COMP7507 Visualization and visual analytics

Project Assignment

Choose Your Dream University

Part 1: Overview of Visualizations

Part 2: New Insights

Part 3: Difficulties

Part 4: Different Methods and Justification

Part 5: Future Expectation

Part 6: Limitation of Existing Tools

Part 7: Contributions

Group by:

Zhou Shuo 3035561510 (leader)

Liu Mingshan 3035562198

Du Ningxin 3035562095

Ye Jiantong 3035562239

Abstract

Studying abroad has become the choice of more and more people. This has become a hot topic around us. Usually, we have to search for information about each university from different websites one by one. How to visually compare the differences and diversity of the world's top universities has become our research interest. We are trying to use the visualization tools to analyze and present data on all aspects of the user's consideration when they choose an university for postgraduate degree. It provides users with simplified and intuitive information, query and matching capabilities to provide users with the convenience of understanding the advantages of the university.

Figure1 Home page

Part 1: Overview of Visualizations

Purpose:

The main purpose of this project is to allow students to better select universities for postgraduate courses by analyzing and visualizing data. The analyzed data is divided into 6 parts.

Visualizations:

1. Personal information & Radar chart of five application standard

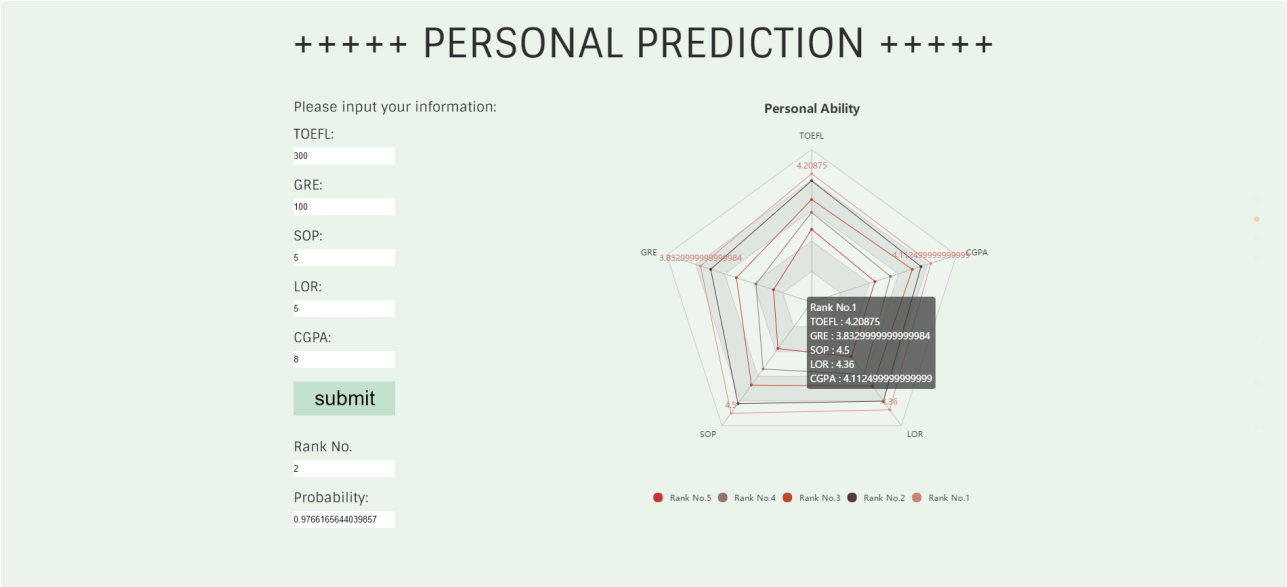

Figure2 Personal information & Radar chart of five application standard

This data set contains a total of 500 student data including GRE Scores (out of 340), TOEFL Scores (out of 120), Statement of Purpose (SOP) and Letter of Recommendation (LOR) Strength (out of 5) and Undergraduate GPA (CGPA) (out of 10).

In this section, we use the radar chart to present data. The radar chart used here is also called Debra. A traditional radar chart is considered to be a graph showing multidimensional (more than 4 dimensions) data. It maps the data volume of multiple dimensions onto the axes. These axes start at the same center point, usually ending at the edge of the circle. It is called a radar chart that connect the same set of points by using a line. It can display multidimensional data, but the angle between the relative position of the point and the axis is not any amount of information. The area enclosed by the radar chart can show some amount of information when the axis is set properly.

When using radar charts, people often artificially combine multiple axes into a single metric, such like a percentage, to make the chart easier to understand. The radar chart can also show the weight of each variable in the data set, which is very suitable for displaying performance data. But too many variables can cause readability to drop. So the number of variables need to be controlled to keep the radar chart simple and clear.

