Thomas Muamar

Data Visualization EN.605.662.81.FA23

Project #4: Interactive Visualization using Python

Introduction:

In this project we are to use the data visualization libraries that are in python, r or java script to create interactive visualizations. These interactive visualizations were created based on the text analysis that was conducted in python or R. Insightful conversation with ChatGPT, we engaged in a dialogue that delves into the complexities of automation's impact on society and the economy. The conversation is structured around a pivotal topic, the Economic Implications of Automation. We navigate through a series of thought-provoking questions and insights, each shedding light on distinct parts of this transformative concept. My exploration begins by unraveling the comprehensive dimensions of automation's influence on the economy. I seek to understand how automation is reshaping economic dynamics. With ChatGPT the conversation got into the ways automation alters employment patterns, job prospects, and the skills demanded by industries. It led to strategies that governments can adopt to mitigate potential job losses resulting from automation. I then asked a question on how educational and training programs can adapt to prepare individuals and the workforce for the changing job landscape. In essence, I will leverage this conversation with ChatGPT to conduct text analysis and create interactive visualizations using Python.

Analysis:

In my first text analysis the code reads the contents of the chat.txt file and changing the values to lowercase to allow for a consistent word frequency analysis performed. Using the Natural language toolkit, I used the method word_tokenize. This is a script that splits the text into individual words. This process allows for a breakdown of the text into manageable units for analysis. Using freqdist from NLTK it calculates the frequency of each word in the text. The text is also filtering out stop words which allows for a more accurate depiction of the results. These frequencies are put into a data frame which will allow me to plot these results using plotly. Plotly is used to create this interactive bar chart visualization. The results produced from the word frequency analysis from the talk with chatgpt illustrate the key themes and topics related to the impact of automation on the job market and the evolving needs of education and training. In the below table are the top words in the frequency.

	Word	frequency
3	automation	37
62	skills	27
74	job	24
120	education	20
121	training	17
47	may	16
41	jobs	14

40	workers	14
124	programs	14
58	new	11
110	businesses	11
126	workforce	10
215	learning	10
42	automated	10
24	work	9
109	opportunities	9
26	need	9
137	impact	8
59	also	8
21	lead	8
0	implications	8
229	encourage	8
1	economy	8
303	individuals	7
36	displacement	7
75	market	7

