Science of Science (Sci2) Tool. Indiana University and SciTech Strategies, <http://sci2.cns.iu.edu>.

**[5]** In Xu and Pekelis\_Chinese Interpreting Studies A Data Driven Analysis\_Peerj 2015

**[6]** Doors and Gambier\_Measuring relationships in Translation Studies\_Perspectives 2015.

**[7]** https://en.wikipedia.org/wiki/Tf%E2%80%93idf