Margarita Prikhodko

EN 605.662 Data Visualization Project  
  
  
PRoject #4:   
Interactive Visualization of ChatGPT conversations using Python, R, or JavaScript

*“Some people call this artificial intelligence, but the reality is this technology will enhance us. So instead of artificial intelligence, I think we will augment our intelligence.” —*Ginni Rometty

1. **Introduction**

Goal of this paper is to analyze conversation with ChatGPT by using NLT libraries and then visualizing the information. An additional focus to of the research is to analyze difference between human and AI text.

ChatGPT is an application based on Natural Language processing. This language processing tool is powered by AI technology. ChatGPT creates human-like conversations and can answer questions and complete tasks, such as composing emails, essays, and code.

The introduction of ChatGPT has created an internet level distribution. It is the fastest-growing app ever released. The app exceeded one million users in its first five days, and surpassed 100 million users two months after its launch.

ChatGPT is currently in its research and feedback-collection phase. It is powered by a deep neural network. These types of software architectures are inspired by the structure of the brain. The app can learn to perform tasks by processing multiple examples and identifying patterns.

However, the app is currently not sophisticated enough to completely replace the elements of human creation. Although ChatGPT can often write elegant prose, it cannot currently match the human intellect. The app often struggles with basic word problems.

My hypothesis is that by analyzing lemmatization, word proximity, word frequencies, I can identify text created by current versions ChatGPT and other AI programs with a 99% accuracy. Graphs will be used to demonstrate the difference between AI and human authors. It should be noted that these applications are rapid advancing and may quickly outpace our ability to distinguish between human and artificial authors.

For this paper, I analyzed the works composed by human and AI authors for the following subjects.

1. Write review of The Three Musketeers by Alexandre Dumas
2. Which historical events have been described in the Three Musketeers?
3. Who are main heroes in The Three Musketeers?

This paper visualized the output of conversation. We run natural language analysis if conversation and compare it with human review of the book. In a new paper, cognitive scientists and linguists address this dissonance by separating communication via language from the act of thinking: Capacity for one does not imply the other.

I took this human review of book from here:

https://www.goodreads.com/book/show/7190.The\_Three\_Musketeers

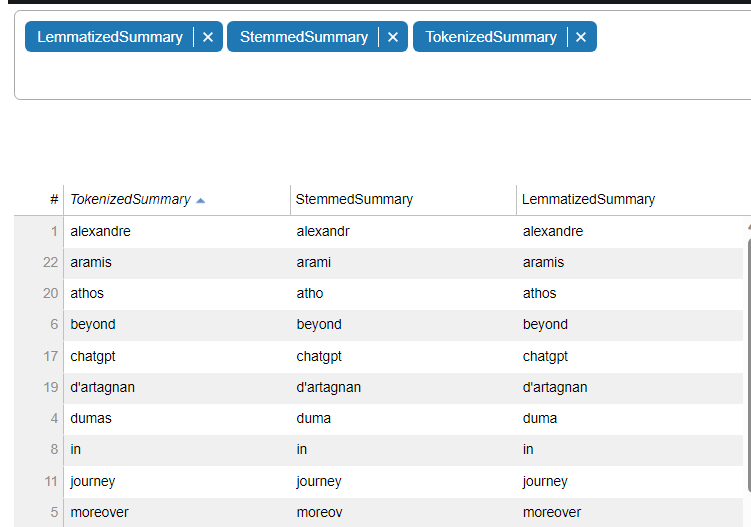
1. **Methodology**

Natural languages libraries have several tools such as lemmatization, word proximity, word frequencies, number of lemmas for the same words, topic discovery. The list of Python visualization tools includes Bokeh and Gensim

1. Lemmatization: NLTK Lemmatization is the process of grouping the inflected forms of a word in order to analyze them as a single word in linguistics. NLTK has different lemmatization algorithms and functions for using different lemma determinations. Lemmatization is more useful to see a word’s context within a document when compared to stemming.

Unlike stemming, lemmatization uses the part of speech tags and the meaning of the word in the sentence to see the main context of the document. Lemmatization is important for understanding a text and using it for Natural Language Processing, and Natural Language Understanding practices. NLTK Lemmatization is called morphological analysis of the words via NLTK. To perform text analysis, stemming and lemmatization, both can be used within NLTK. The main use cases of the NLTK (Tuğberk, 2021)

Output of research represented as table 1:



*Table 1 Lemmatization, stemming*

This method will identify the original sentence and preprocessing output. Some redundancy is noticed. The heatmap in graph 1 represents the unique Lemmatization versus number of words.

A screenshot of a computer screen

Description automatically generated

*Graph 1 Numberof tokens versus lemmas*

1. Statistics: The “Bag of Words” visualization identifies the frequency of words that occurred in text for top 20 words. Histogram representation was used. Generally, human authors will use a higher variety of word variations (graph 2) than OpenGPT (graph 3).

A graph with different colored bars

Description automatically generated

*Graph 2 – Human Review*

A graph of words visualization

Description automatically generated

*Graph 3 – OpenGPT*

The human authors have more variety and richer vocabulary. These graphs are based on collocations. <https://www.nltk.org/howto/collocations.html>

In graphs 4 and 5, I measured the connection between word by using Graph methods. I observed a significant difference. The proximity of words is much more unique in human written text. The word choices used by ChatGPT is pointedly more uniform.

1. Graph Visualization:

For this research, I applied Graph Visualization to picture the proximity of word usage. These graphs calculate the proximity of words based on graph theory. The nodes are words and edges are distance between words (Herman 2000).

A computer screen shot of a network

Description automatically generated

*Graph 4 – ChatGPT word usage*

A screenshot of a computer

Description automatically generated

*Graph 5 – Human Author word usage*

1. **Genism visualization**:

Gensim software automatically detect topics in text and makes connections with different underlying data. Gensim utilizes saliency terms and relevance terms in discovering 3 topics. We analyzed frequency and connection between topics [2,3].

A screenshot of a graph

Description automatically generated

*Graph 6 – Topics discovery in text. We find 3 main topics in text.*

**III. Conclusion**

The paper provide visualization by using all available tools such as Gensim, Boken for visualization in Python Heatmap, Histograms, Tables, Graphs techniques were implemented for visualization.

This paper concludes that current AI applications currently do not have the depth of language to compete with human authors, and AI authors can easily be identified using the techniques described in this research paper. This would conclude that programs such as ChatGPT do not truly “understand” the application of language and are simply rewriting data from books and Wikipedia entries without comprehending it.

If ChatGPT were truly intelligent, it would display a more sophisticated method of combining language and thought—not just layering different algorithms. However, OpenAI continues to release weekly [update](https://help.openai.com/en/articles/6825453-chatgpt-release-notes)s to ChatGPT. ChatGPT continues to improve its mathematical, linguistic, and cognitive capabilities. There is a point where AI applications may truly become “intelligent”. At this point, AI will be capably disrupting every aspect of how we live and work. Hopefully for the better.

**References**

1. [NLTK Lemmatization: How to Lemmatize Words with NLTK? November 7, 2021 by Koray Tuğberk GÜBÜR

2. Termite: Visualization Techniques for Assessing Textual Topic Models Jason Chuang, Christopher D. Manning, Jeffrey Heer Stanford University Computer Science Department 2013

3. Ivan Herman, Guy Melancon, M. Scott Marshall. IEEE Transactions on Visualization and Computer Graphics, 2000.