# Process Book

This process book is for the project “Whisper” which includes the analysis for the topic and review about top 50 companies from the Fortune 500 companies list.

**Haixuan Guo**, **Bishal Sainju**, **Kumar Akshat**

## ● Overview and Motivation:-

The Whisper Project tends to provide the visualization into the employees’ reviews, the project does represent the data regarding topic modelling and reviews via graphical visualizations. The project also does include topic modeling charts which does represent the various topics clustered as per similarity and documents. The pro and cons topics are described to understand how the employees rate the companies based on various attributes. The reviews also does include on the various attributes which vary from a score of 1 to 5. In summary the project does provide an overview summary for each companies including the basic information, reviews on various attributes and topic wise segregation on the pro and cons via the former and current employees.

In summary the project’s objectives are as follows:

a. The project’s audience is the company it refers to.
b. Various attributes can be used to compare among the companies, like reviews in various aspects, revenue, profits, employees, offices, and more general information. The comparisons can help us understand how people rate various companies in aspects like culture, work life balance, work breaks and many. Comparisons can also help job seekers to understand reviews and think about the work culture and decide in applying.
c. Whisper does provide insight into the general information about the companies, i.e. Share price, revenue, profits, employees, headquarters with a map representation of their offices.
d. Ability to compare two companies with each other in terms of general information.
e. The ability to compare various attributes in terms of graphs, like profit distribution.
f. Various topics segregated under pro and cons and clustered to provide a precise insight of how the topics are related to each other and help companies understand how the correlation and reviews can help understand the internal management and policy changes.
g. Visualize the ratio of topics present in documents and sentences.
h. How many former and current employees does provide the reviews using those topics in which section, i.e. pro and cons.

The online media is becoming a popular medium for employees to leave comments about their satisfaction and dissatisfaction about the company they are working for, or they had previously worked for. Thus, with the proliferation of large amount of online reviews today, we can look to mine these reviews, to explore the hidden factors that could contribute to boosting employee’s satisfaction. The traditional methods like questionnaire and surveys are becoming increasingly

outdated, and cumbersome in present day. They are firstly hard to form, takes a lot of time to prepare, and furthermore, they have narrow scope, because the user need to provide their opinions in limited set of topics. So, online reviews allows user to form open-ended survey, and thus allows employees to provide their opinion in an unbiased and uncontrolled way i.e. their opinion is unaffected by external factors. Online reviews provides efficient platform to collect the employees review, and furthermore the established sites like indeed provides millions of such user reviews, thus, providing plenty of data to decipher meaningful results from these data. Furthermore, these reviews are not just limited to certain set of factors as in traditional approach thus providing users freedom to expand their critique to wide range of topics, thus helping employers discover wide varieties of aspects to take into consideration to boost their employees satisfactions. Although online reviews are quite difficult to process and requires lots of computational effort to come up with any meaningful conclusion, however, with the increasing computational power to process big data, and advancement in machine learning models like topic modeling, these are being increasingly efficient to analyze, and as a result researchers are now pursuing their interest in mining these qualitative reviews.
For each company in Fortune 50, we want to build a persona to show the company's general information such as revenue, profit, ranking and size over the years. By depicting the time series data, we will see the overall operating condition of each company. Therefore, perhaps we can find some interesting connections between the company's performance and its employees' satisfaction.

- ## Related Work:-

  We came through a lot of research papers and links where we saw various ideas and we tried to learn from them. We have attached the links and tools that we found amazing and inspiring.

  Tools
  https://datavizcatalogue.com/ with d3 code

  Vis design
  http://measure.fathom.info/
  http://www.noceilings.org/education-income/#AFG
  https://fortune.com/fortune500/2016/visualizations/

  Awesome vis design
  https://www.makeuseof.com/tag/data-visualization-sites/
  A collection
  https://anvaka.github.io/pixchart/?d=4&ignore=0_12_14&link=https%3A%2F%2Fi.imgur.com%2Fq8BWCRJ.jpg&groupBy=rgb.b&initial=collapsed
  This one is amazing
  https://paleobiodb.org/navigator/

Cool map design
https://flagstories.co/
Pretty interesting story about flag
https://media.hhmi.org/biointeractive/click/virus-explorer/#/vaccine
Nice UI

Topic modeling Visualization
http://4humwhatevery1says.pbworks.com/w/page/104256241/Topic%20Modeling%20Systems%20and%20Interfaces
https://techxplore.com/news/2018-07-complex-network-based-approach-topic.html
https://dig-eh.org/dig-eh/TopicModelling/CircularDisciplines/#
http://www.datalab.uci.edu/papers/topicnets.pdf

- ## Questions:-

There have been a variety of questions that came to us while sharing ideas and planning this project, the most prominent ones are listed with the time we got into them, how we solved those questions and why did we choose the exact solution.

1. Who is our audience?
   Ans - We got this question a lot of times and we were quite confused about it but finally we arrived at the conclusion that the companies are the audience as to include a variety of audience would have made the visualization more complex and difficult to understand. The companies who are involved in the data are the audience and the platform can be useful for them

2. What are we trying to display?
   Ans - We got to this question while selecting the various chart methods and techniques to represent the way the charts will work. The selection of chart is the most crucial step in visualization. We were stuck in understanding which is the most relevant information to display and help the audience to get to a conclusion after using the platform. We discussed a lot and arrived at a point where we were clear that we are focusing our project on the reviews and topic modelling.

3. Why are we doing this project?
   Ans - This as been the existential question since the time we thought to moving ahead with this project. We thought of online platforms and saw that people not only gave star rating short reviews about companies but also did provide information regarding their ethics, work culture, environment, breaks and numerous attributes which if can be analysed and used with graphics can be very fruitful and easy to understand the pro and cons for each company. Thus we found it to be very exciting and started to work on this project.

4. Why use visualization and not just statistical representation to show data?
   Ans - We came through this question while we were planning, why are we considering that visualization should be considered as a good option for data representation and not the other methods, one of the reasons to do was that we are creating a project for data visualization but later on as we went more into planning and meetings, we came to realise that visual representation will always be easiest to understand even for a person who has no prior knowledge of the field.

5. How to make the topic modeling more easily understandable?
   Ans - The data analysis produced a categorization of topics. The topics were segregated as per the pro and cons for each company and while using visualization it did seem to make sense but was cluttered and difficult to understand. Thus we came up with the idea to use clustering to use with the pro and cons topics which created a more understandable method to help understand the topic modeling.

6. Why is the general information required and is it helpful?
   Ans - While creating the project we were focused on topic modelling and dedicated towards the use of visualization techniques for various topics, we then realised that understanding the topics and reviews would be more interactive and interesting if we put more information like company profile and their revenue.Then we planned to put general information about the various companies.

7. Can the platform be used for the candidates who are searching for jobs?
   Ans - We thought about this initially and came to understand that the conflict in the idea of who the audience is not a small problem thus we limited the platform usage to the companies and so it can not be quite fruitful for the candidates searching for jobs.

8. Is map relevant when we are discussing is about companies and their reviews?
   Ans - We thought of putting a map which will show the 50 companies present in the states and while clicking on them will provide the general information, clicking on two circles will provide information regarding two companies which can be used for comparing their general information.

9. Does star rating reviews important and fruitful for the companies to understand?
   Ans - Aside from topic modeling we had data for company review in various attributes in terms of star rating, we thought to use that too for review and put it inside the general information.

- ## Data:-

We are collecting our data from Fortune 50 companies site, and Indeed. We will gather information about the company's rank, their profit and revenues over the year, and company size in terms of employee number and it's assets, to analyze the basic information about the company, and how their ranking have changed over time.

Similarly, the other part of our project will be focused on finding out aspects of employee satisfaction, and employee turnover determinants from Fortune 50 companies employee reviews on Indeed. So, for this we will collect our data of the employee reviews for each of the Fortune 50 companies.

**Fortune 500 (https://fortune.com/fortune500/2019/):**
The Fortune 500 is an annual list compiled and published by Fortune magazine that ranks 500 of the largest United States corporations by total revenues for their respective fiscal years. The list includes publicly held companies, along with privately held companies for which revenues are publicly available.The companies that are included in the survey are those companies that are incorporated in the U.S. and operate in the U.S. and file financial statements with a government agency.

The site provides various information regarding companies Profit, Revenue, Assets, Total Stockholder Equity and so on. However, for our visualization purpose we will only be using data related to profit, revenue, employee number, and total assets of the company for the fiscal year ended on or before Jan 31, 2019.

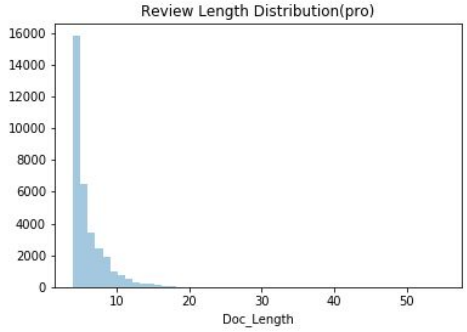
The information could be scraped from fortune site: "https://fortune.com/fortune500/2019/*/", where '*' is replaced by company name identifier. Only attributes like 'Number of Employees', 'Revenues', 'Revenue Percent Change', 'Profits', 'Profits Percent Change', 'Assets', 'Market Value', 'Sector Type', 'Industry Type' will be scrapped for further analysis..

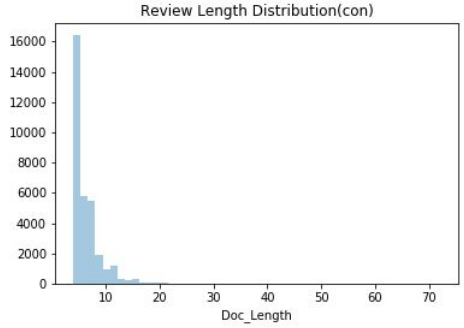
**Indeed (https://www.indeed.com/):**
Indeed is an American worldwide employment-related search engine for job and has the most listing compared to its other competitors (Glassdoor)(https://www.reviews.com/job-sites/). Furthermore, with its easy-to-use user interface, and extensive features, it has become one of the best platforms for the employee's to express their opinions regarding the companies that they work for (current employee) or that they previously worked for (former employee). Moreover, Indeed has one of the highest counts of the number of reviews available, which are readily served for analysis. That's why, Indeed's employee reviews will be used for the purpose of our analysis for figuring out latent determinants for employee satisfaction.