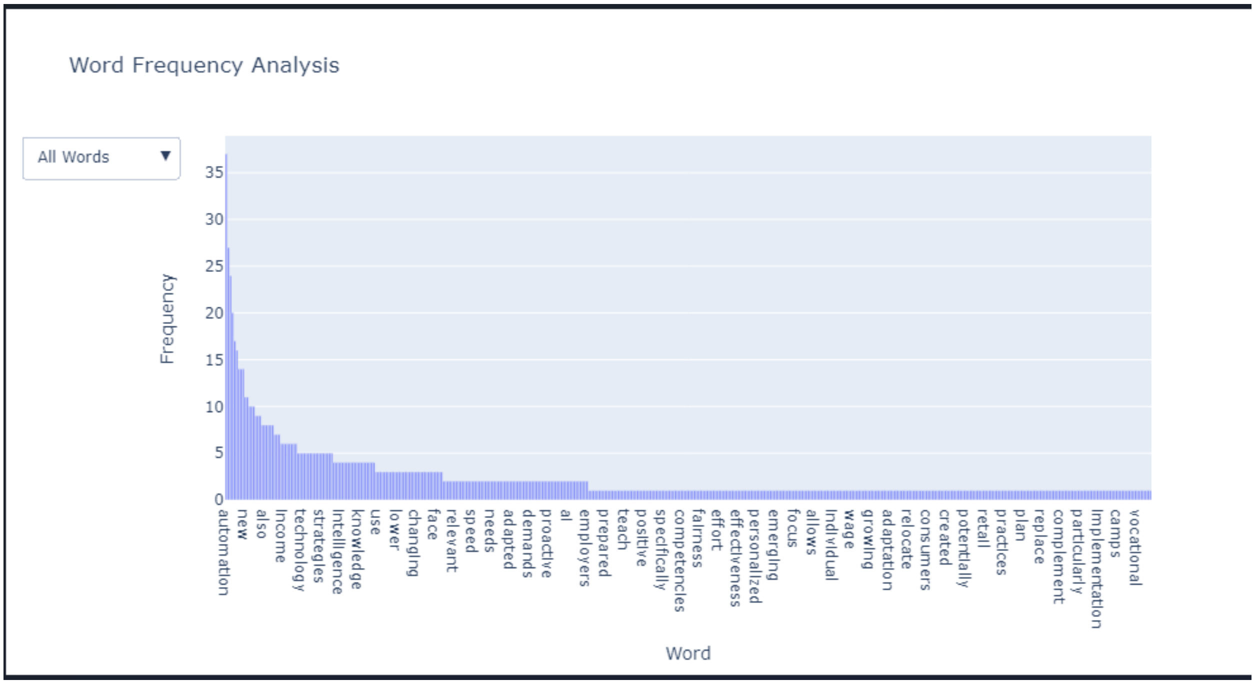
The next analysis I performed focused on sentiment analysis of the chatgpt talk, aiming to understand the emotional tone and subjectivity within them. The libraries used in this one are textblob, pandas, and plotly. The steps for this analysis started with reading and preprocessing the data. It read the text data from the file and stored it line by line in a list of named texts. In order to perform the sentiment analysis I took the text in the list and created a textblob object. I then extracted the entiment from each text blob which gives polarity and subjectivity. The pandas library helps create a dataframe df with the columns text, polarity and subjectivity. From this sentiment analysis, the polarity score is close to 0 which indicates a neutral tone. This is usually more common in factual and informative text. It seems like some of the negative polarities shown in this analysis were more critical view regarding certain aspects of automation. The subjectivity scores were shown to be high in some instances which indicates personal opinions or judgements that are expressed. Most of the text indicated a moderate subjectivity score so it suggested a balanced mix of information and opinion. Deep diving in the results the topics related to job displacement, skill shifts, and income inequality show varying polarity, reflecting the complex and multifaceted impacts of automation. Then there were the discussions on strategies to mitigate job losses, like investing in education, have generally positive or neutral polarity, indicating an optimistic view on these solutions. Overall, the sentiment analysis paints a picture of a nuanced and comprehensive discussion around automation. It reflects a range of emotions and subjectivities, highlighting both the concerns and potential solutions in the realm of automation's impact on jobs, education, and the broader economy.

The next analysis I used the text tokenization from NLTK and the named entity recognition from nltk. The text tokenization allows me to break down the text into manageable units to be able to process it further. Then the name entity recognition is used to identify named entities in the text which helps extract structured information. The results would be the frequency and the type of entities, and the graph produced will help display this connectivity between these entities. This type of analysis allows for understanding the relationships and structures within the text of data and it is making it very valuable information.