The radar chart on the right shows the average of the grades of different parameters of graduate students admitted to different ranks of universities. Students can understand which aspects of they need to improve to reach which level will be better accepted by the dream school. As shown in the radar chart, universities with a rating of 5 (the best universities)

tend to enroll the students whose TOEFL scores is about 113.67 and their GRE Scores is about 328.33.Their Statement of Purpose is about 4.5 and Letter Of Recommendation is near 4.36. And their undergraduate GPA is around 9.29.

In terms of vertical comparison, we found that students who have been admitted to different universities have a large difference in GRE. Therefore, students can choose to improve in this aspect if other grades are basically up to standard.

This data set has multiple parameters. What we want to know is the relationship between the school rating and the individual's various kinds of grades, that is, to show multiple data mapping to one type of data. So we use this type of graph to analyze this part of the data. This graph allows students to understand the goals they need to achieve in order to successfully apply the school they want to go when they have not started to prepare for the application.

The input boxes on the left provides the user with a query channel. Students can enter their own grades to check applying for which school they will have a higher probability to admit and find out what is the success rate.

The radar chart is designed to allow students to set goals by using them as a reference. The input box function is used by students to learn applying for which level of school is more reasonable when they know their own grades.

2. Word cloud of AC/FC

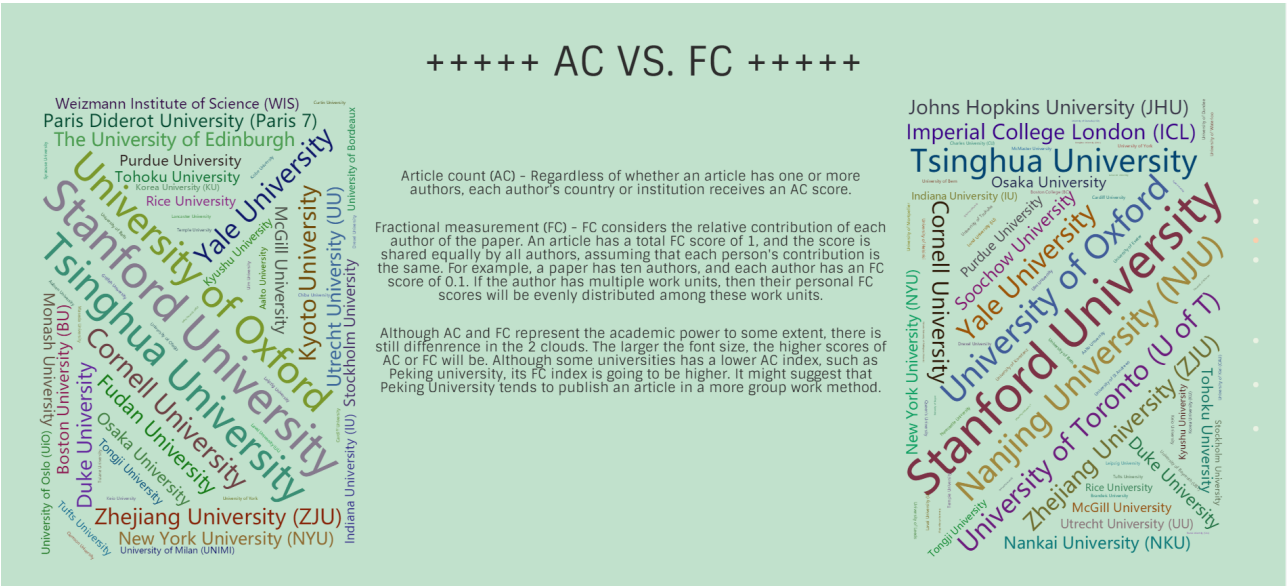

Figure3 Word cloud of AC/FC

A word cloud is a visual representation of textual data. It displayed a large amount of text data by a cloud-like color graphic which is composed of words. It's usually used to describe keyword metadata (labels) on a website, or to visualize free-form text. The importance of each word is displayed in font size or color. Using word cloud to visualize the data can quickly perceive the most prominent text and locate the relatively prominent parts of the alphabetical text.

The data set used in the second part contains three parameters: school name, article count (AC), fractional count (FC) which are presented in the form of word clouds. AC represents the number of academic articles published by the school. FC reflects the number of people who participated in the published literature research. The school with the larger font size has a larger AC or FC value. Through the word cloud of AC on the left, we can learn the academic level of the school. Through the word cloud of FC on the right, we can learn whether the academic research of the school used to be an individual research or a group research. Students can choose the school according to their preference.

Word clouds can be used to present a larger amount of data than a histogram. It displays text more intuitively and maps more categorical fields on the text style.

Select it to visualize this part of the data, so that we can see the academic level of those universities more intuitively and clearly. The relatively low academic level is relatively less prominent because of the smaller font size.