And, also it is legal to scrape “https://www.indeed.com/cmp/*/reviews”. So, the companies reviews will be scraped using this URL, replacing ‘*’ with the company name that indeed used in its site for company reviews. Only 50 companies of Fortune 50 companies will be used for our job satisfaction analysis purpose, the list of which will be obtained from Fortune 500 site. For our analysis purpose, only the attributes like “Review Title”, “Reviewer Job Status”, “Review Text”, “Pros Texts”, “Cons Texts”, “Ratings” will be scraped.

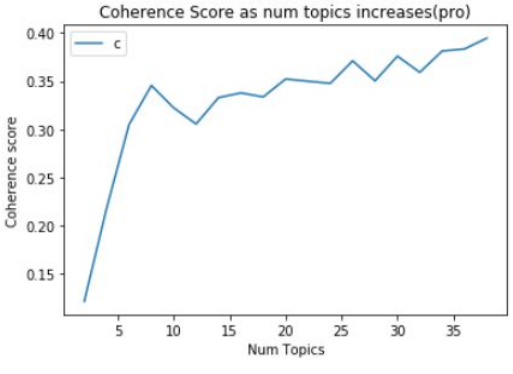
- ## Exploratory Data Analysis:-
- What visualizations did you use to initially look at your data? What insights did you gain? How did these insights inform your design?

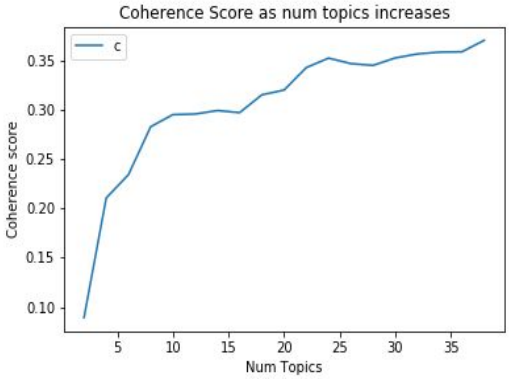
  So, basically we used distribution plot on seaborn to analyze the distribution of the terms for feeding those terms on the Topic Modeling algorithm. We did this as a text pre processing step for building the model.




Then we fed the texts to the LDA algorithm, and extracted the topic term distribution and document topic distribution. We need to feed in 3 parameters to LDA, beta, gamma and the number of topics to cluster into. We set the beta and gamma value to auto, and plotted the coherence score, with the number of topics to find out the best value of topics for each “pro” and “con” models.



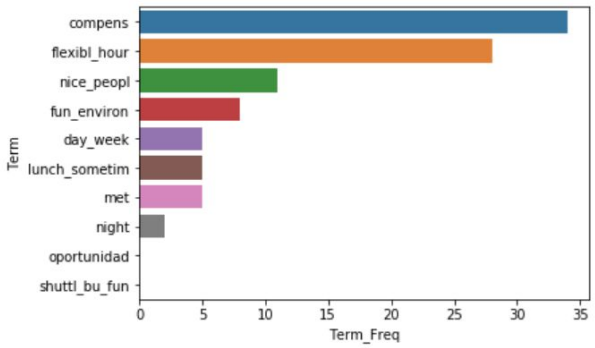
So, we decided to pick 10 number of topics for the prototype visualization, however, we might tweak this value later. Or have the option for the user to choose the number of topics so that they can see how the models looks like for various number of topics.

We then first plotted the PCA components using bokeh scatter plot, which looks as follows:

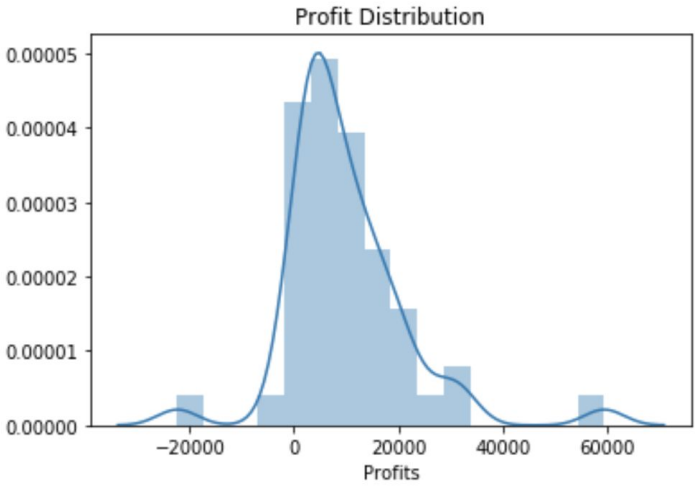

The topics looks quite cluttered, some of the topics overlap on top of each other. So we rescaled it in d3 before plotting them.

We thought of merging both of those models together onto a single plot but it did not make sense, as they were cluttered, and the similar pro and con topics were not together.


Then we also precomputed the relevance values and figured out if it was working or not.



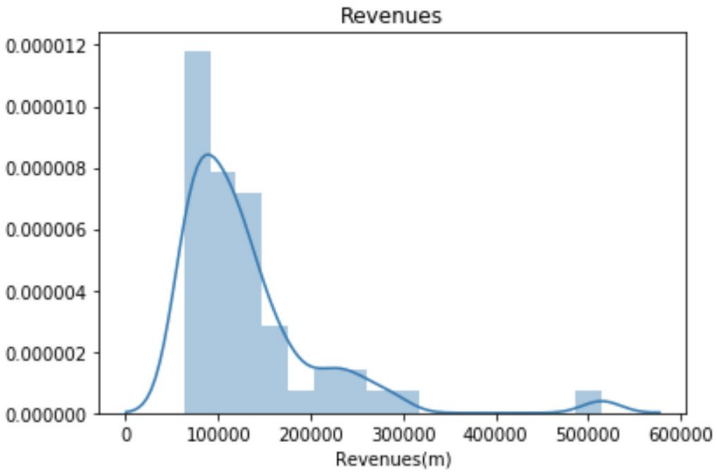

We also did some analysis on companies general information. We looked at the profit distribution of a company using KDE plot.




And it was surprising to see that some companies had even negative profits.

| | Rank | Company_Name | Employee_Number | Revenues(m) | Revenues(pc) | Profits | Profits(pc) | Assets(m) | Mar |
|---|---|---|---|---|---|---|---|---|---|
| **7** | 8 | CVS Health | 295,000 | 194579.0 | 5.3% | -594.0 | -109.0% | $196,456.00 | |
| **20** | 21 | General Electric | 283,000 | 120268.0 | -1.6% | -22355.0 | - | $309,129.00 | |
| **33** | 34 | Dell Technologies | 157,000 | 90621.0 | 15.2% | -2310.0 | - | $111,820.00 | |
| **56** | 8 | CVS Health | 295,000 | 194579.0 | 5.3% | -594.0 | -109.0% | $196,456.00 | |
| **69** | 21 | General Electric | 283,000 | 120268.0 | -1.6% | -22355.0 | - | $309,129.00 | |
| **82** | 34 | Dell Technologies | 157,000 | 90621.0 | 15.2% | -2310.0 | - | $111,820.00 | |

Also we did distribution plot on Revenues using KDE.




These various analysis gave us the insights about whether or not our data manipulation were correct or not, and provided us with the guidance to move ahead with our visualization. Furthermore, it also provided us insights about what should and should not be done in the visualization.

- **Design Evolution:** What are the different visualizations you considered? Justify the design decisions you made using the perceptual and design principles you learned in the course. Did you deviate from your proposal?

  Sunburst chart - The sunburst chart was the idea implementation to show the sector division and the sector the companies belonged from whenever they are compared or put to selection. The chart functions as - whenever a circle on the map is selected, the sunburst chart would show the sector the map company belonged from and also will

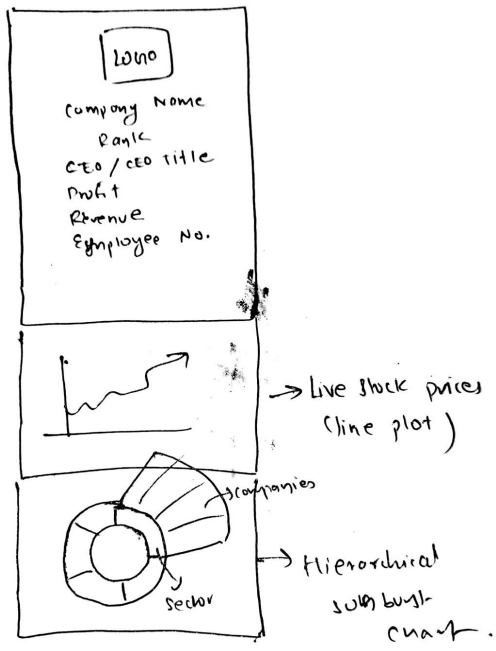
highlight the other companies that belong to the same sector. For this we considered a variety of different visualization techniques, like pie charts and statistical graphs, which will highlight the companies belonging to the same sector. The design decisions made during the selection included pie charts, statistical graphs and bubble clusters for represent this scenario. While selecting this we did deviate a bit and started selecting various techniques which were irrelevant, but as the project went further we were able to find the

Company map - The company map was one of the most crucial steps as the project loads, it shows the company map with the points to visualize the company names and then helps in considering the company names and the comparisons. Since the beginning, we had a plan to use the map to visualize the company list.

Company profile:

We have for now loaded the rough profile card for each company. We are still thinking of adding some more information there. We haven't yet completed the hierarchical sunburst chart of companies sector and industry. We will do that in the next phase. Similarly, we will include live stock price graph later on. We are still sticking to the same previous design.

Our design for company profile looks like below:



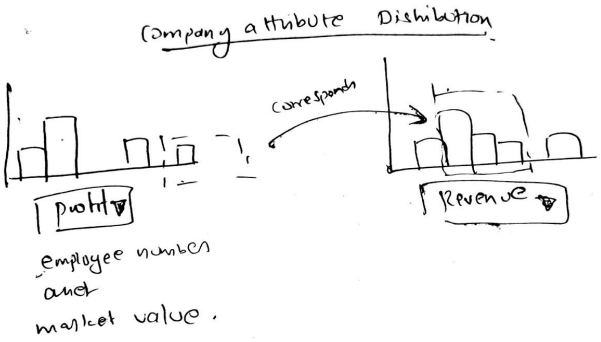

Company Profile Card.

