University of Arkansas – CSCE Department

Capstone I – Final Proposal – Fall 2019

Data Visualization

Forest Tennant, Michael Fahr, Bryce Mendenhall, Pao Yang, Rafael Del Carmen

Abstract

Large amounts of data are generated on an everyday basis and can be identified in different categories. At Sorcero, it could be about the corpus or the user interactions with the corpi. With such a large amount of data, it is hard to navigate and be able to find the information that is needed. For example, this information could be useful to help find insights on key performance indicators (KPIs). The objective is to analyze the large volumes of data and providing a meaningful visual context that could help provide means of navigation and insights into key performance indicators.

The approach is to design and implement a program that collects the data, develops relations in the data using a set of algorithms, and outputs a straightforward visual context based on the information provided. The data visualization resulting from this approach is important because it will assist the user in processing and understanding the data shown. By being able to navigate through the data in an understandable way, it becomes easier to detect trends or patterns. It also communicates the data quickly and effectively to other people who may not be familiar with the information.

1.0 Problem

Corporations gather and generate large volumes of data, including information about documents and how users interact with these documents. It is crucial for this data to be stored so that it can

be analyzed. However, as this amount of data grows in size, it becomes progressively more difficult to visualize the core information provided from this data and navigate through the data in a meaningful way. Also, the data doesn't provide a clear way of measuring performance towards a specific goal. It is important for companies to be able to understand the data that is being generated to fully capitalize on the information that the data is providing.

Without a solution to this problem, there will not be a quick way to understand and visualize large amounts of data. It will be necessary to manually search through the data looking for information that shows progress towards specific goals. This manual search of data will be extremely inefficient compared to an automated program searching the data, ultimately wasting company resources.

2.0 Objective

The objective of this project is to receive large volumes of data gathered from corpus, analyze the data using natural language processing, develop algorithms to create relations between the data, and use the created relations to provide a meaningful visualization, means of navigation, and insights into key performance indicators.

3.0 Background

3.1 Key Concepts

One of the key technologies related to this problem is natural language processing (NLP). NLP focuses on how computers interpret and analyze natural language data. NLP systems have been designed by using a rules-based system but have more recently moved into a machine learning approach. By analyzing large inputs of data, rules can be inferred and created to help develop an approach more specific to that system. Examples of this include having computers answer human created questions, speech recognition, text-to-speech, etc.

Apache Hadoop[6] software library is a framework that allows for efficiently storing and processing very large data sets across clusters of computers in parallel. One example of an advantage from using Hadoop is for processing weather data. Weather sensors around the country that collect data every minute can easily generate billions of bytes of data, and traditional relational database management systems aren't capable of processing this amount of data. Hadoop can easily be used in different languages like Java, Python, or C++. Apache Hadoop is a good solution for handling Big Data.

3.2 Related Work

Many modern software solutions exist for data visualization, ranging from simpler user friendly drag and drop programs such as Tableau[1] to more programmatic and adaptive platforms such as MATLAB[2]. Even AI driven applications like IBM Watson Analytics[3] exist for displaying such data.

All of these platforms are established with dedicated user bases. However, there is always room for improvement in software. In particular, all of these platforms offer broad, mostly generic functionality due to that they are meant to appeal to as broad a spectrum of users as possible. However, this means that these platforms are complicated, intricate things that pose many features and functions that a single user will likely never use.

IBM Watson, for example, offers the promise of powerful AI computing. However, the average user will simply not require that sort of power. Tableau offers a powerful and diverse range of functionality, but requires the user themselves to decide what is and is not right for them. This task could be daunting for some users. In the case of MATLAB, the powerful features of the language come with the caveat that truly tapping into its potential requires the technical skills of someone familiar with using MATLAB.

Our program then offers the appeal of a targeted application. Instead of building their own set of data visualization tools from the broad offerings above, Sorcero will benefit from the use of a target set of tools developed specifically for their needs. This will allow them to spend less time determining which tools and skills they need to analyze their data.

