NPS CHAT: A CHATBOT FOR ANSWERING QUESTIONS ON US NATIONAL PARKS

Betzalel Moskowitz
MSDS453-DL: Natural Language Processing
December 3, 2023

Introduction

Over the last decade, US National Parks have seen a 20% increase in visitation while suffering a 16% loss of employees due to insufficient federal funding (Bennett, 2022). These factors pose challenges for National Park Service (NPS) staff to keep up with visitor demand for information about park regulations, recommendations for recreational activities, and logistical information. NPS would likely benefit from a chatbot that visitors can ask questions to and receive answers in a matter of seconds, regardless of the time of day or availability of staff. This paper overviews the implementation of such a chatbot, NPS CHAT, experimenting with diverse NLP approaches towards developing an effective question-answering chatbot focusing on US National Parks. The paper finds that an OpenAI GPT Model outperforms alternate approaches on this task. The paper also offers additional areas for exploration to further adapt the chatbot to important NPS needs.

Methods

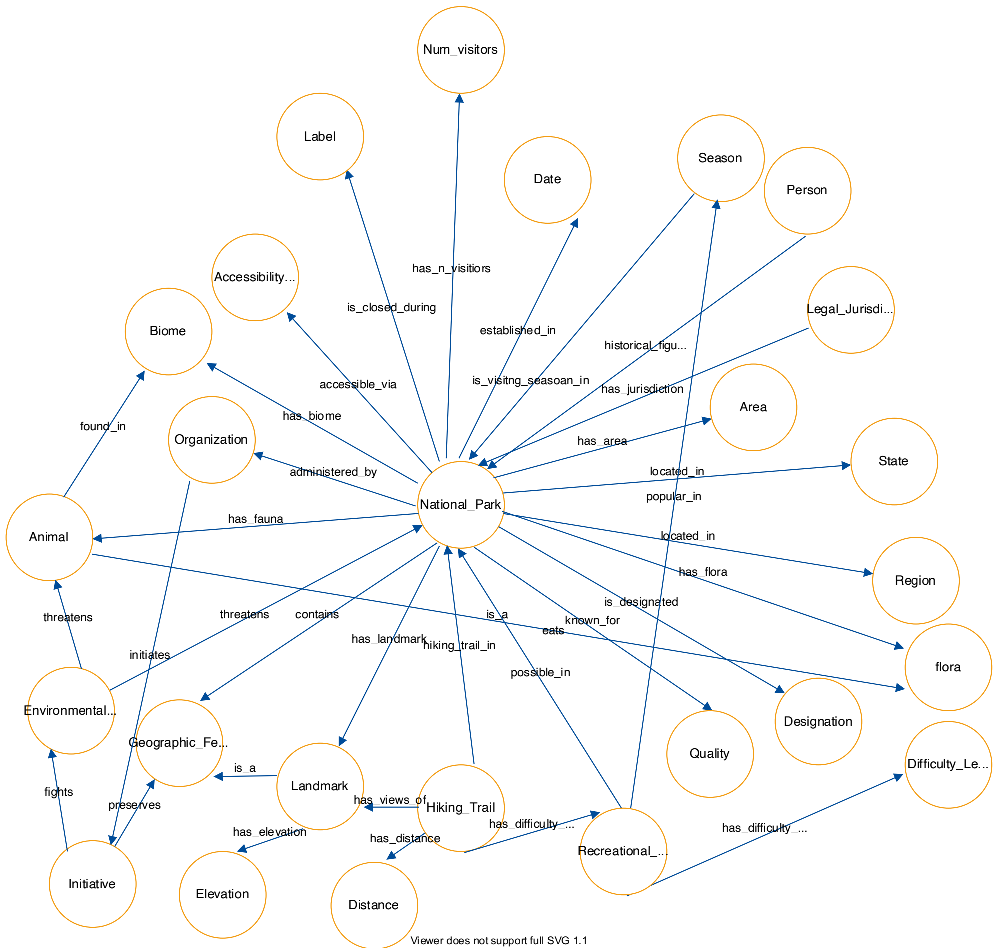
Initial Ontology

To aid in the selection of a dataset for the chatbot, an ontology was created to map out the types of entities and relationships that should be present in the dataset. To define the ontology, it was necessary to think about the following questions:

- “What is a National Park?”
- “What is National Park is made up of?”
- “Who manages the National Park?”
- “What is a National Park known for?”
- “What can one do in a National Park?”
- ...