The live stock prices should show the live share prices update using line plot. The hierarchical sun burst is a slightly modified version, where it's a sunburst just that it will not burst, but it will just highlight the sector and all the companies in that sector so that user can easily navigate to other companies in the same sector, which should be interactively highlighted on the map, and profit/revenue distribution bar chart. Also on hovering on the right side and left side of the company profile container, we are planning to make right and left arrow pop up so that one can easily access the companies one higher rank then the current, and company one rank below the current.

Company Profit/Revenue Distribution:

We haven't yet implemented this part. We are still working on normalizing data for this part. We already have company profit and other variable values, we just need to bin it and create histogram values, so that we can easily load this up in js. And then we will have to make it interactive.

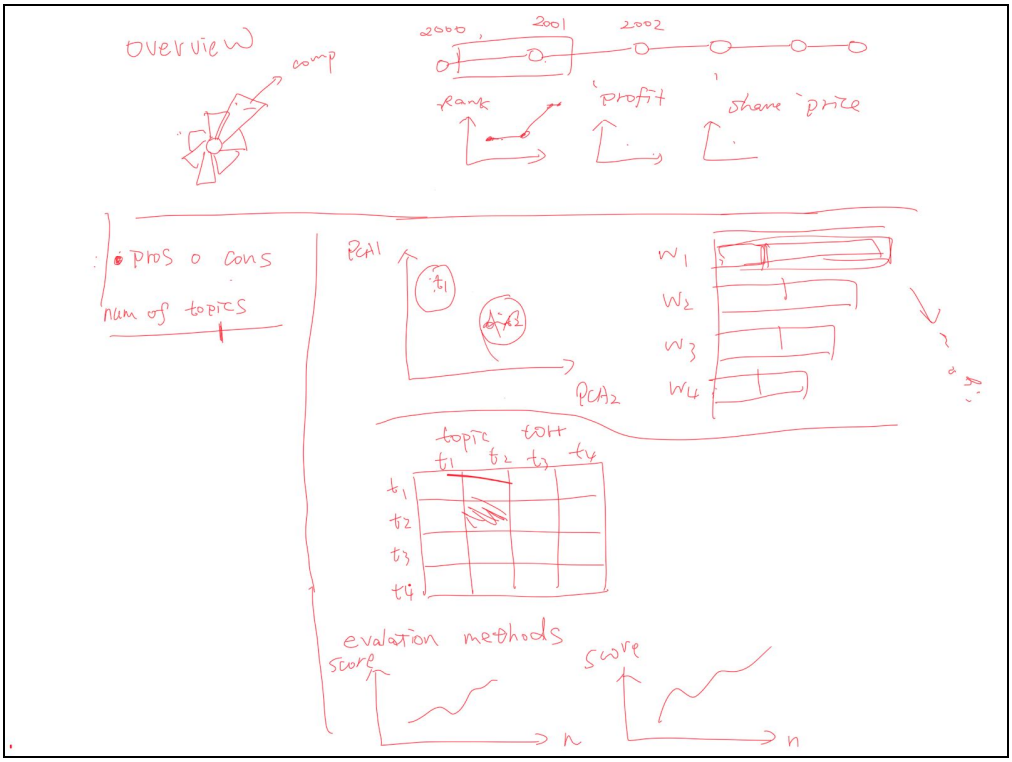
Our visualization should look something like the following:





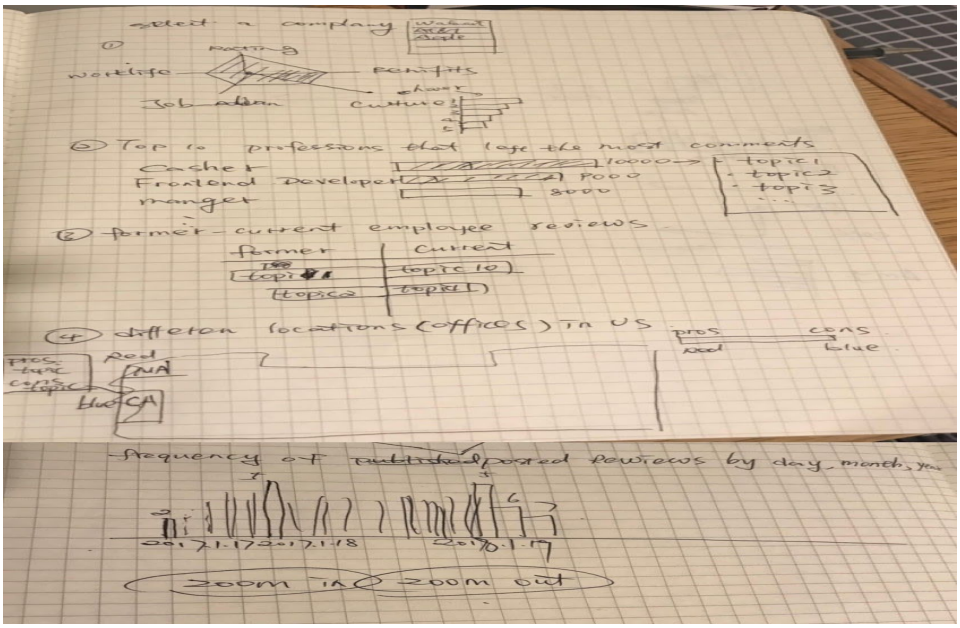

So we will have brushing and selection feature on barchart selecting on some bars will display corresponding companies in the map, and also there will be a drop down menu indicating the various attributes that we can select.

(bishal and haixuan)Topic modeling:



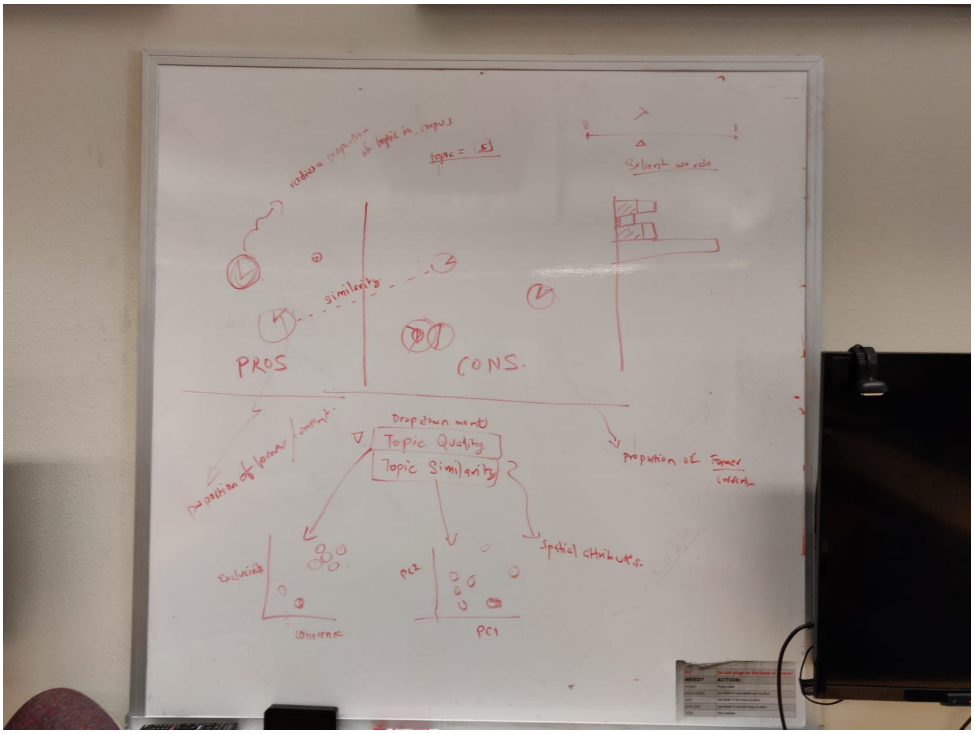

Basically this was what we thought of initially, but there were so many problems. There were not many time series data and also, there were so many information, and our audience was not clear.

Then we came up with the following idea.