Muriel Cooper presented a project at a TED5 conference in Monterey that changed the way designers look at the possibilities of electronic media.[4] Cooper presented the work that she did and took it to another dimension.[4] Information Landscapes was a series of projects that happened over a period of time led by Cooper and her students. It was a dynamic and malleable element. This project used different viewpoints and space to show the various ways of visualising sets of data. It challenged the possibilities of conveying information through computer and graphic design in the digital realm.[5]

4.0 Design

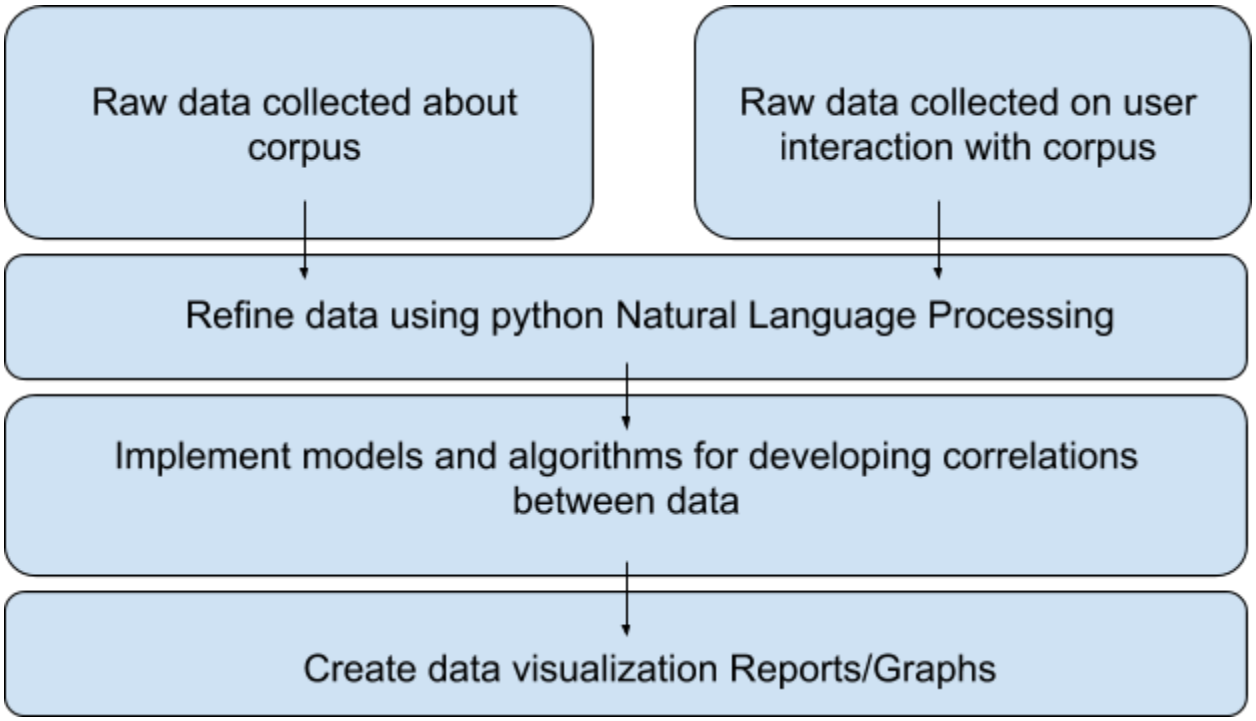
4.1 Requirements

- Store the large volumes of data
- Group data into similar data sets
- Use NLP to create correlations between the data sets
- Provide additional means of navigation through data sets
- Generate a communicable visual context based on the data provided
- Show significant results with insights into key performance indicators