3. World map of Top 100 universities & Line chart

Figure4 World map of Top 100 universities & Line chart

1)World map: Distribution and total score of universities

This part use scattered dot map to visualize the data set. Scattered Dot Map, also known as Dot Distribution Map and Dot Density Map, is the method for representing the distribution of data in geospatial by plotting points of the same size on a geographic background.

We use one-to-one type of Dot Map, that is, a point represents only one data or object. Because the position of the point corresponds to only one data, so it must be ensured that the point is in the correct spatial location.

Dot Map is an ideal way to observe the distribution of objects in geospatial space. Clusters of points formed on the map can display some data patterns. With the help of dot map, it is convenient to grasp the overall distribution of data.

This visualization graph is used because it can be seen intuitively where the top universities are concentrated. Through the map, we can find out the top schools gather in the United States and Europe. It can be seen from the enlarged map that in the United States, schools are mainly concentrated in the west, and schools in the east are mainly located in the coastline. After zooming in on Europe region, we can see that the number of the top 100 schools in some countries is increasing over time, such as Netherlands, which indicates that the quality of education in the country is constantly improving. By analyzing these data, we can better understand in which country, and in which part of the country, the quality of education will be higher. Having an overall and detailed view will help students apply to universities in areas with relatively high quality education.

2)Line chart: Five scores (Teaching, Research, Citation, Industry Income and National Outlook) of each university over nine years

A line chart is used to show the change of data over a continuous time interval or time span. It is characterized by a tendency to reflect things as they change over time or ordered categories.

The line chart can show whether the data is incremental or decremented. The rate of increase and decrease, the law of increase and decrease (period, helicity, etc.), peak and other characteristics can be clearly reflected. Therefore, line charts are often used to analyze the trend of data over time, and can also be used to analyze the interaction and interaction of multiple sets of data over time.

Compared with the histogram, the line chart is mainly used for multiple trends that can be integrated into one graph. In this way, users can have a more specific comparison among different lines or attributes.

In this project, we have used the line chart to visually see the different component and its influence for a the overall university score. And we have found that students can selectively view the trend of some of the data according to their emphasis on the quality of education, academic level, industry level or other elements.

3) Combination of world map and line chart

The data set used in the whole section 3 contains the attribute including the times, location, university, and the world ranking. The points displayed on the map are locations of the top 100 universities in the world. The color of the dots represents the level of the comprehensive score of the world university rankings. The higher the score, the more the color of the dots tends to be red. The lower the score, the more the color tends to be blue. As indicated on the figure, the position of red dots on the map will change by the timeline below. When the red dot reaches a certain year, the data displayed on the map is the world university ranking data for that year. When the mouse selects a point on the map, a small tag appears showing the university's name, latitude, longitude and composite score. When the

mouse selects a certain area, the name of the country will appear, and users can also select a corresponding university in this country by the drop-down menu. After that, the line graph on the right side will also show the trend of the evaluation data of the school over time, such as education level and academic level. Users can zoom in and out of the map by scrolling the mouse wheel to see more clearly which country a university belongs to. Users can un-display a line (that is, a group of data) by clicking the legend below the line graph.