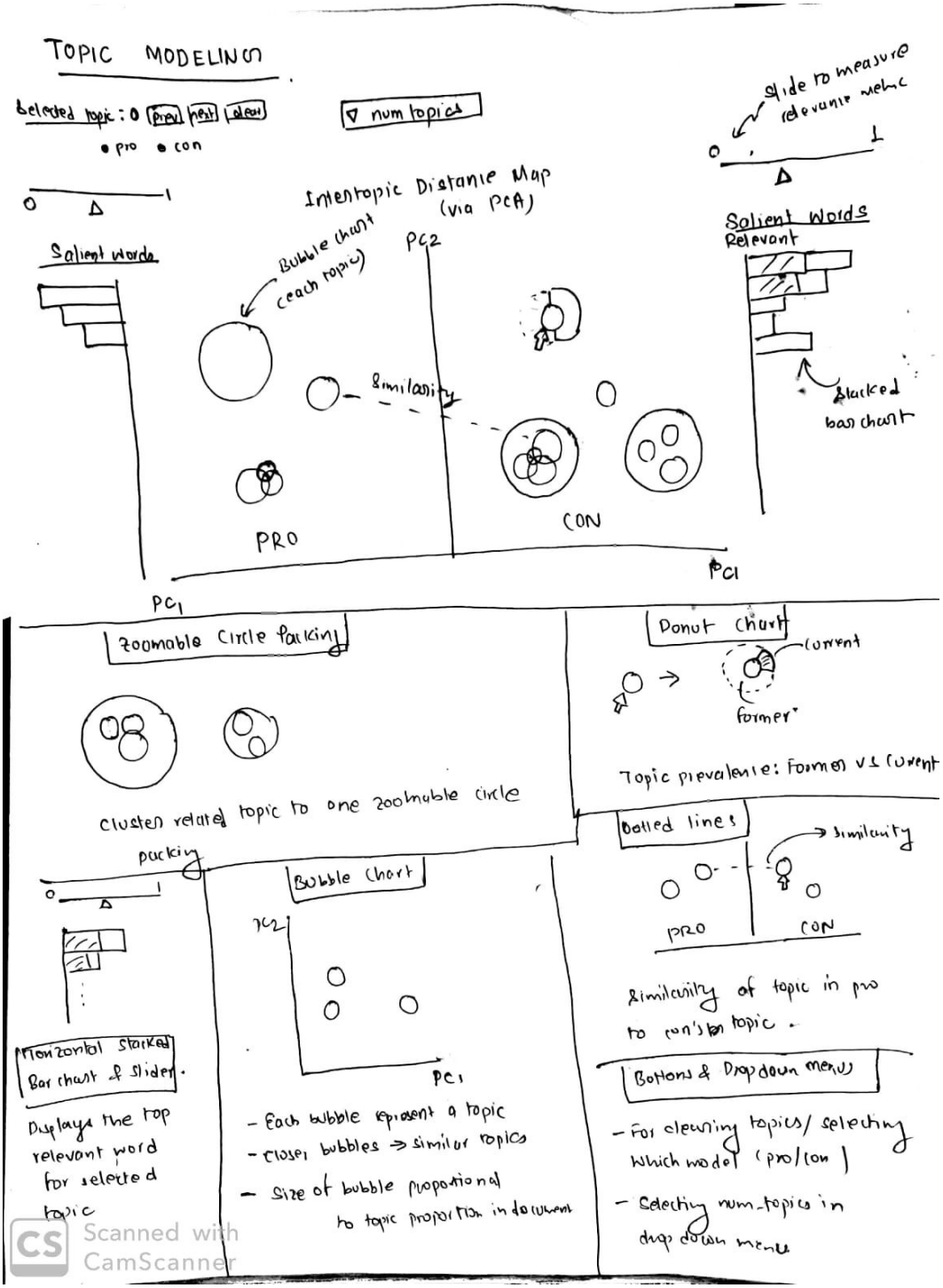

Here we broadened the topic by taking professions and working locations into account. However, it was hard to implement some of the ideas because of data availability, so we gave up on this.

Then this was what we came up with the following:



There were lots of interesting things going on in here, however it would be too difficult for the user to understand everything.

So we finally came up with the following design for the topic modeling, which is much more cleaner, and effective in conveying the purpose of our project:



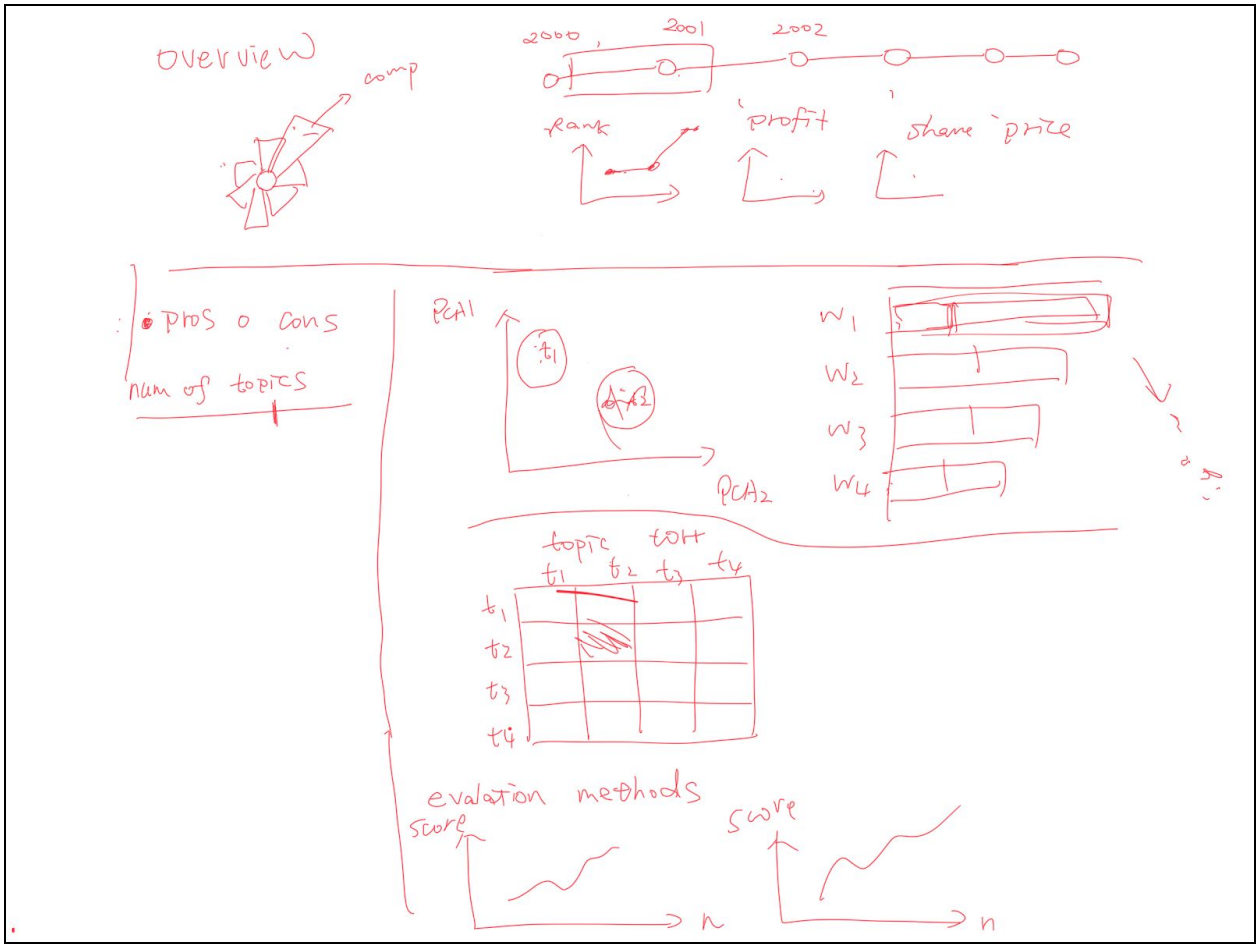
For this visualization we will basically be using different visual encodings like Zoomable Circle Packing, Stacked bar charts, Bubble Charts and Donut Charts. We will also use various controllers like buttons, radio button, text box, drag down menu and slider. We will also enable interaction between various components.

So, the main aspect of this visualization is to show the several topics in our models for pro and con, to one spatial space. We will utilize the spatial position attribute to cluster similar topics together. This will basically be a bubble chart with each bubble being one single topic, with the proximity between the bubbles representing similarity between those topics, while distant bubbles are very different topics. And with the size of bubble we aim to encode the proportion of this particular topic in the whole corpus. We will also use zoomable circle packing to cluster similar topics together, in order to remove the clutter and make our visualizations much more cleaner. On clicking any of the circle packing (which contains clusters of other topics), it will zoom in to display to show topics within it, and on clicking any of these circles will display the most relevant words for that topic, adjusted using the slider on the side for relevancy metric. Also on clicking one circle we will highlight only that circle and gray out all other circles, and also highlight the similar topic in another model ie. if we click on food related topic on “pro” model we aim to connect it to food related topic on “con” model using dotted line and also display weight on it. Also on clicking on any topic will expand it to a donut chart showing the proportion of this topic among former and current employees, helping us identify where this topic is more prevalent in, either the former or current employee. Similarly buttons, drag down menus, radio buttons will be added for control over what we will display.

## ● Design Evolution:-

There are two main aspects we want to present in our project - general information of companies and topic modeling. In general information section, some financial factors that indirectly affect employees’ satisfactions are included such as companies’ profits, revenues and share prices. In topic modeling section, we will show topics from pros model and cons model respectively, as well as key words in each topic.
Here are three alternative design sketches: -