4.2 High Level Architecture

At the forefront of our Data Visualization tool is two sets of raw data. One set is the data about the corpus, and the other set is the data from the user interaction with the corpus. Our visualization tool will collect this data from the corpus provided. This will be our raw data. We will then clean and refine our raw data by converting it to proper data types and sorting it into one of two sets of clean data. The next step is to run the data through our natural language processing algorithm. This algorithm will take our natural language data input and process it, recording the

paragraph in which it occurs and how many occurrences it has had. This is where our database storage will be important. It will have a record of each paragraph from our PDF input. Each word that occurs within the paragraph will be linked to the paragraph accordingly. Furthermore, each word will correspond to its other matches in the document. If we visualized this connection, it would look like a spiderweb spanning across the PDF connecting each word to its corresponding paragraph and other word matches. Each set of data, data about the corpus or user interaction data with the corpus, will go through this process of building these connections and being recorded in the database separately. The tool will then generate a straightforward visual-based report allowing the user to select which dataset they would like to view. The user will then be able to select certain words or paragraphs and select how they would like the data visualized. For example, if they select a landscape visualization and click on a paragraph, an information landscape will be built showing the weight of the words within that paragraph.



4.3 Risks

Risk	Risk Reduction
Misrepresentation of data due to incorrect correlations from the Natural Language Processing	This risk can be minimized by inspecting to make sure that incorrect correlations don't appear in the outcome. If we find that this has happened, we can alter our algorithm to help prevent this.

Algorithms becoming biased to the test data set and producing undesired results with other data sets	This risk can be minimized by testing our code on multiple data sets to ensure that a bias is not developed for a specific data set.
Difficulty using Natural Language Processing to analyze graphs and tables and accurately represent the data through visualizations	This risk can be minimized by researching multiple different techniques for analyzing graphs and tables and selecting one that produces the best result.

4.4 Tasks

1. Explore and understand the background of data visualization and natural language processing
2. Research other modern implementations to get an idea of other approaches
    a. Explore other alternatives for processing large amounts of data other than using traditional relational databases
    b. Determine the advantages and disadvantages for using each of the explored alternatives
3. Finalize architecture design and language of implementation
4. Develop code to process the large volumes of data
    a. Decide the best method of storage for this data to prevent running out of memory
    b. Determine the best way to process large volumes of data on commodity hardware which won't require expensive servers with high processing power and large memory
5. Create an algorithm for grouping/sorting the data into related fields
    a. Filter out stop words to help reduce the natural language that will have to be processed
    b. Investigate multiple methods of grouping/sorting to determine which results in the best relations
        i. Use a Grouping Method
            1. Determine fields that data can be grouped into
            2. Use a combination of each data set's type, title, and content to assign it a group
        ii. Paragraph Method
            1. Analyze each paragraph separately
            2. For each word, record the:

        a. Paragraph number it appears in
        b. Frequency of the word in that specific paragraph
    3. When selecting either a paragraph or a specific word in a paragraph from the original text, display the relations to other paragraphs where the word/words appear
        a. Prioritize displaying paragraphs that contain the highest frequency of the word
  iii. Other Methods
6. Use the data to provide a meaningful visual context that suits the data
    a. Decide what parameters of the data are important to provide a visual representation
    b. Use NLP to parse each data group for indicators of the parameters
    c. Utilize provided libraries to create a visualization for the data
7. Finalize the program by testing the application on multiple large sets of data
    a. Ensure the results and visualization provided from the application are relevant and significant
8. Document the final results

4.5 Schedule

Tasks	Dates
1. Do some research and understand the background of data visualization and natural language processing.	1/13-1/20
2. Research other modern implementations to receive an idea of other approaches. This includes exploring other alternatives for processing large amounts of data and determining the advantages/disadvantages based on the different alternatives.	1/21-1/27
3. Finalize architecture design and language of implementation.	1/28-2/10
4.Develop the code to intake the large volumes of data by deciding the best method for data storage	2/11-2/24

without running out of memory. In addition, determine the best way to process large volumes of data that does not require expensive servers with high processing power and large memory.	
5. Determine the group/sort method that will be used in the algorithm. Create an algorithm that sorts/groups the data into related fields based on the method that was chosen. In addition, incorporate a filter to reduce the amount of natural language that has to be processed.	2/24-3/9
6. Use the data to provide a meaningful visual context that suits the data by choosing different parameters to find which data is important to provide a meaningful visual representation.	3/9-3/23
7. Finalize the program by testing the application on multiple large sets of data.	3/23-4/6
8. Document the final results.	4/7-4/17