4.World map of countries that contain Top 100 universities & Line chart

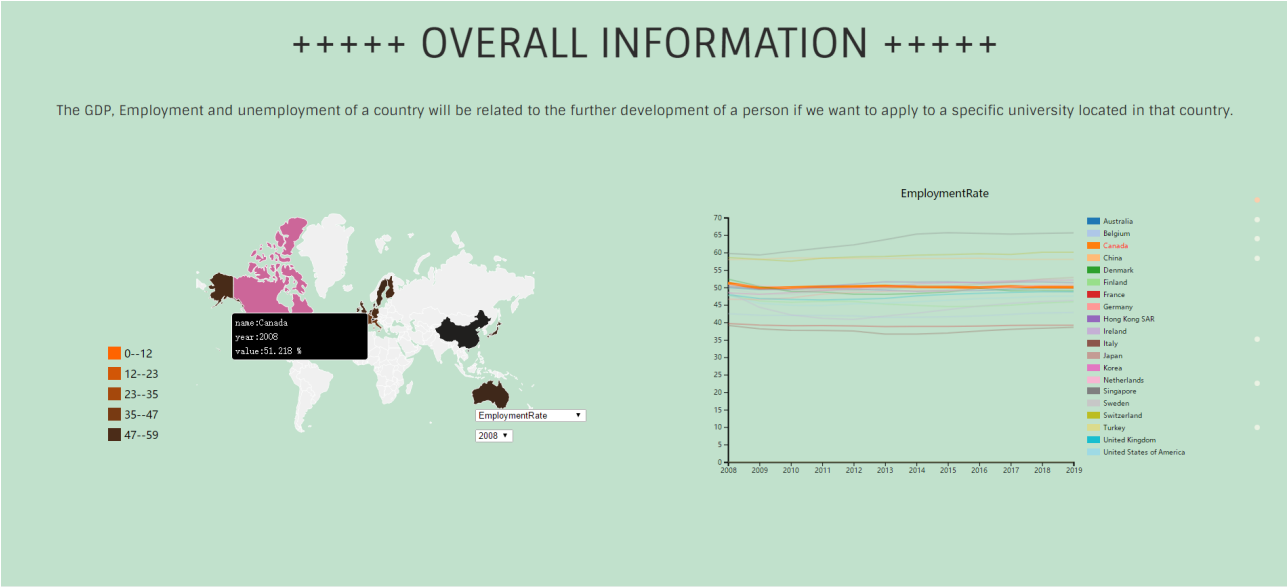

Figure5 World map of countries that contain Top 100 universities & Line chart

The data set contains employment, unemployment, population and gross domestic product in countries with 100 schools worldwide. There are two option boxes in the diagram, the first provides the ability to select these four elements, and the second option box provides the ability to select the time.

1) World map: Four aspects (Population, GDP per person, Employment, Unemployment) over nine years

The first image is the Choropleth Map, which is a map that uses visual symbols (usually colors, shadows, or different dense halos) to represent the distribution of a range of values on a map partition.

Within a number of small zoning units (administrative zoning or other zoning units) throughout the mapping area, grading according to the number (relative) indicators of each

zoning. And use the corresponding color level or different dense halo to reflect the concentration of the phenomenon in each area or the distribution of development level. It is most commonly used for the visualization of election and census data, which is based on geographic regions such as provinces and cities. It is usually used color level to represent.

Here we use the geographical region of the country to reflect the distribution of unemployment, employment, population and GDP in various places, and use a single color gradient to rank.

When the mouse moves over an area, a prompt box appears showing the country name, time, and value of the corresponding element. The lightness of the color on the map shows the size of the corresponding element of the region.

2)Line Chart: Four aspects (Population, GDP per person, Employment, Unemployment) over nine years

The line chart on the right shows the trend of an element over time. Each line represents a country. Use color to make a distinction between countries. When the mouse stays on a country label next to the line chart, the country's geographic location on the map and the corresponding polyline on the line chart are highlighted.

The line chart mainly shows the development trend of an element, so as to observe the development prospects of a certain area. For example, users can observe the trend of employment rate and unemployment rate to consider the future opportunities to get a job if they want to study and stay in this country after completing their studies. Users can also learn the economic development of the country by observing the trends in GDP because it’s related to their future income closely. When a university locates in a more potential and promising country or area, students will get more opportunities for their future development.

5. HeatMap of Number of Top 100 universities each country having in relation with GDP per person, Employment and Unemployment

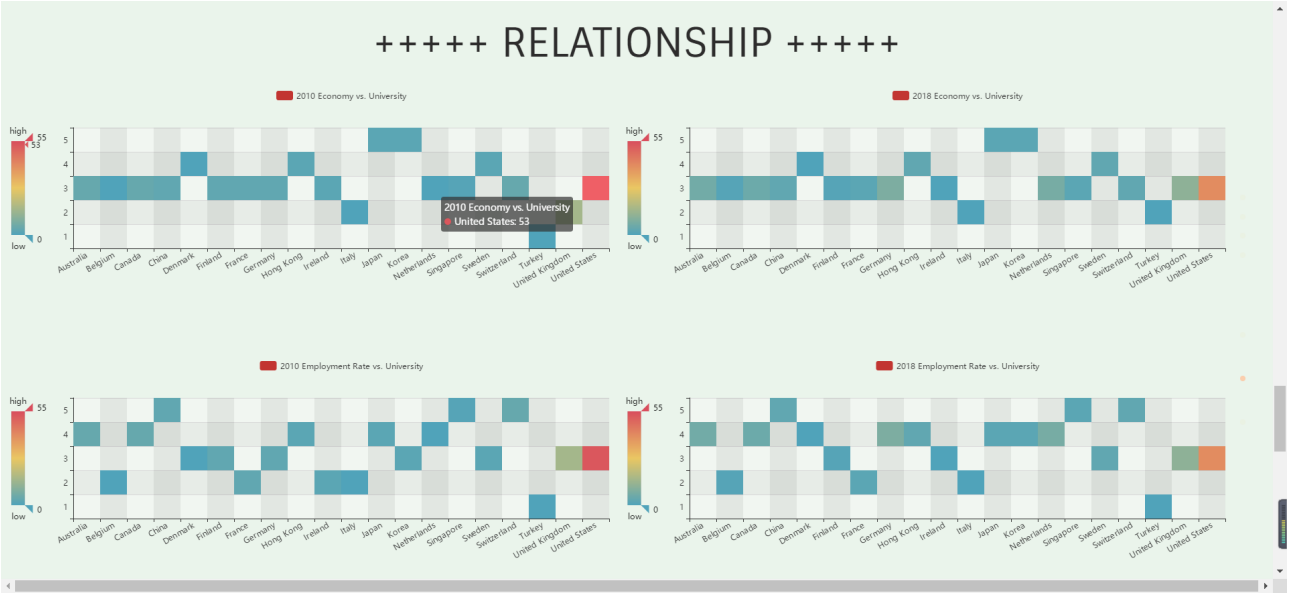

Figure 6 HeatMap of Number of Top 100 universities each country having in relation with GDP per person, Employment and Unemployment

In this section, we can use the heat map to link the GDP, unemployment, and employment rates of each country to the total number of schools in the country. Through the comparison of 2010 and 2018 data, the potential links between GDP, unemployment rate, employment rate and the number of universities are found.