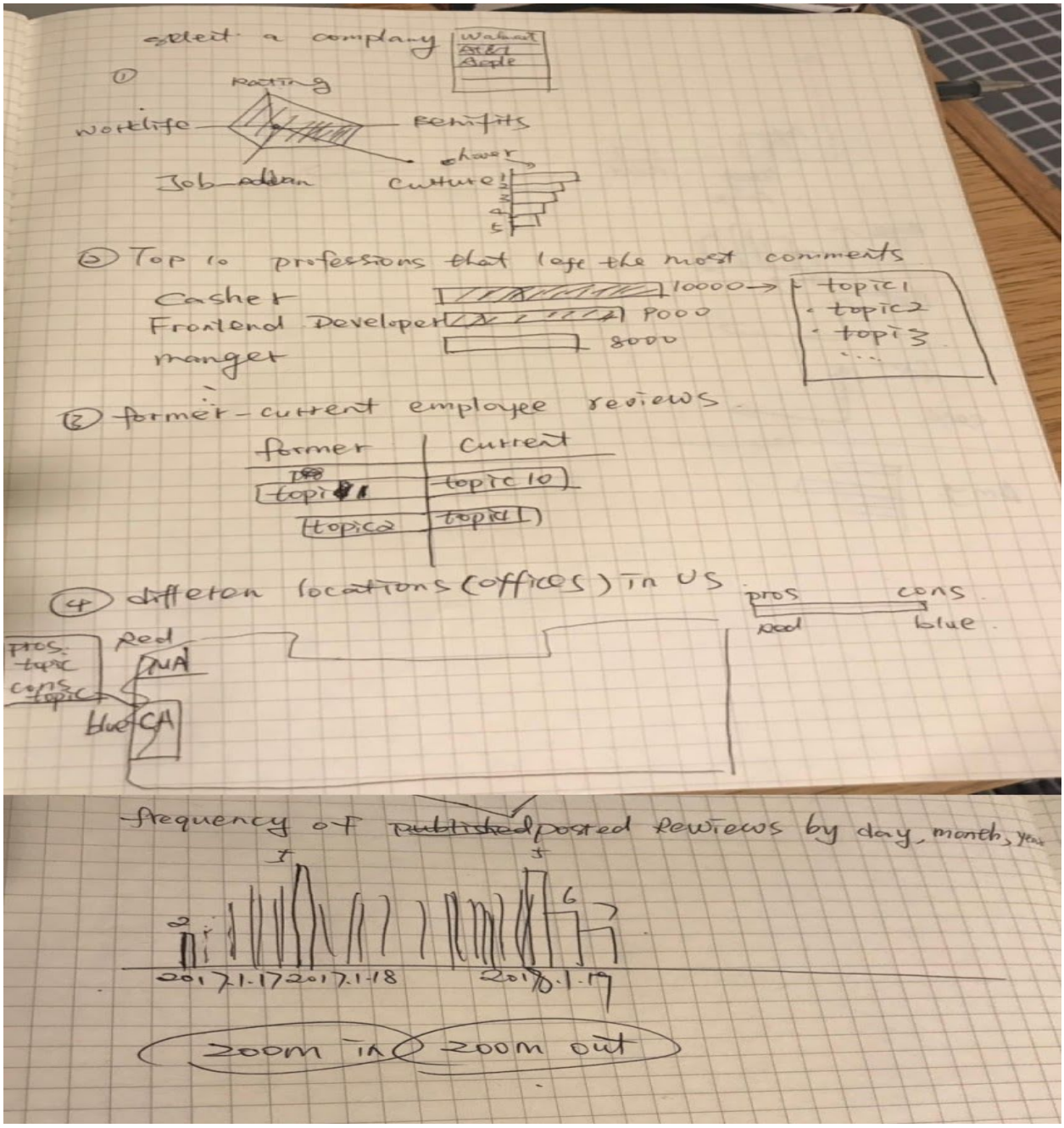
a. Vis1




Advantages: Basically it shows all the information we want to display

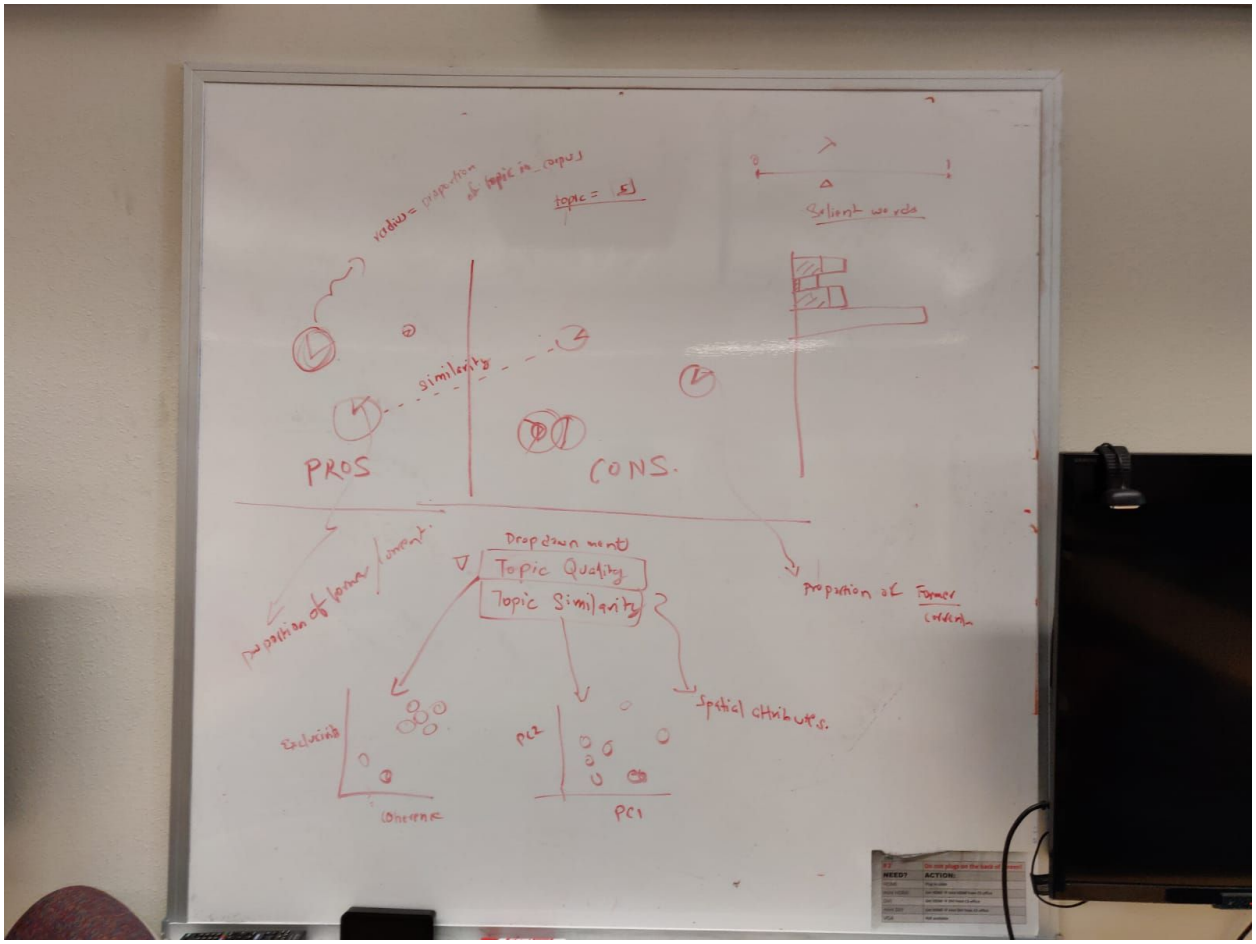
Disadvantages: There is not enough time series data by year.