With these questions in mind, an ontology was created. It is visualized in *Figure 1* and Appendix *A1*.

Figure 1: First-Pass Ontology for National Parks



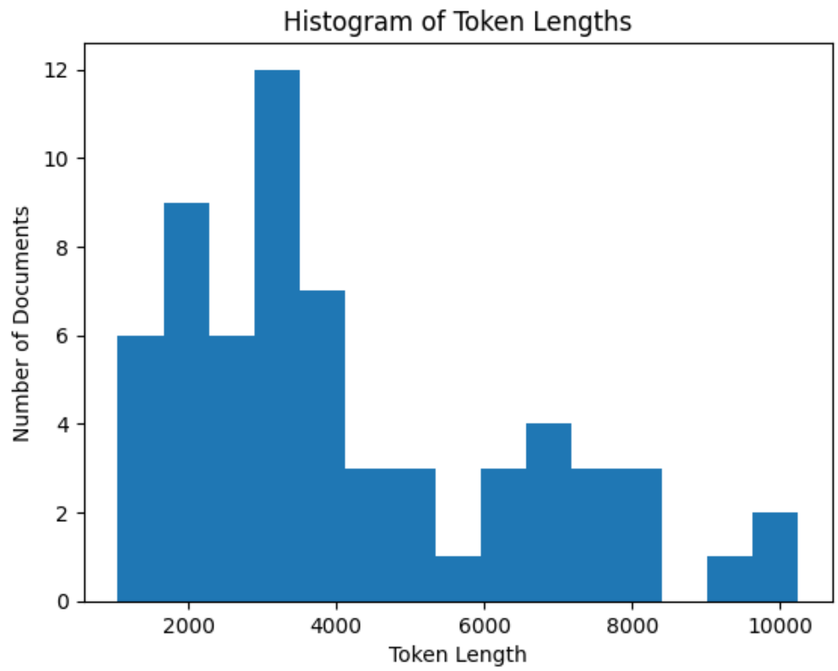
Dataset

Based on the types of entities and relations described in the ontology, it was determined that Wikipedia articles about each of the National Parks seemed to contain most of the entities and relationships modeled in the ontology.

To extract the text for the corpus, a list of the 63 national parks were collected from "List of National Parks of the United States" (Wikimedia Foundation, 2023). Python's Wikipedia module was then utilized to extract each Wikipedia page for every national park on the list. Next, special Wikipedia characters and new line characters were removed. After inspection of the texts, it was observed that many of the documents contained citations and other extraneous text at the. These pieces of text were removed. All of the national park texts were then compiled into a Python Pandas DataFrame.

The documents were rather long and were of varying lengths, spanning multiple thousand tokens each. *Figure 2* displays a histogram of token lengths for the documents within the corpus.

Figure 2. Histogram of Document Token Lengths in Corpus



There are inherent challenges working with documents with long sequence lengths. Many models have limits on sequence lengths that can be processed due to fixed size input layers in deep learning-based models. Other models allow long sequence lengths, but the quality of answers may degrade with extraneous context. A common workaround is to split up documents into smaller chunks, but this may hinder the ability to collect contextual information that may be important in question-answering systems. The decision of whether documents should be split up into chunks or if the context should include all of the documents in the corpus is best made on the constraints of the model and the data available.

For the purposes of the experiments, the decision of whether to chunk up these documents were left to the constraints of the models themselves. However, in an effort to ensure that sufficient context was provided to answer questions that may require context across multiple documents, the context fed to each model included all of the documents from across the entire corpus, regardless of how each document was tokenized.

TF-IDF Sentences with Cosine Similarity

The first approach was inspired by a Medium article "*Creating your own chatbot using python NLTK.*" (Dass, 2018). This first chatbot used a TF-IDF vectorizer to vectorize sentences in the documents into TF-IDF vectors. When the chatbot receives a query from the user, it vectorizes the query using the same vectorizer and returns the sentence from the dataset with the highest cosine similarity score to the query. This sentence is used as the chatbot's "response". Prior to performing this task, each document in the corpus was converted to lower case, had punctuation and stop words removed, and underwent lemmatization before being tokenized into sentences using NLTK's sentence tokenizer. The same preprocessing techniques are executed for the user's query before performing the cosine similarity comparison against the documents.