1）Economy

On the economic front, the economic development level of these countries in the past nine years has been relatively stable and there has not been much fluctuation. At the same time, the number of schools in each country is relatively stable.

2）Employment

In terms of employment rates, the employment rates in China, Singapore and Switzerland are always high, indicating that there is a greater likelihood of getting jobs in these three countries. It can also be observed from the figure that the employment rate in Germany has increased significantly and the total number of top100 universities has increased by five. It can be speculated that in the development of Germany in the past nine years, the improvement of education level may have a certain improvement effect on the employment rate. The employment rate in the United States and the United Kingdom is medium, but there are many top100 universities. Perhaps it implies that the trend in the two countries is not looking for a job locally.

3）Unemployment

Observed by the images, the unemployment rate in Singapore and the Netherlands has improved significantly and the number of top100 universities has increased. This shows that the development of education may provide better talents to participate in social work.

Overall, the decline in the number of schools in the United States and the United Kingdom indicates that education levels in other countries are improving. This also implies that the development of education in the world tends to be globalized and balanced.

6. Interactive Bubble Chart of AC/FC of Top 500 Universities in four disciplines

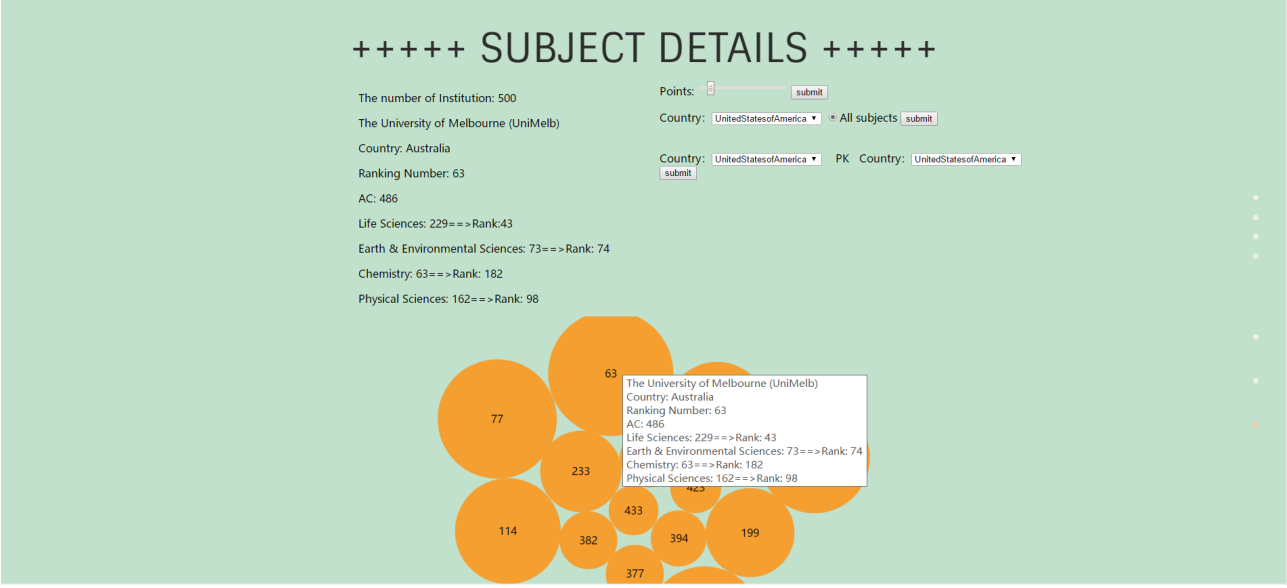

Figure7 Interactive Bubble Chart of AC/FC of Top 500 Universities in four disciplines

The data set contains the AC and total values for each of the top 500 universities ranked by AC value.

The chart is called interactive bubble chart. Users can generate different bubble charts grouped by countries or subjects through choosing different attributes and their response combination. There are three functions.

The first function is that people can drag the "points" slider to the right to change the number of bubbles that represent the universities in the generated image. Users can click "submit" to generate the image. Each bubble represents an university. The number on the

bubble represents the rank of the school. The size of bubble represents the AC values. When the mouse stays on one bubble, a label containing information such as the AC value of the dominant subject of the university appears. When one of the bubbles is clicked, the information bar on the left side also outputs the corresponding data of the university.