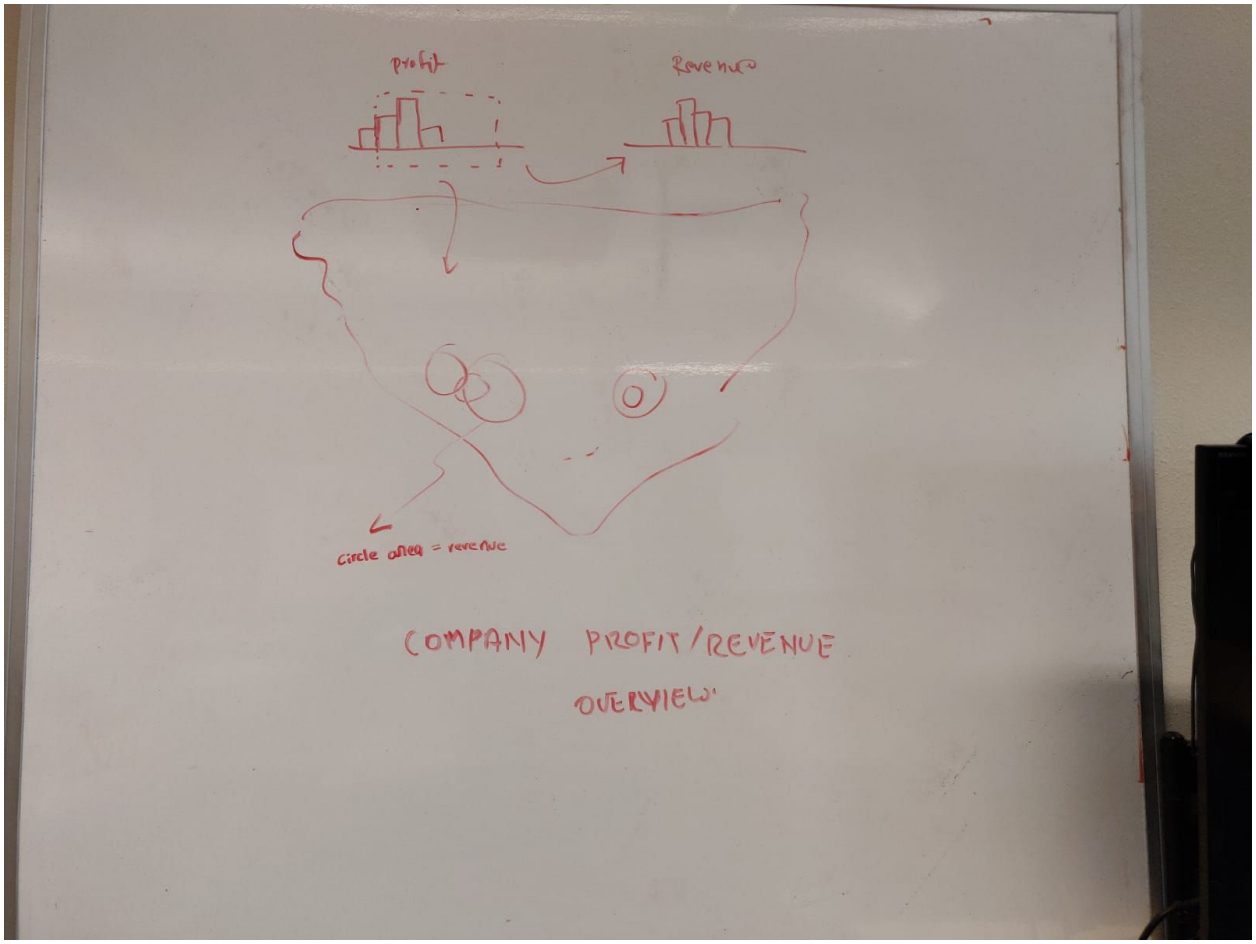
b. Vis2: -

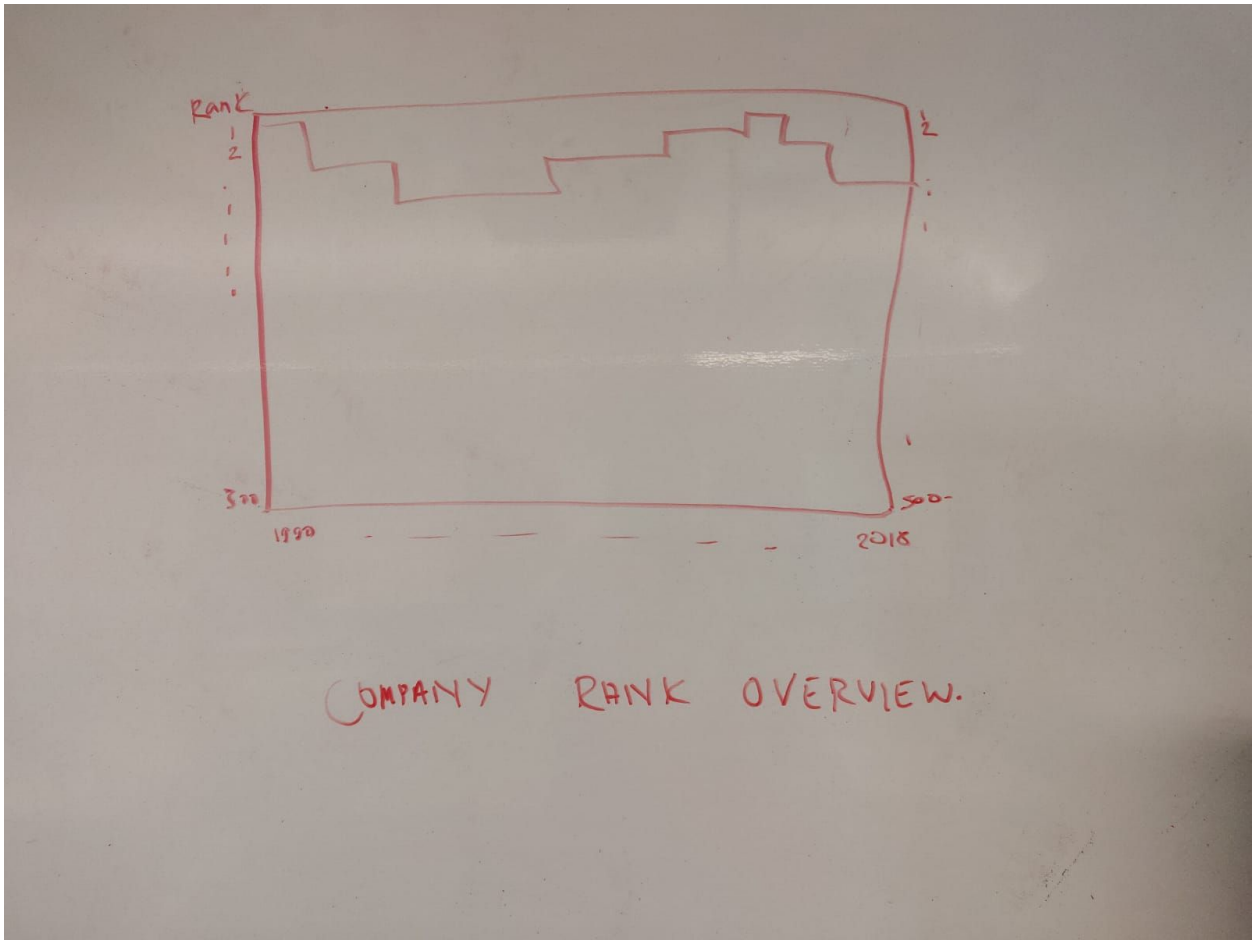



Advantages: broaden the topic by taking professions and working locations into account
Disadvantages: it is hard to implement some of the ideas because of data accessibility.

c. Vis3: -

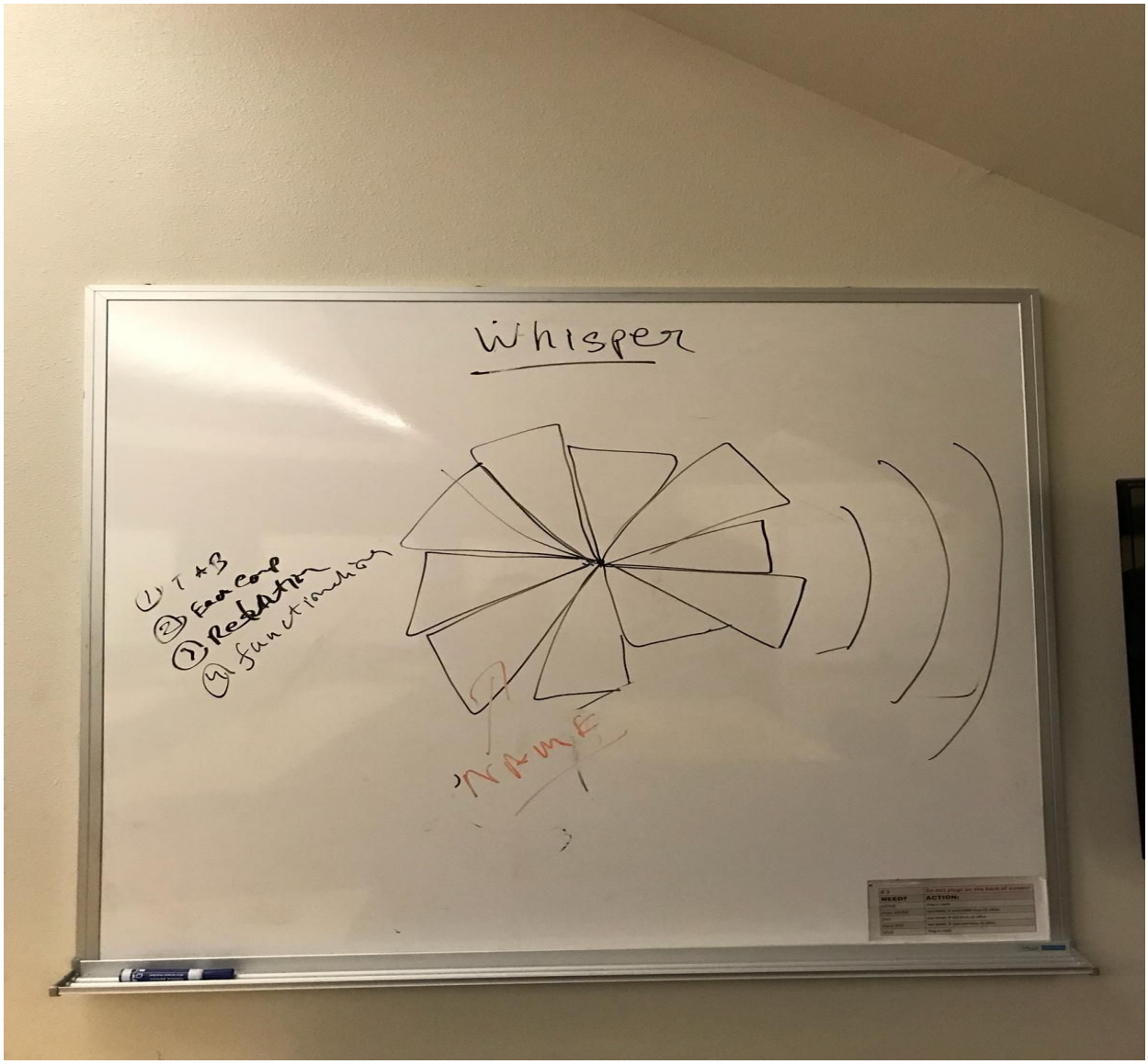

radius = proportion of topic in corpus
topic = .5
Salient words
similarity
PROS
CONS.
Dropdown menu
Topic Quality
Topic Similarity
proportion of former comment
proportion of former comment
Spatial attributes
PC2
PC1
coherence


profit
Revenue
circle area = revenue
COMPANY PROFIT/REVENUE
OVERVIEW

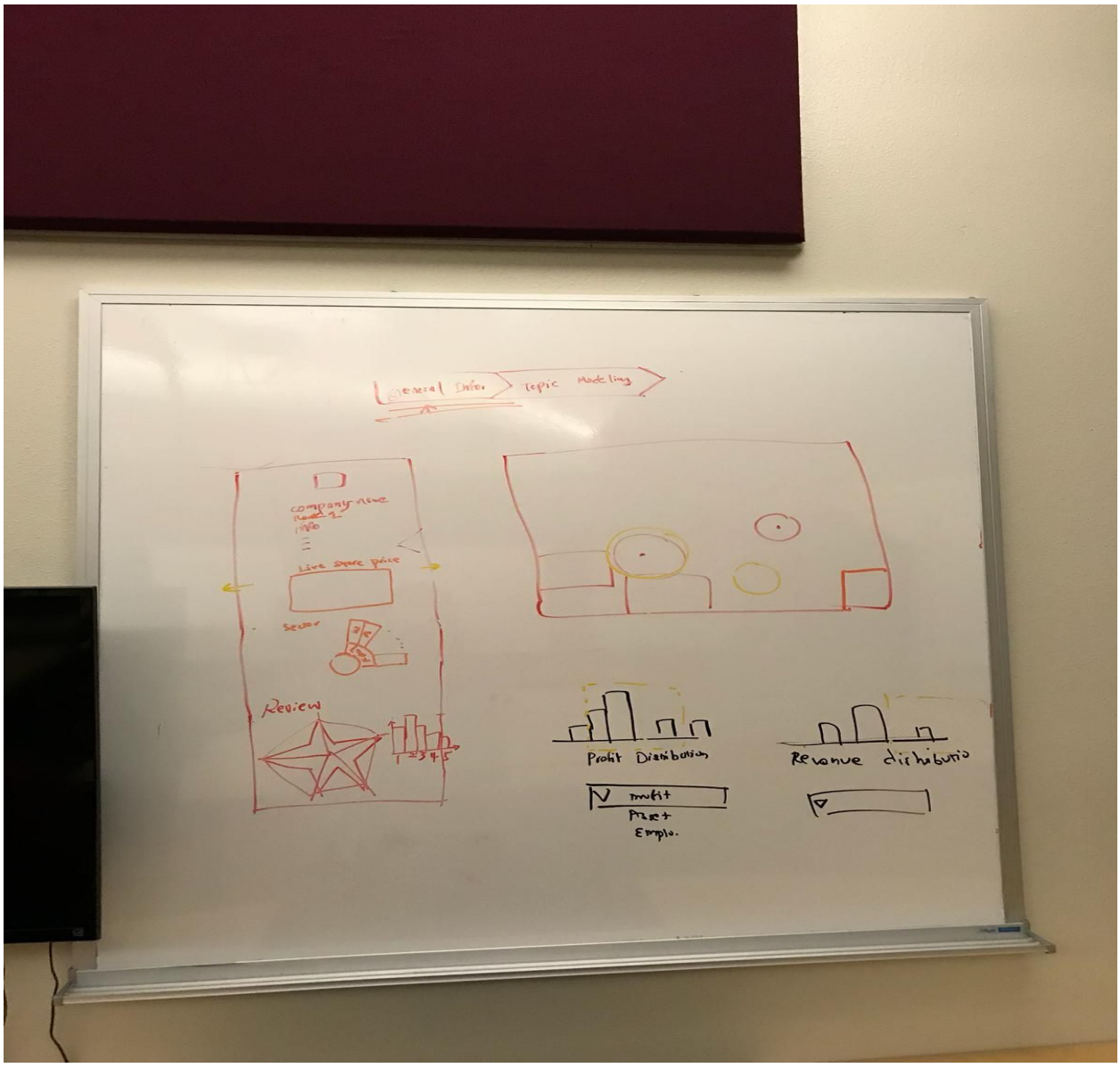



Advantages:  some interesting and diverse charts
Disadvantages: a little bit weak connection in two aspects.