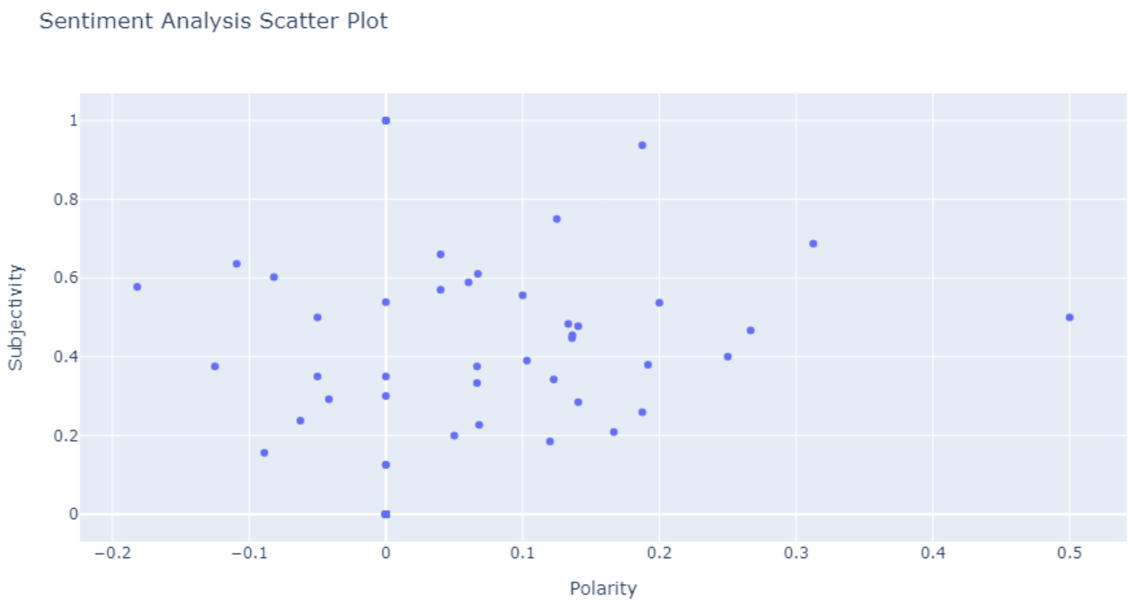
The fourth analysis conducted is a part-of-speech tagging analysis which was using NLTK. The analysis involves taking text documents and reading and tokenizing it to get the smaller units of words for further analysis. Then with POS tagging each word in the text is assigned this tag using NLTK's pos_tag function. The POS tagging in this process of marking up a word in a text as corresponding to a particular part of speech like nouns, verbs, adjectives, and so on. In the code I count each of the occurrences of each tag to allow for a quantitative overview of the distribution. With in the code, it also grabs 5 examples for each of the tags to give more detailed and a more interactive chart that allows user to understand the frequency of each tag in the text. Overall, POS tagging highlights the basic grammatical composition of the text, such as the prevalence of nouns, verbs, adjectives, and adverbs. This can give a general sense of the text's nature.

The fifth analysis conducted was using word frequencies and identifying the most important sentences based on these frequencies. The libraries involved in this analysis are NLTK, plotly, and pandas. The analysis was conducted by doing both sentence tokenization and word tokenization. I then used filtering of stop words and punctuation to get a more accurate frequency. Each sentence is scored based on the frequency of the words it contains. The score of a sentence is the sum of the frequencies of its words. I then got the highest scores and formed a summary. These sentences are then detokenized to form a coherent summary of text. The summary provides a quick overview of the text, which can be particularly useful for understanding long or complex documents without reading them in full. In summary, the results from this script offer a quick and surface-level understanding of the key contents and structure of the text, which is particularly useful for large datasets or preliminary data exploration.