The second function is that people can choose the country and all subjects. People can select the country to show the universities that belong to that country and can choose all subjects show the AC values of all the subjects instead of the total AC values of the school.

The third function is that users can choose two countries to compare the AC values of the schools in those two countries in order to compare their academic research level.

A bubble chart is used to visualize a data set with two to four dimensions. The first two dimensions are visualized as coordinates, the third as color and the fourth as size.

Using an interactive bubble chart here to visualize the data can be very straightforward to compare the level of academic research by comparing the size of the bubbles. Students can also get an accurate value by clicking on a bubble. There is a lot of fun because the bubbles can be dragged anywhere and will be recalculated with the weight of power coming from different attributes.

Part 2: New Insights

Interactive Charts

Use Pyecharts and Data-Driven Documents to do interactive charts rather than statistic charts. Interactive charts is a solution for intelligent business analysis, making static charts as dynamic as possible, thus improving the actual use value of the report. Interactive charts provide users with interactive features that allow users to enter or select values before running a chart to determine data and form. Users can use interactive charts to not only show or hide content in a report, but also to access other visualizations by clicking on the links. In short, interactive reports add user-operable features to static reports, making them interactive.

Charts are no longer displayed in static form and presentation becomes interactive by adding user operations such as dynamic sorting, filtering, and drilling in static charts.

Making Charts interactive is the need to solve the data analysis of end users. Interactive charts can provide business intelligence with the intelligence needed to analyze data. Users can actively control the rendering of data and make reports more vivid and intelligent.

In our charts, you can select which university you are interested in, then click “submit”, it will show the line chart of five scores of this university. If you want to know number of Top 100 universities is relevant to which attribute, you can select “Employment”, then it will show the different degree of employment each country in the world map and show the line chart of changes over nine years. If you want to know AC/FC in four disciplines of some countries, you can filter other countries.

Part 3: Difficulties

Data Processing

When we downloaded the original data, it was very messy and had many losing and uncorrected data. We needed to deal with the data first before doing the visualizations. For all the tables, we re-integrated the attributes to generate several new tables. For the missing data, we processed it with 0. The original dataset was very large, so we spent a few days to generate tables that we could apply well in our expected visualizations.

Interactive Bubble Chart

After deciding this topic, our group started to consider how to express information of the dataset. Our dataset includes a list of institutions. Every institution consists of four main subjects i.e. Life Science, Earth and Environment Science, Chemistry and Physical Science. There is an obvious proportion relationship, which is easy to be expressed in the pie chart. However, we have a list of universities, so it means that we would do a lot of pie charts. Moreover, it cannot show the rank of universities.

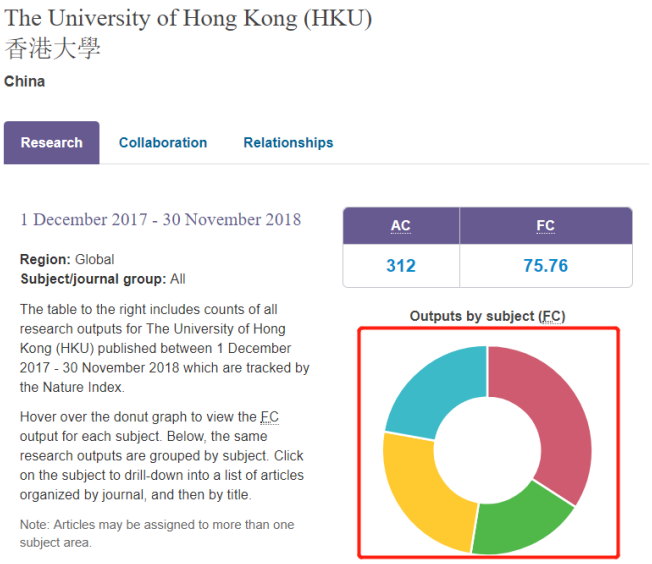

Figure8 pie chart example

When we did the literature review, we find several examples which could help us to complete our visualization. The first example is from a TV and films visualization. It utilizes a bubble chart to express information. The area of the circle represents the value. The color of the circle could show different kinds. The left of the button can select those objects which include the same field. It is suitable for our dataset to present that one university involves 4 subjects.