# Final design

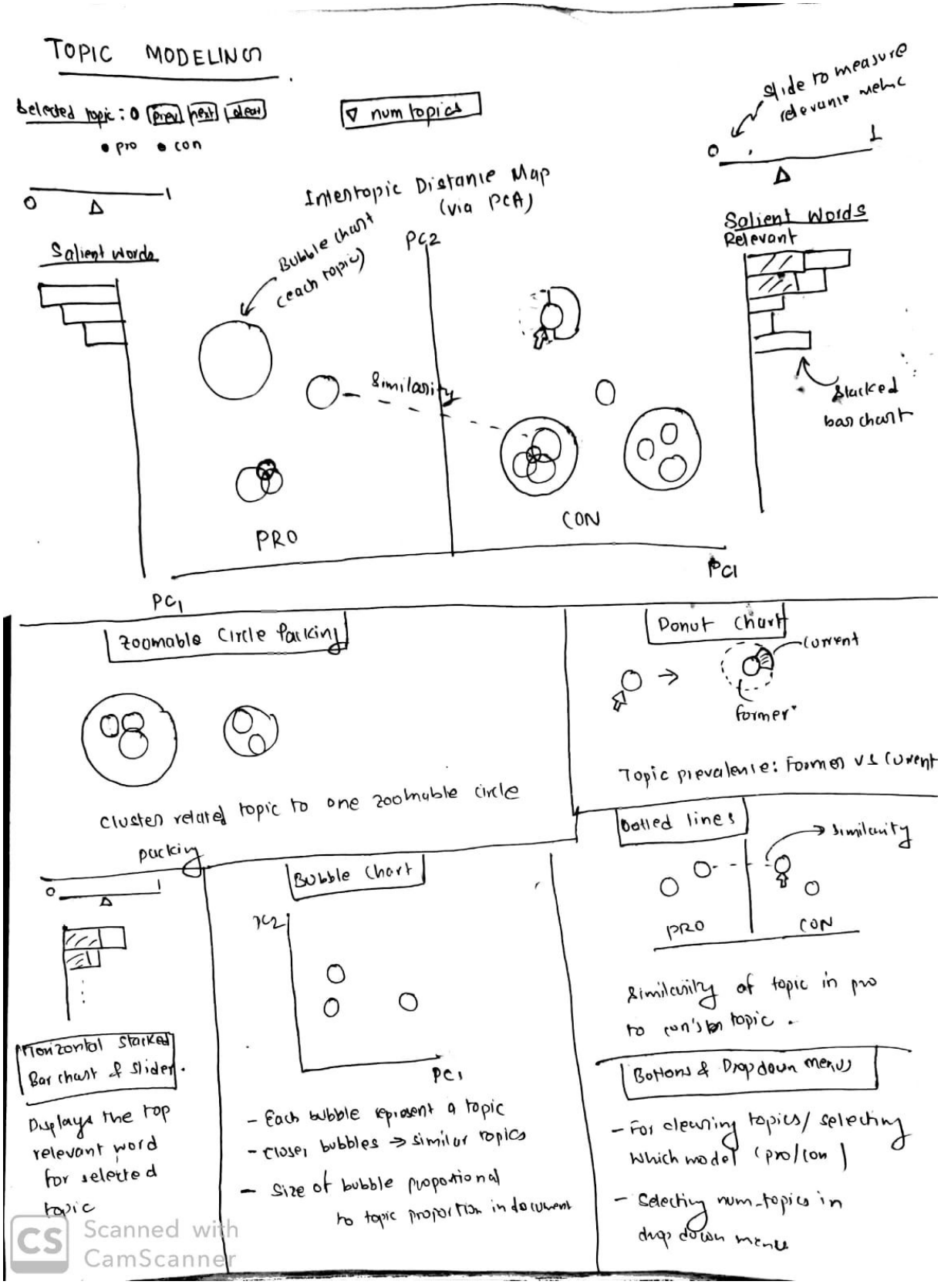



The design will be the one to load at first as the project loads up. The radial chart will load slowly showing the project it loading, with each part loading in few milliseconds, as the whole chart completes and the circle is complete, the Whisper project loads up and the signal three bars show up showing a signal. Then the whole chart does of and the general information tag comes up on the screen.




In this view, we will show the geographical distribution of Fortune 50 companies in the US. Circle size on the map encodes company's revenue. On the left, we will use line chart to show daily share price, and a hierarchical zoomable sunburst diagram to show the sector and industry which each company belongs to. Furthermore, a radar chart is designed to show average scores of five covariates like work life, benefits(1-5). When hovering on one covariate, a histogram that shows the frequency of the covariate from one score to five scores.

**Topic Modeling:**



**Topic Modeling:**

For this visualization we will basically be using different visual encodings like Zoomable Circle Packing, Stacked bar charts, Bubble Charts and Donut Charts. We will also use various controllers like buttons, radio button, text box, drag down menu and slider. We will also enable interaction between various components.

So, the main aspect of this visualization is to show the several topics in our models for pro and con, to one spatial space. We will utilize the spatial position attribute to cluster similar topics together. This will basically be a bubble chart with each bubble being one single topic, with the proximity between the bubbles representing similarity between those topics, while distant bubbles are very different topics. And with the size of bubble we aim to encode the proportion of this particular topic in the whole corpus. We will also use zoomable circle packing to cluster similar topics together, in order to remove the clutter and make our visualizations much more cleaner. On clicking any of the circle packing (which contains clusters of other topics), it will zoom in to display to show topics within it, and on clicking any of these circles will display the most relevant words for that topic, adjusted using the slider on the side for relevancy metric. Also on clicking one circle we will highlight only that circle and gray out all other circles, and also highlight the similar topic in another model ie. if we click on food related topic on “pro” model we aim to connect it to food related topic on “con” model using dotted line and also display weight on it. Also on clicking on any topic will expand it to a donut chart showing the proportion of this topic among former and current employees, helping us identify where this topic is more prevalent in, either the former or current employee. Similarly buttons, drag down menus, radio buttons will be added for control over what we will display.

# Must-Have Features

- Sunburst chart that shows revenues of companies
- US map that shows the geographical distribution of all the companies, circles whose size shows the revenue of companies, and interaction with company’s profile.
- Company profile including profits, revenue, share price, and 5 basic measurements of employee’s satisfaction such as work-life, rating, benefits and so on.
- Pros and cons topic circles that shows topic prevalence and similarity, and corresponding stacked bar chart that shows word distribution in documents.
- A bigger circle that clusters topics with high similarity , zoom in to see specific topics and zoom out to see overview.
- Pie charts to compare prevalence of topics among former employees with that among current employees.
- A track bar to select the number of topics for modeling.

# Optional Features

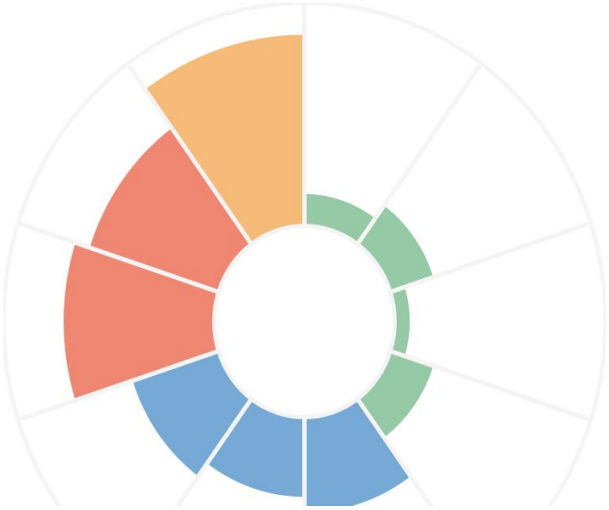
- Comparison of two companies' profiles
- Document Clustering
- Company Rank change over the years

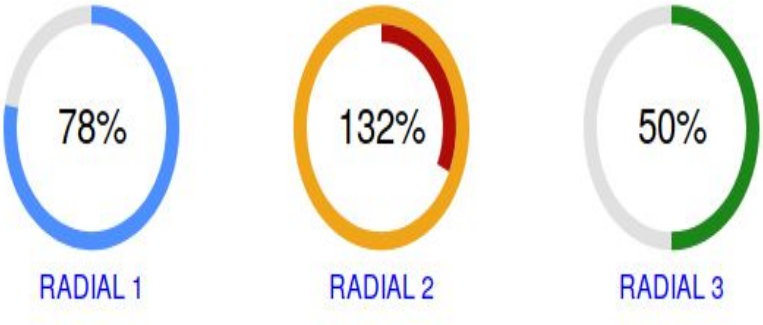
- **Implementation:-**
- Describe the intent and functionality of the interactive visualizations you implemented. Provide clear and well-referenced images showing the key design and interaction elements.