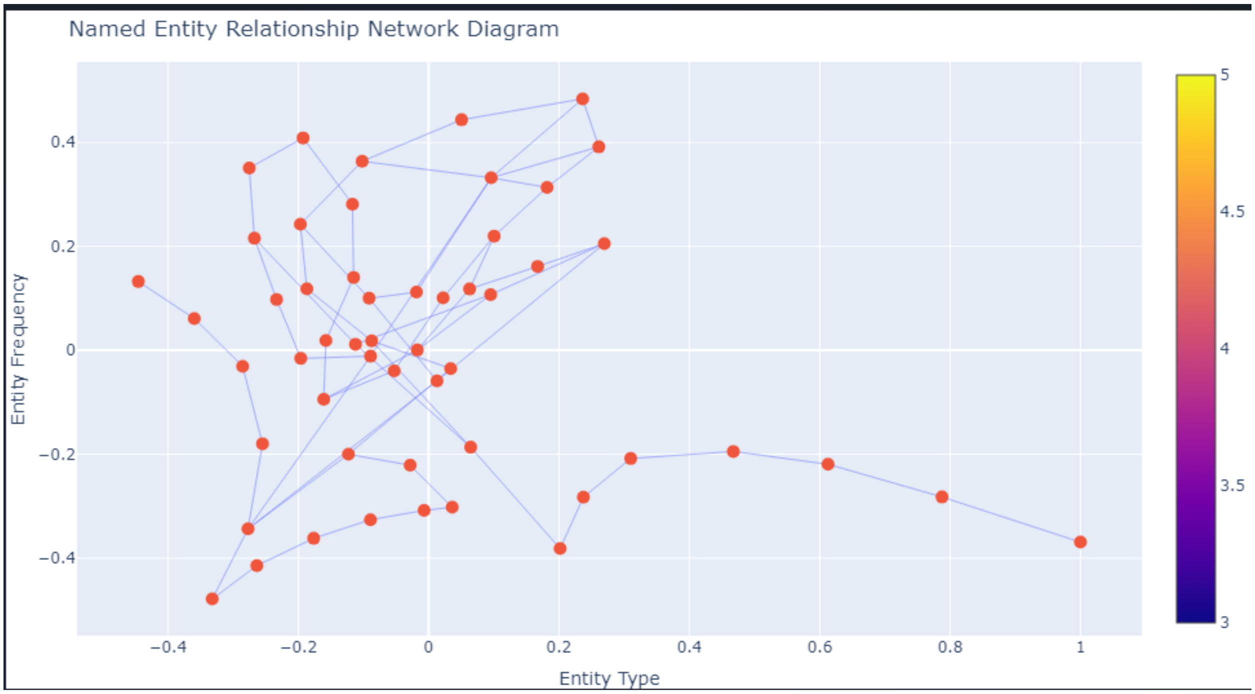
Visualization of text analysis:



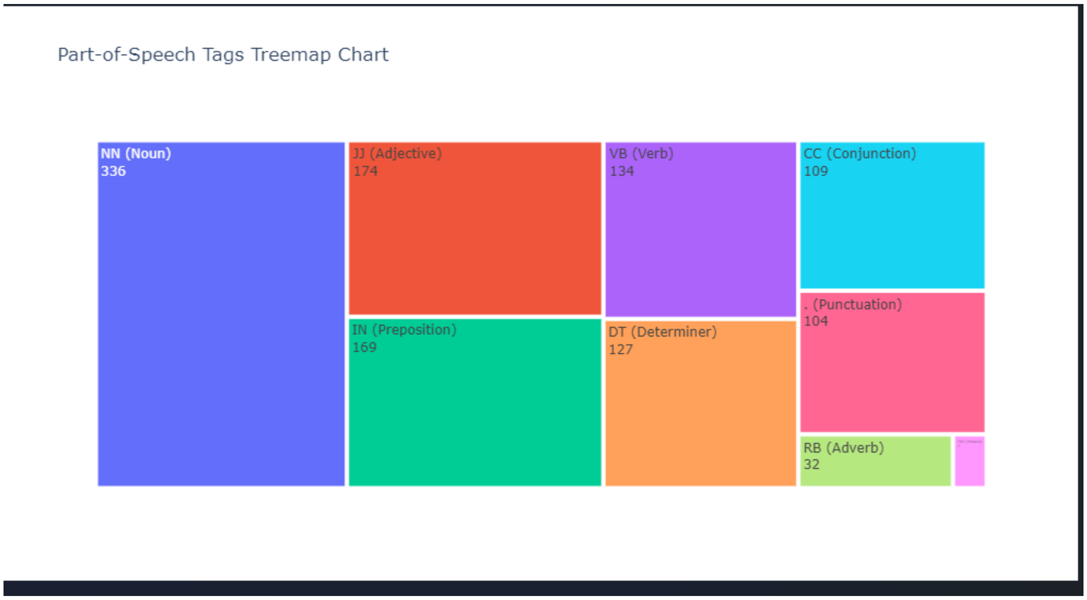
The above visualization is a bar chart that presents a word frequency analysis. The chart visualizers the frequency of different words from the text. Words are on the x-axis are ordered based on their frequency, with the most frequent words on the left and least frequent on the right. Hovering over each bar users can see the exact frequency count of each word. There is also a dropdown filter which allows the users to filter the view for top 10 or top 20 words to get an easier visualization of the data. Users can zoom in to focus on sections of the chart and to pan to navigate through the zoomed in areas. This is important because it identifies high-frequency words and readers can quickly determine primary themes or focal points of the text. Basically, words that appear more frequently might indicate areas of emphasis or importance in the text. Overall, the bar chart is a powerful visualization technique for word frequency analysis. It provides an intuitive overview of the prominence of words in the text which allows users to quickly understand the main themes and potential areas of emphasis.



