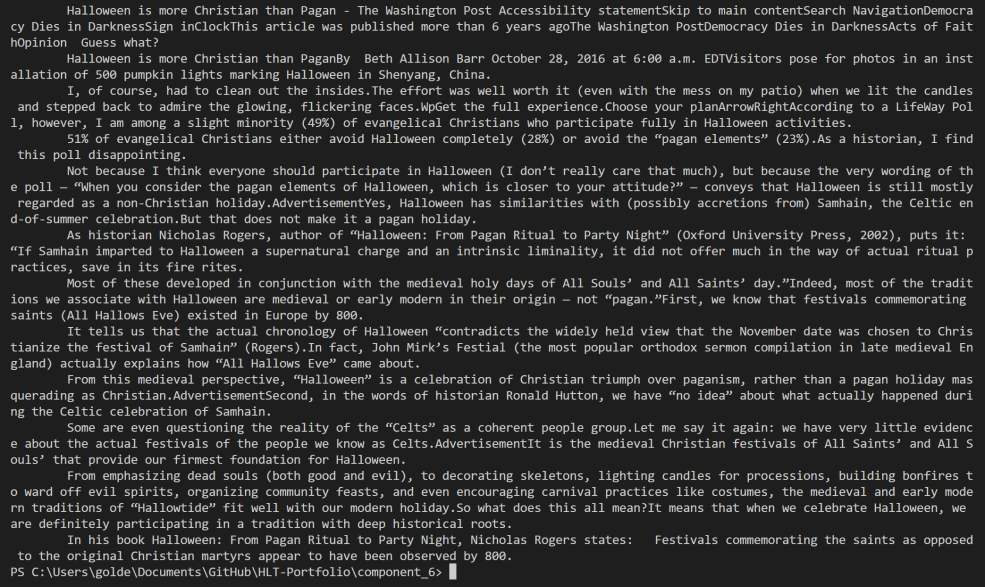
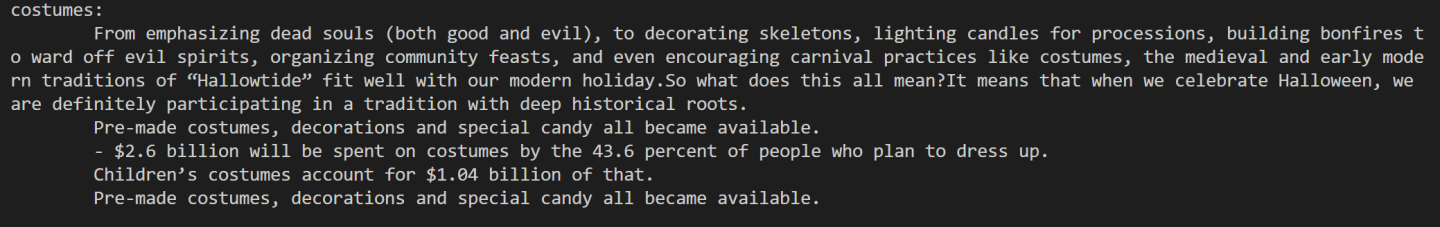
We create this knowledge base by calculating the tf-idf value for each word in the corpus and sorting it. We first did this be calculating the term frequency value for a term and putting that in a dictionary. We then normalized these values by the number of token. The next step was calculating the inverse document frequency for the terms and normalizing that as well. Once we got these values, they were multiplied together to get the tf-idf value for each term. After which we sorted the terms and stored the first 30 values, so we get the most important terms only. Then we manually looked at the list and chose the top Ten most important values, not only by their tf-idf value, but also just on their importance based on our knowledge of the topic. By doing this we got our top ten terms and their tf-idf values. The top ten terms are ['samhain', 'costumes', 'festival', 'halloween', 'illustration', 'church', 'saints', 'crafts', 'disney', 'pagan']. Then we got all the urls and the clean sentences from each url. We checked if a word from our top Ten was in the sentence, and if it was, we added the sentence to our list of facts pertaining to that word. By doing this process, we obtained the dictionary/knowledge base containing the important term and facts related to the term.

Example of Knowledgebase:

Halloween:



Costumes:



1. ChatBot Conversation:

Me: Hello

Chatbot: Hello! What would you like to know?

Me: Can you tell me more about Halloween?

Chatbot: Sure! Halloween is more Christian than Pagan.

Me: Cool! So what can you tell me about the costumes?

Chatbot: $2.6 billion will be spent on costumes by the 43.6 percent of people who plan to dress up.