Sentence Transformers with Cosine Similarity

The Sentence Transformers with Cosine Similarity chatbot was similar in that it converted document sentences and user queries into vector embeddings and then returned the sentence with the highest cosine similarity score when compared to the vectorized query. For this chatbot, the documents were converted to lowercase and then tokenized using NLTK's sentence tokenizer. Instead of using a TF-IDF vectorizer, this approach uses a pre-trained Sentence Transformer model ('multi-qa-MiniLM-L6-cos-v1') for sentence embeddings. This transformer model was designed for semantic search purposes, mapping sentences and paragraphs to a 384-dimensional dense vector space. In particular, this model was trained on input text of up to 250 tokens, and has a 512-token limit, with any longer text truncated (Hugging Face, n.d.-a).

Pre-Trained Question Answering Model

The third chatbot utilized a pre-trained question-answering model from the Hugging Face Transformers library (Hugging Face, n.d.-b). The chatbot takes in a query from the user and runs it against a context. The model was trained to return the answer to the supplied question based on the context. While the implementation allowed either a single document or concatenated documents to be passed in as context to the query, the concatenated documents were selected for experimentation purposes to remain consistent with the other approaches that utilized context from across all documents.

OpenAI GPT

In this final approach – inspired by a Medium Article "*Creating your own AI-powered Second Brain: A Guide with Python and CHATGPT*" (Hafiz, 2023) – a GPT vector store index was created for each document in the corpus, using LLaMa's GPTVectorStoreIndex (Liu, 2022). This index is a representation of the text data in a format suitable for similarity searches using the GPT model. The GPT vector store index was then saved to disk in JSON format. Finally, a query engine was created from the previously created index, allowing for efficient querying of the GPT vector store index. The user then queries the query engine to receive a generated response based on the context from the corpus. No other preprocessing steps for the documents were conducted other than removing newline characters. It is also worth noting that this approach was not free – it required an OpenAI API key which charged about $0.01 per query (OpenAI, n.d.).

Quality Assessment Criteria

In order to produce a quality chatbot for users, the responses need to be accurate, helpful, contextual, support synthesis from multiple sources (if necessary), fast, and inexpensive.

Five areas were defined to assess the quality of answers provided by the different chatbots.

The first criteria, *information extraction*, seeks to test the chatbots' ability to answer questions merely using information with terms that are directly stated somewhere in the corpus. The ability to achieve this represents the most basic but necessary functionality of the chatbot.

The second area – *understanding context* – assesses the ability of the chatbot to go beyond the literal terms in the query. To perform well in this area, the chatbot must be able to use context and recognize semantically similar parts of the corpus to use as context for the response.

The third area – *providing a range of information* – evaluates the ability of the chatbot to return multiple pieces of information when necessary. Strong performance in this area requires the chatbot to combine multiple pieces of information from a document when necessary for crafting a complete and helpful response.

The fourth area – *retrieving relevant information across multiple documents* – measures the chatbots ability to synthesize information sourced from multiple documents. This is often important for queries that require knowledge from disparate topics.

The final area – *generative ability* – examines how well the responses are formatted and articulated. This is important to make sure that the chatbots' responses are coherent, grammatically correct, and well-articulated using the same quality of speech expected from a professional.

Figure 3 provides a table of these five areas along with a corresponding question designed to test performance in these areas.

Figure 3: Quality Assessment Criteria and Questions

Criteria	***Question for Assessment***
Information extraction	*"When was Yellowstone National Park established?"*
Understanding context	*"What factors make the Grand Canyon unique?"*
Providing a range of information	*"What activities are available in Shenandoah National Park?"*
Retrieving relevant information across multiple documents	*"Which national parks are located in California?"*
Generative ability	*"Explain the potential effects of climate change on Glacier National Park's glaciers and ecosystems. How can the park adapt to these changes?"*

All four chatbots were evaluated on these five questions. When evaluating chatbot responses, the responses were identified as either “Successful”, “Partial”, or “Failure” and are represented in the results table (*Appendix A2*) as green, yellow, and red respectively. A column is also provided to display the average response time across the five queries for each chatbot approach.

Results

For reference, the chatbots’ responses to each of the questions are available in *Appendix A2*.

The TF-IDF Cosine Similarity Chatbot seemed to struggle across almost all questions. It answered the first three questions and the fifth question incorrectly with answers that didn’t make any sense. It achieved a partial success on the question “Which national parks are located in California”, where it correctly identified Redwood National Park but failed to name any of the other eight national parks located in California. Despite the low-quality answers, this chatbot’s average response time was relatively quick at 7.57 seconds per query.