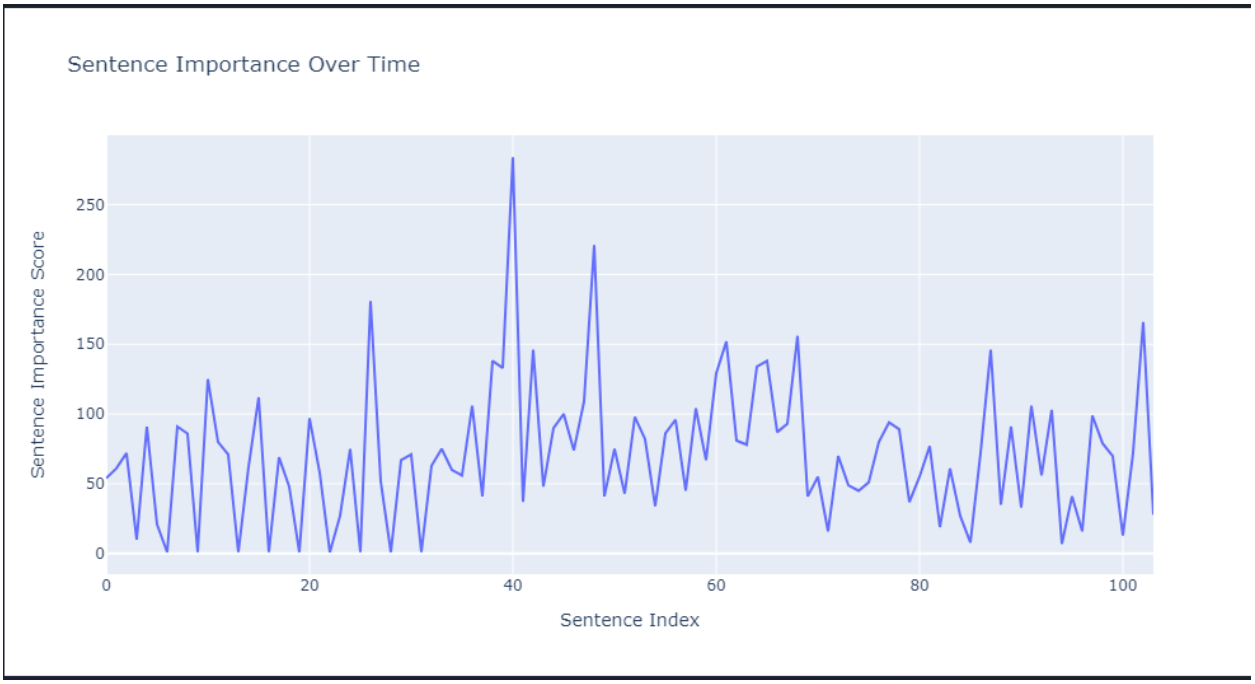
This next visualization is a scatter plot of the sentiment analysis. The scatter plot visualizes sentiment based on the two metrics which are polarity and subjectivity. Hovering over individual points typically displays detailed information about that data point which is the specific text and the polarity and subjectivity value. This scatter plot is relevant because it gives an immediate visual overview of the general sentiment of the data whether it leans towards positive, negative, or neutral. It can be used to identify outliers such as the point of the far right which can tell us that they might have a unique or unusual sentiment characteristic. The concentration of points in certain quadrants can give insights into the nature of the text. For example, many points in the top-right quadrant (high polarity, high subjectivity might indicate a lot of positive and emotional content. In summary, the scatter plot is an effective visualization technique for sentiment analysis as it allows for a quick visual assessment of both sentiment and the degree of subjectivity in the data. It provides an immediate understanding of the overall tone and nature of the content being analyzed, making it a valuable tool in sentiment analysis tasks.