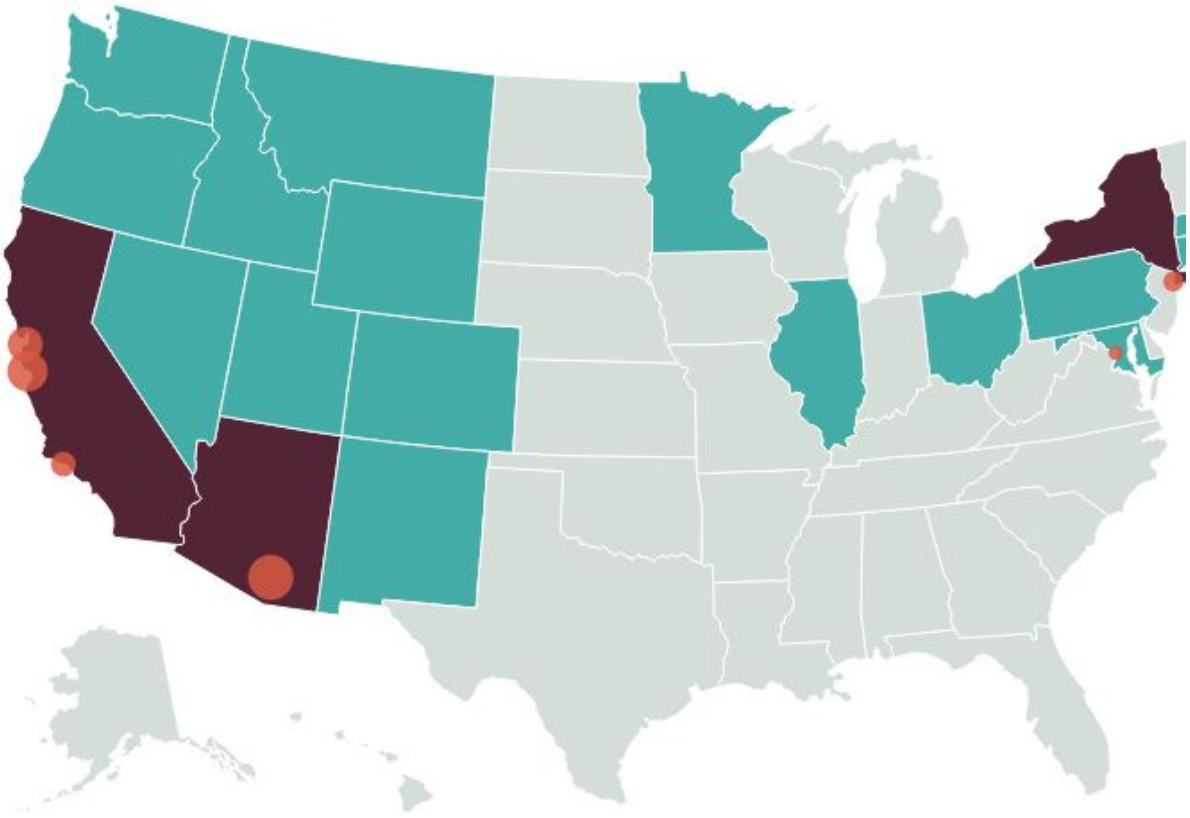
  Sunburst Chart - This chart will represent the sectors and the companies that belong to the specific sectors. The clicking will show the zoomable scatter plot.

-

Company map - As the project loads the aster plot shows up with the loading sign.







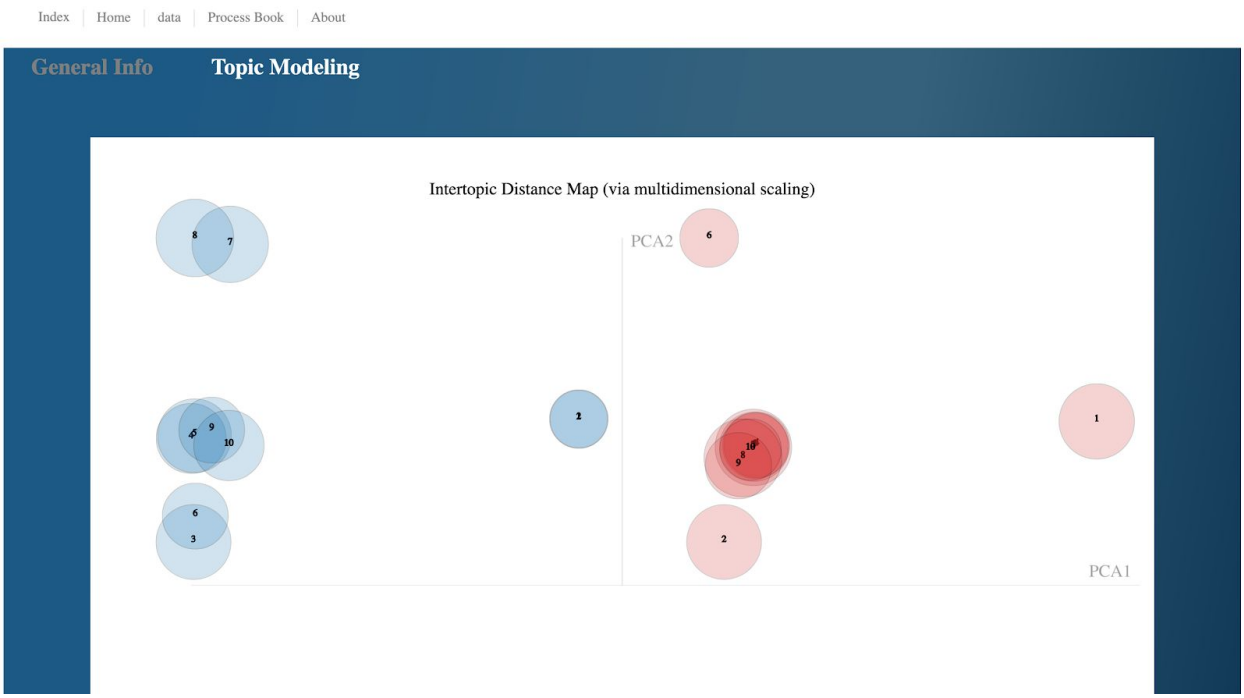
The next part is the map which will show the points as circles which will represent the companies and clicking on them will provide the general information.

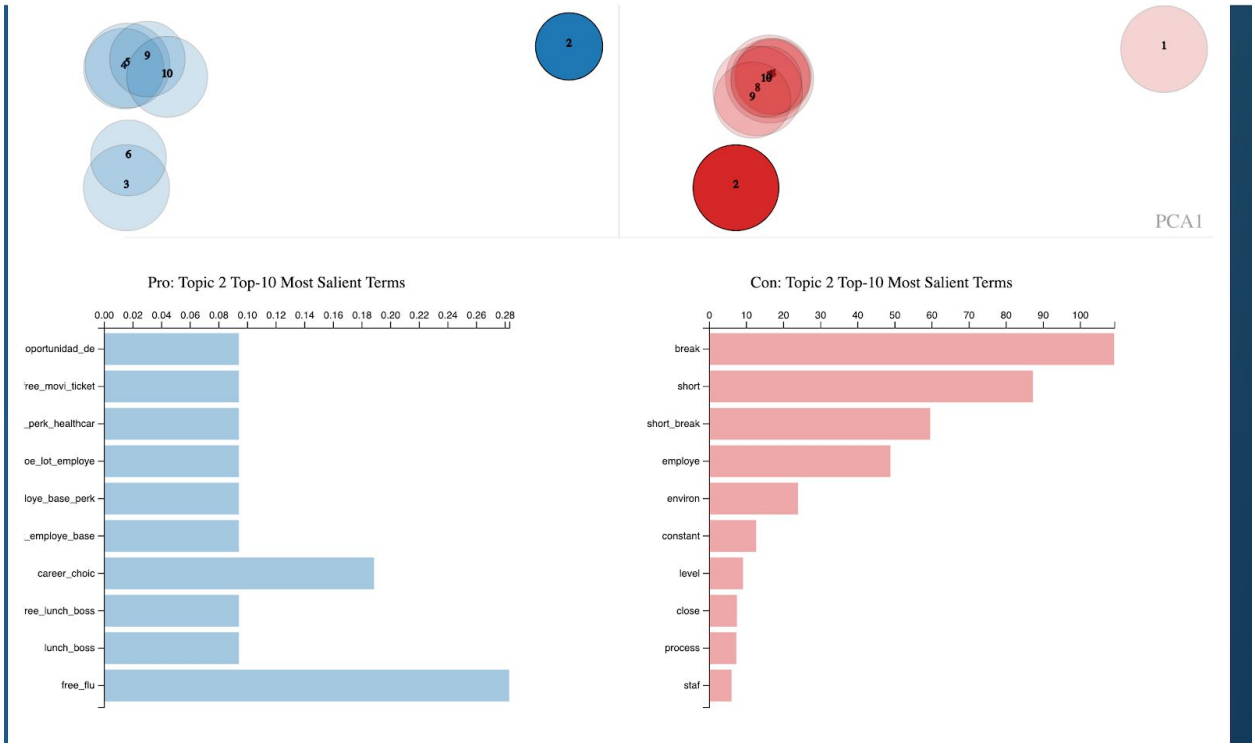
Company profile

We still have hierarchical sunburst (showing company section information) and live stock prices to put in there, however, for now we have just added a skeleton structure of how this section should look like. We still need to make this visually appealing.


Topic modeling

We have created a bubble chart, where the size of bubble is proportional to Topic Proportion. Also we have created barchart for the most relevant terms. So, when a topic is clicked, the most relevant terms barchart shows up in the bar graph panel, as can be seen in the following pictures. The first figure shows pro and con models in two PCA components. After clicking a bubble/circle, bubbles with the same topic in pro and con models will be highlighted and the corresponding terms with high frequency will pop up below the models. But we realize that bubbles overlap together. In the next step, we will add an option box for selecting topics.




We still need to implement donut chart for each individual topic showing the proportion of the topic that came from former or current employee. Similarly, we also need to implement Removable Circle Packing to cluster similar topic together. We also need to

show the connection line connecting topics between “pro” and “con” models. Similarly, we also need to show the overlapping barchart, we have one bar already, we need another overlapping bar showing the total term frequency over all the company, we haven’t implemented that yet for now.

- ## Evaluation:-

  What did you learn about the data by using your visualizations? How did you answer your questions? How well does your visualization work, and how could you further improve it?

  The data is in a very raw format, and needs a lot of preprocessing to have it in a form usable in d3. So data handling part took a lot of time. So we had to do a lot of exploratory data analysis to look delve deeper into the data.

  We still haven’t completed the visualization, but from what we have now, we are able to extract some useful topics that shows which topics are more prominent in “pros” and “cons” model, signifying, what topics make employees satisfied and dissatisfied. We are still in the process of what topic the “Former” and the “Current” employee are talking about more. We are hopeful that this question will be answered once our visualization is complete. Similarly, from what we have now, we can clearly extract some topics like “free foods”, “free lunches” and so on that employees have written in the pros section. Similarly, long break hours, good location, flexible schedule, good management, better pay and incentives are other factors that has been shown to satisfy the employees the most. However, on the con side, employees tend to leave negative feedback regarding, bad schedules, work loads, management issues, short breaks and so on. So, if companies can work on this aspect their employees would be more satisfied.

  Yeah for now the visualizations are working pretty well. However, there are still issues regarding whether we can fit everything into one space or not. If we have to divide up the location, we might not be able to convey all the stories as fluently as we are supposed to. So, that is still the issue, however, we should get it sorted. We should still work on the interactivity, and also the design part. That is our next step. And also there are other elements we need to add, which is supposed to be done in the next phase.