The Sentence Transformer Cosine Similarity Chatbot performed better. It correctly identified when Yellowstone National Park was established (March 1, 1872) and correctly responded to the “What activities are available in Shenandoah National Park?” question mentioning its status as a “national scenic byway” and its offering of “196,000 acres of backcountry and wilderness camping”. It was partially successful at answering the question “Explain the potential effects of climate change on Glacier National Park's glaciers and ecosystems. How can the park adapt to these changes?”, mentioning research efforts to attain a broader understanding of climate changes in the park. However, it also failed to identify the potential effects of climate change. It failed miserably at the questions “What factors make the Grand Canyon unique?” and “Which national parks are located in California?” responding “the national parks: america's best idea: parks-grand canyon” and “the national parks” to these questions respectively. This approach had the fastest average response time at 6.53 seconds per query.

The Transformers Question Answering Pipeline performed poorer than expected. It incorrectly responded “1947” to “When was Yellowstone National Park established?”, “increased visitation and traffic congestion” to “What factors make the Grand Canyon unique?”, “farming or mining” to “What activities are available in Shenandoah National Park?”, and “everglades” to “Which national parks are located in California?”. It also failed to correctly identify the disappearing glaciers in Glacier National Park. All of the responses from this chatbot were too concise – every response was either a single word or a short phrase. Overall, this method performed the worst out of all of the chatbots – its responses were not helpful or accurate, and its average response time at 26.75 seconds was more than three times as long as any of the other three chatbots.

The OpenAI GPT Model was by far the most performant model. It successfully and thoroughly answered all of the questions, with the exception of the fourth question, “Which national parks are located in California?”, to which it responded, “Pinnacles National Park and Yosemite National Park are located in California.” While the response mentioned two of the nine national parks in California (more than any other chatbot), it missed the other seven. This result was rather surprising for a model using GPT, a massive large language model (LLM) with 175 billion parameters. Regardless, this chatbot produced superior performance in all criteria areas.

Furthermore, this chatbot produced well-crafted responses that coherently answered questions in full sentences. This chatbot also responded quickly with an average response time of just 8.13 seconds.

Analysis

The performance of the TF-IDF Cosine Similarity chatbot was not surprising. This method was the most simplistic – it has very limited contextual ability and is best suited for answering questions where the answer to the question shares many of the exact terms in the query. Due to this method's selection of the sentence with the highest cosine similarity score, this chatbot is unable to fuse together multiple sentences to answer questions when required. It is also not based on a generative model that can output well-crafted responses from the context. The performance of this chatbot also struggled in part to its use of sentences from all documents, as the TF-IDF vectors were constructed across all sentences in the entire corpus. This might have made it more difficult for this approach to retrieve relevant information and could have also introduced more noise. The approach may benefit from an initial cosine similarity comparison between the query and each national park document to find the most relevant document, using the designated document's TF-IDF vectorized sentences to retrieve more relevant and focused answers.

The Sentence Transformer Cosine Similarity technique saw increased performance relative to the previous method due to its richer semantic capabilities, boosted by its dense embedding space after having been trained on 215 million question answering pairs (Hugging Face, n.d.-a). This improved semantic understanding of language allowed it to find answers from the corpus that required semantic awareness. Its slightly improved average response time was likely due to fewer preprocessing techniques being employed or faster cosine similarity calculations due to a lower dimensional embedding space than the TF-IDF vectors. That said, this method likely suffered similarly to the previous technique in that using the entire corpus made retrieving relevant information more difficult due to broader scope and additional noise. Similarly, this method used sentence tokenizers, making it impossible for this method to produce answers that synthesized multiple sources together, an important task for some questions, like "Which national parks are located in California". Furthermore, this method merely returns the sentence with the highest cosine similarity compared to the query, instead of generating new text that summarizes the answer or rephrases the answer in full sentences.

The performance of the Transformers Question Answering Pipeline was also likely hurt by the concatenation of all of the documents into one context. Hugging Face even details that this model was "trained on input text up to 250-word pieces" (Hugging Face, n.d.-b) which has surely been exceeded in using the whole corpus. The model underlying this approach relies on extracting an answer from the text that answers the question. With such a large context, it becomes difficult to assess relevance, especially when there are many answers that may answer the question. This is evidenced by the incorrect answers that could be reasonable answers, but they are likely based on information that has been sourced from irrelevant documents. For instance, the chatbot answered "1947" for the question "When was Yellowstone National Park established?". This sounds like a reasonable answer, but it is not correct. It is likely that this answer was sourced from parts of the corpus containing information about Theodore Roosevelt National Park and Everglades National Park, which were both established in 1947. It would be interesting to vectorize the query and documents individually in order to perform a cosine similarity search to determine the most relevant document to query. That document can then be used alone as the only context for the Transformers Question Answering Pipeline and may yield superior performance. This model would also likely benefit from fine-tuning, but this was not possible given the lack of labeled datasets for this task. Regardless, this approach will never yield the generative ability provided by the OpenAI GPT Chatbot.