This is a network diagram that is visualizing the named entity relationships. It basically maps out the relationships between various named entities in the text. Each node represents a distinct entity while the connecting lines or edges depict relationships between them. The color gradient on the right suggests a scale for measuring the importance or prominence of entities, with warmer colors (red to yellow) indicating higher importance. Hovering over a node provides details about the named entity such as the name, and type. The network diagram provides an overview of relationships and interdependencies between entities, which is difficult to ascertain from just the textual formats. The network diagrams can reveal patterns, such as clusters of closely related entities or outliers that might be indicative of certain themes or topics in the text. Overall, it aids in understanding the relationships between named entities.



This treemap provides a visual representation frequency of different part-of-speech tags within the chatgpt text. Each tag corresponds to a category like noun, verb, or adjective. Each of the rectangles area corresponds to the frequency of each of the tags. Larger the rectangle the more of that certain tag. Hovering over the rectangles displays information like example words, name, and the frequency. The treemap offers an immediate visual snapshot of the frequency of different linguistic categories. Understanding this distribution of pos tags can guide decisions on further analyses like focusing on specific categories or comparing text. In summary, the Part-of-Speech Tags Treemap Chart is a effective visualization tool that translates linguistic data into a visually inducing and interactive format. It benefits in the rapid assessment of a dataset's linguistic characteristics and serves as a foundation for more nuanced and detailed analyses.



This is a sentence importance over time line chart. This line chart provides a visual representation of the perceived importance of sentences in a sequential order from the text. Hover over each of the points on the line gives the score and the index at which the sentence is at. It allows for quick pinpoint of sentences or sections with higher importance scores. It helps with extracting key points from the text. This chart can reveal the flow and structure of the text by showcasing peaks and troughs of important information. If someone were reviewing the text from chatgpt they can now tell which sentences carry higher importance and can basically guide an individual to pay attention to these parts. Overall, the sentence importance over time line chart serves as a crucial tool for understanding the distribution of information within the text from chatgpt.

Conclusion:

In my conversation with ChatGPT, I decided to employ five distinct analytical methods, each complemented by a unique visualization. Initially, word frequency was assessed, and its visualization using a bar graph provided insights into the predominant themes by highlighting frequently used words. This was followed by a sentiment analysis, where a scatter plot effectively showcased the text's overarching sentiment and its subjectivity. The named entity recognition, illustrated through a network graph, offered a clear view of the interrelationships between the named entities. A treemap was then used to present the part-of-speech distribution, emphasizing the text's grammatical composition. Finally, the line chart depicting sentence importance over time proved invaluable in understanding the distribution and emphasis of information throughout the text.

References:

1. Arora, S. (2023, July 24). Sentiment analysis using Python. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2022/07/sentiment-analysis-using-python/

2. *Getting Started with Plotly in Python*. Getting started with plotly in Python. (n.d.). https://plotly.com/python/getting-started/
3. Burchfiel, A. (2023, May 11). *What is NLP (Natural Language Processing) tokenization?*. tokenex. https://www.tokenex.com/blog/ab-what-is-nlp-natural-language-processing-tokenization/#:~:text=Tokenization%20is%20used%20in%20natural,into%20understandable%20parts%20(words).