CSC 404/604, Fall 2024, Assignment 11

handed out Nov. 25, due Dec. 5 at 11:59pm on Canvas

In this assignment, we will learn how to use some elements of a package called NLTK to perform some textual analysis. In particular, we will learn how to perform sentence and word tokenization, compute and plot word frequency distributions, learn to remove stop words and punctuation, perform stemming and lemmatization, as well as part of speech tagging, and named entity recognition. A good place to start to learn about the tools you will need for this assignment is: https://machinelearninggeek.com/text-analytics-for-beginners-using-python-nltk/

Here is the description of the assignment:

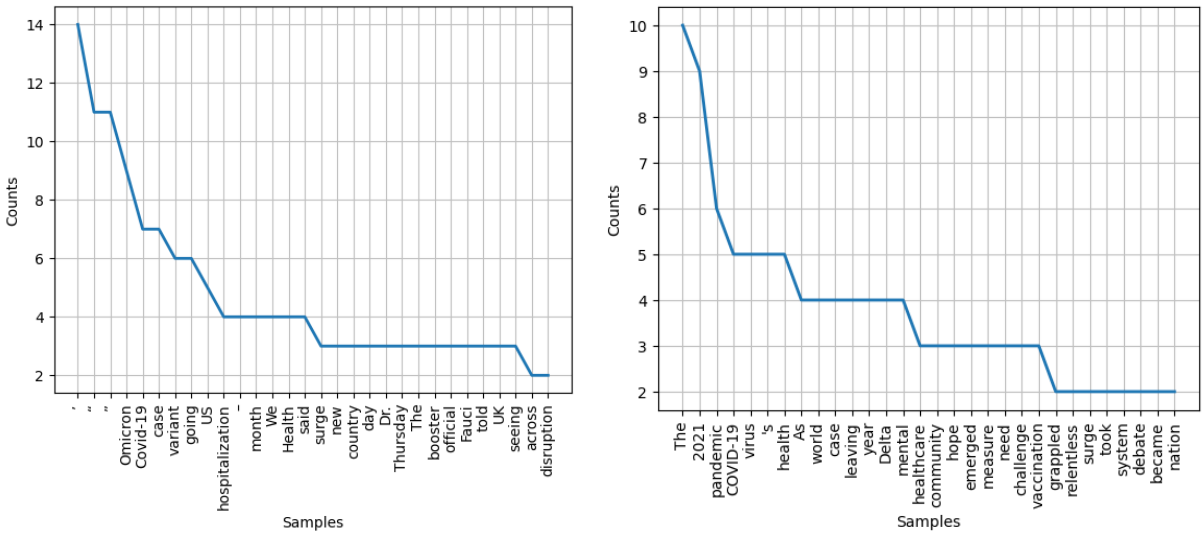
In this assignment, we will analyze two articles. Both are about Professor Lichtman a historian who teaches at American University and his presidential election predictions. The first one was written by a real person, Patty Housman, Assistant Director of Communications for CAS at AU, on September 5th 2024. The second one was written by Chat GPT on November 6 based on the following prompt: "write a news release about Allan Lichtman, the American University Professor who has been correctly predicting presidential election outcomes on many occasions". Note: I was having fun with this: it turns out that this year, Professor Lichtman was wrong in his prediction, but even though the outcome of the elections was known on November 6, ChatGPT had not yet been updated. The two articles are available along with this assignment.

I am showing an example of a similar study I did on two articles (one, genuine, the other one ChatGPT generated) about Covid-19. The first one was written by a real CNN journalist in 2021. The second one was written by ChatGPT two summers ago. Please run a similar study on the articles about Professor Lichtman. Please feel free to add more features if that helps you reach conclusions about the human/machine nature of both articles.

In the earlier study, the first article is entitled: "Covid-19 is raging across the US, with a surge in cases and hospitalizations causing new disruptions" and was written by Dakin Andone, for CNN. It was last Updated 8:20 PM EST, Thu December 16, 2021.

The second article is entitled: **"**Unmasking the Pandemic Surge: A Sobering Look at COVID-19 Cases in 2021". It was generated on *July 31, 2023* by ChatGPT prompted as follows: "write an article about a surge in covid cases in 2021 in the cnn style"

After performing word tokenization, removing stop words and punctuation, performing lemmatization, computing and plotting word frequency distributions, I obtained the following two graphs for the True (left) and Generated (right) articles:



I made two observations of interest: The true article mentions Omicron 9 times (which started in November 2021) whereas the Generated article doesn't mention it or at least mentions it fewer than twice. The Generated text mentions 2021 9 times. The true text doesn't.

I ran a grammatical analysis using Part-of-Speech Tagging and found the following:
True:
Counter({'NNP': 56, 'NN': 44, 'NNS': 37, 'JJ': 24, 'RB': 24, 'VBG': 22, 'IN': 22, 'CD': 16, 'VBN': 14, 'VB': 12, 'DT': 10, 'VBP': 10, 'PRP': 8, 'VBD': 7, 'VBZ': 4, 'CC': 3, 'WP': 2, 'NNPS': 2, 'WDT': 2, 'MD': 2, 'PRP$': 2, ',': 1, '.': 1, 'JJR': 1, 'TO': 1, '#': 1, 'RBR': 1, 'RP': 1, 'JJS': 1, 'EX': 1})
Generated:
Counter({'NN': 77, 'JJ': 62, 'NNS': 52, 'VBG': 28, 'VBD': 26, 'NNP': 22, 'IN': 22, 'VB': 17, 'RB': 16, 'VBN': 13, 'DT': 12, 'PRP$': 3, 'PRP': 3, 'VBP': 3, 'VBZ': 3, 'CD': 2, 'CC': 2, 'NNPS': 2, 'MD': 2, ':': 1, ',': 1, '.': 1, 'WP': 1, 'POS': 1, 'TO': 1, 'WDT': 1, '(': 1, ')': 1, 'WRB': 1, 'JJS': 1})

The true text uses proper nouns (NNP) much more than the generated text. The Generated text uses many more adjectives (JJ) than the True text.

I ran a Named entity analysis which confirms the first finding of the grammatical analysis:

The number of Named Entities in the True Text is: 51

The number of Named Entities in the Chat Text is: 20

Are any of these results significant? Hard to say since we performed the analysis on a single pair of documents. If we repeated this study on many other pairs and found similar patterns, then we may start to believe that ChatGPT follows certain rules and not others. By the way, Machine Learning, the subject of the last assignment gives us tools to draw such conclusions about the data we analyze (text or other). An example of such an approach was created by AU Computer Science Professor Roberto Corizzo in the following article: https://ieeexplore.ieee.org/document/10386674