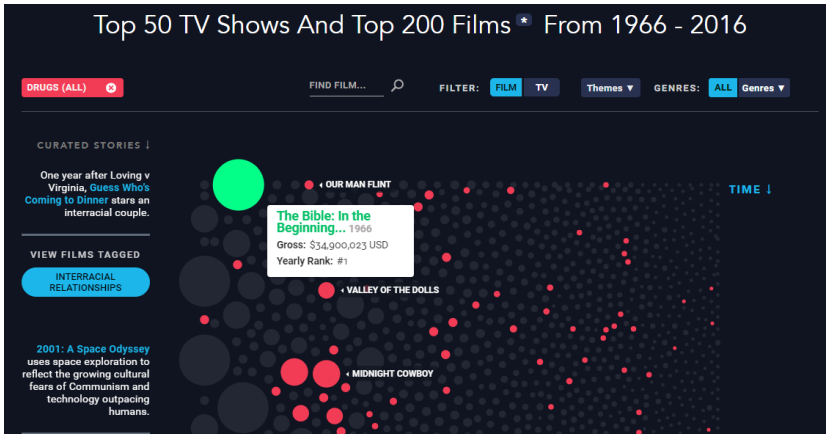

Figure9 static bubble chart

In order to implement this bubble chart, we also browse the D3 website. The interactive bubble chart attracts our views. The interactive bubble chart includes force system. Except for color, it is also able to use force to cluster elements. Moreover, interactive bubble charts are more interesting than static charts. Therefore, our group decided to achieve an interactive bubble chart to express our dataset.

However, D3 API is not friendly for a new user. Although the D3 website has plenty of examples, the different D3 version is incompatible. A lot of old version example cannot run in the current environment. Through try a lot of examples, we learned about the architecture of the interactive bubble chart and force system. As we final demonstration, it is a good way to show our dataset.

Part 4: Different Methods and Justification

Different Methods to Build Visualizations

1. Pyecharts

On the one hand, Echarts is an open source data visualization which is offered by Baidu. With good interactivity and ingenious graphic design, it can generate many excellent visualized images. On the other hand, python is an expressive language that is very well suited for data processing. When data analysis is associated with data visualization, Pyecharts is the best choice for us to generate some various visualized images to meet our different needs on data.

Pyecharts has many advantages for using. It has very simple API design, smoothly using and support chained calls. Pyecharts includes more than 30 common charts such as pie chart, line chart, bar chart, tree map and so on. It almost offer most basic charts that you need. Pyecharts has highly flexible configuration items which are very easy for you to match then generate beautiful charts. Pyecharts can also import data in bulk rather than import data one by on manually.

2. Data-Driven Documents

Data-Driven Documents is a JavaScript library for producing dynamic, interactive data visualizations in web browsers. It makes use of the widely implemented SVG , HTML5 , and CSS standards. Data-Driven Documents use scalable vector graphics or SVG format to render enlarged or reduced shapes, lines and fills without sacrificing quality. Data-Driven Documents is used on hundreds of thousands of websites. Some popular uses include creating interactive graphics for online news websites, information dashboards for viewing

data, and producing maps from GISmap making data. In addition, the exportable nature of SVG enables graphics created by D3 to be used in print publications.

The functional goal of D3 is to allow everyone to encapsulate a Javascript utility function that implements their own data for rendering ideas, allowing any node, placeholder in HTML to be instantly transformed into an icon or dynamically rendered content.

Different methods to build same data visualizations & Justifications

1. Use radar chart to present different application requirements that universities in five levels need.

At the beginning, we consider to use which chart to show requirements that each university need in five aspects(TOFEL, GRE, Statement of Purpose, Letter of Recommendation and Undergraduate GPA). But we find that if we just use a map of all universities and click to show its requirements, when users input a university name and then it outputs requirement of this university. Users can not compare their own information with universities that are almost at the same level in a general view. So we choose radar chart to show which level of university that you have more possibility to apply. Radar chart artificially combines multiple axes into a single metric to show average requirements of universities in this level need. Users can input their information, then we can compute which level of universities they are probable to apply.

2. Use word cloud to present AC/FC of a university.

Use Pyecharts to import all data of AC/FC and then generate the visualization. We tried to show AC/FC in a world map by using size of circle point of this university. But if two university are very close to each other in map, their circles will be overlapped and increase the difficulty of comparing AC/FC. Besides, it is not aesthetic if all circles crowd together because top 100 universities are mainly distributed in the United States and western Europe, especially area of western Europe is very small. For students they probably want to have a

overall view of which universities have higher AC/FC, so we choose word cloud to give a general view of AC/FC among all universities.

3. Use HeatMap to present relationships between number of Top 100 universities in each country and GDP per person, employment, unemployment.

At first, we only use two groups of world map and line chart to present relationships between number of universities in each country and other attributes. But in two sections, it is not very easy to compare relationships in two sections. So we thought about HeatMap which can better show relationships and this can connect two groups of visualizations more closely. Through the HeatMap, we can clearly know that stable economy, higher employment and lower unemployment are relative to number of Top 100 universities in countries.

Part 5: Future Expectation

Augmented-Reality

We would like to apply all our visualizations with Augmented-Reality and then generate a QR code. When users input their personal information of GRE, TOEFL and so on through our website on the mobile phone, it can generate a QR code which hide your personal information. Then if users scan a QR code, they can see 3D stereo images of all our visualizations in front of them. For example, in 3D world map, users can see separate dots which represent Top 100 universities. The 3D globe will rotate so that users can see more clearly and have intuitive feeling that each university locates where. It is more vivid and has more educational meanings if this visualizations used for students in geographical classes.