4.6 Deliverables

- Design Document: Contains information on the programming languages that are used in the implementation process. This also contains the design process for implementing the NLP with the database. Furthermore, it includes different design features that were used

to implement the final software and explains various aspects of the program such as the inputs and outputs.
- Database schema and initial data: Schema for storing the relations concluded from the NLP and the initial data these relations are created from.
- Python code: All of the code that was used to implement the project.
  Includes:
  - code that was used for analyzing the data with NLP
  - code that uses this data to provide the visualization
- Website code: The PHP code for the web site split into three main subdirectories.
- Final Report: A report that summarizes the process and outcomes of the program that was implemented.

5.0 Key Personnel

Michael Fahr – Fahr is a senior Computer Science major in the Computer Science and Computer Engineering department at the University of Arkansas. He has completed Software Engineering and Programming Paradigms. Michael has received relevant experience with software engineering and structuring databases while interning with J. B. Hunt. Michael will be responsible for creating the relations between the data that will be used for the visualizations. (Tasks 1, 2, 3, 4, 5, 7, 8)

Forrest Tennant - Tennant is a senior Computer Engineering major in the Computer Science and Computer Engineering department at the University of Arkansas Fayetteville. He has completed Software Engineering and Programming Paradigms. Forrest will be responsible for implementing visualizations on the collected data. (Tasks 1, 2, 3, 6, 7, 8)

Bryce Mendenhall – Mendenhall is a senior Computer Engineering major in the Computer Science and Computer Engineering department at the University of Arkansas. He has completed Programming Foundation I/II, Software Engineering, and Programming Paradigms. Bryce will be responsible for implementing the Natural Language Processing and visualization of the data. (Tasks: 1, 2, 3, 5, 6, 7, 8)

Rafael Del Carmen – Del Carmen is a senior Computer Engineering major in the Computer Science and Computer Engineering department at the University of Arkansas. He has completed Cloud Computing, Programming Foundations I/II, Programming Paradigms, and Software Engineering. Rafael will be responsible for implementing the Natural Language Processing and develop code for processing large amounts of data. (Tasks 1, 2, 3, 4, and 5)

Pao Yang – Yang is a senior Computer Science/Computer Engineering major in the Computer Science and Computer Engineering department at the University of Arkansas. He has completed Programming Foundation I/II, Software Engineering, and Programming Paradigms. Pao will be responsible for taking the data collected and implementing the visualizations based on this data. (Tasks 1, 2, 3, 6, 7, 8)

6.0 Facilities and Equipment

There are no specialized facilities or equipment that we require to complete this project.

7.0 References

[1] “Tableau: Business Intelligence and Analytics Software.” Tableau Software, www.tableau.com/.

[2] “MATLAB.” *MathWorks*, www.mathworks.com/products/matlab.html.

[3] “Smart Data Analysis and Visualization.” *Watson Analytics*, www.ibm.com/watson-analytics?lnk=hmhm.

[4] David. “Muriel Cooper: Information Landscapes.” *Inventing Interactive*, 28 Apr. 2014, www.inventinginteractive.com/2010/02/01/information-landscapes/.

[5] “Inbox: Muriel Cooper's Information Landscapes.” *The Museum of Modern Art*, www.moma.org/calendar/exhibitions/1654.

[6] “Apache Hadoop.” Apache Software Foundation, www.hadoop.apache.org.