However, we do not have enough time and energy. Also, because of technology limitation among us, we think we can add this function in the future.

Part 6: Limitation of Existing Tools

Tableau

Tableau's simplicity and ease of use are explicit and this is the biggest feature of Tableau but this is exactly its limitation. Users don't need to be proficient in complex programming and statistical principles. They just need to drag data directly into the work log through some simple setting up to get the data visualization that you want, which allows even an unprofessional background person to create beautiful interactive charts for valuable data analysis.

However, without programming request by users, Tableau only does a structural process for the already organized data, which is technically less difficult, so that tableau can just present simple static visualization images, and can not achieve the interaction of charts. When doing some complicated graphs in tableau, it is more troublesome. For example, when doing Sankey diagram, you need to manually write the sigmoid function. Another example, when doing a rose graph, you still need to manually write more complicated calculation fields. Tableau cannot make whatever visualizations that you want.

Data-Driven Documents

The learning cost of Data-Driven Documents is very high, and it takes a lot of time to learn if you have never touched it before. During the most time, we find an example from the web and then modify it to be what we need in the actual project. If we want to create some innovative features, we need to keep searching, trying and changing until it seems to be what we want. But this is very difficult for beginners that learn Data-Driven Documents.

Data-Driven Document was very popular at the time of its release in 2011 because jQuery and Backbone dominated at the time, and the browser only implemented some simple css standards such as “transitions”. The more complex "flex" layout is still in the foreseeable future. However, the current framework uses a more flexible and powerful design concept such as virtual DOM. Also, CSS has become easier in layout and animation for users. Much of the API exposes direct access to the DOM, which might clash a little with how modern frontend frameworks like React or Vue work. There are ways to work around that, though. Also, SVG is not suitable for dynamic rendering and large data volume rendering. This is

because high complexity slows down the rendering (any application that overuses DOM is not fast). So it is also not suitable for gaming applications.

Data-Driven Document doesn't always try to support older browsers. If you want users on those browsers to see your visualization, you may have to use a static placeholder (in the case it doesn't load). Data-Driven Document has some data-source limitations. It cannot easily conceal original data. If you're using data that you don't want shared, it can be challenging to use D3. Data-Driven Document doesn't generate predetermined visualizations for you. If you frequently use Tableau, Matplotlib, Excel and Plotly for a quick turnaround on visualizations. It can be time-inefficient to generate D3 visualizations where a quick chart from an alternative source would perform beautifully.

Echarts vs Pyecharts

The area of the chart is too white to use the space effectively, especially when the space is tight. Echarts cannot import data in bulk. Pyecharts is a class library for generating Echarts charts. Pyecharts can display dynamic graphs that Online reports are more beautiful and display data is convenient. You can hover over the graph to display values, labels, and so on. So we use Pyecharts instead of Charts because our data is very large. If we input data one by one, it is just waste of time. But both of them has the same limitation that users cannot customize the chart whatever they want.

Gephi

Gephi is an interactive visualization and exploration platform for all kinds of networks and complex systems, dynamic and hierarchical graphs. It can visualize all kinds of networks including directed, undirected, weighted, un-weighted, labeled, un-labeled, etc. Gephi is compatible with various operating systems such as, Windows, Linux and Mac OS X.

But it also has some limitations that it only works best for single nodal/edge types. It is less flexible and nuanced than coding languages such as R. Support of Gephi is less robust (updates don’t always happen frequently enough).

To work with data in Gephi, it must be in the form of nodes and edges. Otherwise, there will be no possibility to create a network graph. In theory, you could have nodes only, but this defeats the point of creating and analyzing a network. Gephi provides the ability to convert an edge-only source into nodes, saving you a potential step. However, this approach has some limitations from a node perspective, particularly if you are working with supplemental fields that hold incidental node information to be used for partitioning, ranking, filtering, or any other possible use.

NodeXL

NodeXL will re-running the data extraction on different machines but with the same parameters at the same time. It needs to use multiple seed terms for a specific problem.

NodeXL is an Excel add-in, so you will need Excel to use it which is a bit of a limitation for Mac users for example. It doesn't have all of the flexibility of Gephi in terms of visualization but can produce some quality visualizations.

Part 7: Contributions

Zhou Shuo	3035561510	Data processing, Word cloud, Interactive world map&line chart of university ranking, HeatMap, Video recording
Liu Mingshan	3035562198	Data processing, Interactive bubble chart
Du Ningxin	3035562095	Data processing, Interactive world map&line chart of four other features, Report writing
Ye Jiantong	3035562239	Data processing, Radar chart, Report writing