If one of these chatbots was to be deployed immediately to serve NPS's needs, the OpenAI GPT Chatbot would be the clear choice. This model produced superior results in almost every area, and it even mentioned that it could not find an answer given the current context if the context did not supply the information necessary to answer the question. Its ability to generate accurate, helpful, and full-sentence answers to questions makes this chatbot the most desirable for users expecting staff-level answers to their questions. Additionally, the OpenAI GPT chatbot was able to accommodate the entire corpus, allowing it access to more information that useful for answering questions that require information synthesis. However, this model is still imperfect and can be improved – its inability to name the other seven national parks in California illustrated the limitations of the LLM's ability to provide complete and accurate information.

Additionally, it is expected that as information is added or updated about the parks (closures of trails, environmental changes, or new amenities), NPS will need to ensure that this new information is reflected in its answers. The OpenAI GPT Chatbot is based on data taken from a snapshot in time, and information that is routinely updated will necessitate the need to recreate the index. This is a less desirable option.

Future work may investigate implementing a Retrieval-Augmented Generative (RAG) AI system to provide a more accurate information base that can be queried to find relevant and up-to-date data to answer questions while supporting dynamic content updates.

A RAG-AI system might utilize an LLM to generate cypher code to query a knowledge graph (constructed from a large corpus on subject matter). The chatbot could then query the database to return the data, which can then be passed back to an LLM to summarize into a paragraph. That paragraph can then be used as the context needed to answer the question. Developing and maintaining a robust knowledge graph is extremely difficult and time-consuming, but the initial ontology built may serve as a launching point to begin thinking about how the knowledge graph will be structured and which entities and relationships it must include. Creating an ontology could also help identify gaps in the knowledge graph where data from additional sources could be of use.

It may also be beneficial to integrate a user feedback loop to remember information provided by the user during a chat session. Awareness of a user's previous queries may help in gaining additional context to provide relevant answers.

Conclusion

In conclusion, the OpenAI GPT Model emerged as the most promising choice for deployment, given its superior performance in providing accurate, contextual, and generative responses. The TF-IDF Cosine Similarity Chatbot struggled with contextual understanding and information extraction, providing incoherent responses. The Sentence Transformer Cosine Similarity Chatbot faced challenges in synthesizing information from multiple sources and returned inaccurate answers for specific queries. The Transformers Question Answering Pipeline experienced difficulties in relevance assessment, offering concise and often inaccurate responses, exacerbated by the concatenation of all documents into a single context. Despite the promise of the OpenAI GPT Chatbot, ongoing efforts are needed to refine and update the underlying knowledge base to ensure the chatbot's effectiveness in addressing evolving information about the national parks. NPS CHAT stands as a valuable step towards enhancing visitor engagement and information accessibility within the dynamic landscape of US National Parks.

References

Bennett, G. (2022, April 20). *National parks are swamped, but the park service faces mountainous employment crises*. OnLabor. https://onlabor.org/national-parks-are swamped-but-the-park-service-faces-mountainous-employment-crises/

Dass, R. (2018, September 24). *Create your chatbot using python NLTK*. Medium. https://medium.com/@ritidass29/create-your-chatbot-using-python-nltk-88809fa621d1

Hafiz, I. (2023, April 3). *Creating your own AI-powered Second Brain: A Guide with Python and CHATGPT*. Medium. https://levelup.gitconnected.com/creating-your- own-ai powered-second-brain-a-guide-with-python-and-chatgpt-f5547ef7e136

Hugging Face. (n.d.-a). *Sentence-transformers/multi-qa-minilm-L6-cos-v1 · hugging face*. sentence-transformers/multi-qa-MiniLM-L6-cos-v1 · Hugging Face. https://huggingface.co/sentence-transformers/multi-qa-MiniLM-L6-cos-v1

Hugging Face. (n.d.-b). *Pipelines*. Hugging Face Transformers. https://huggingface.co/docs/transformers/index

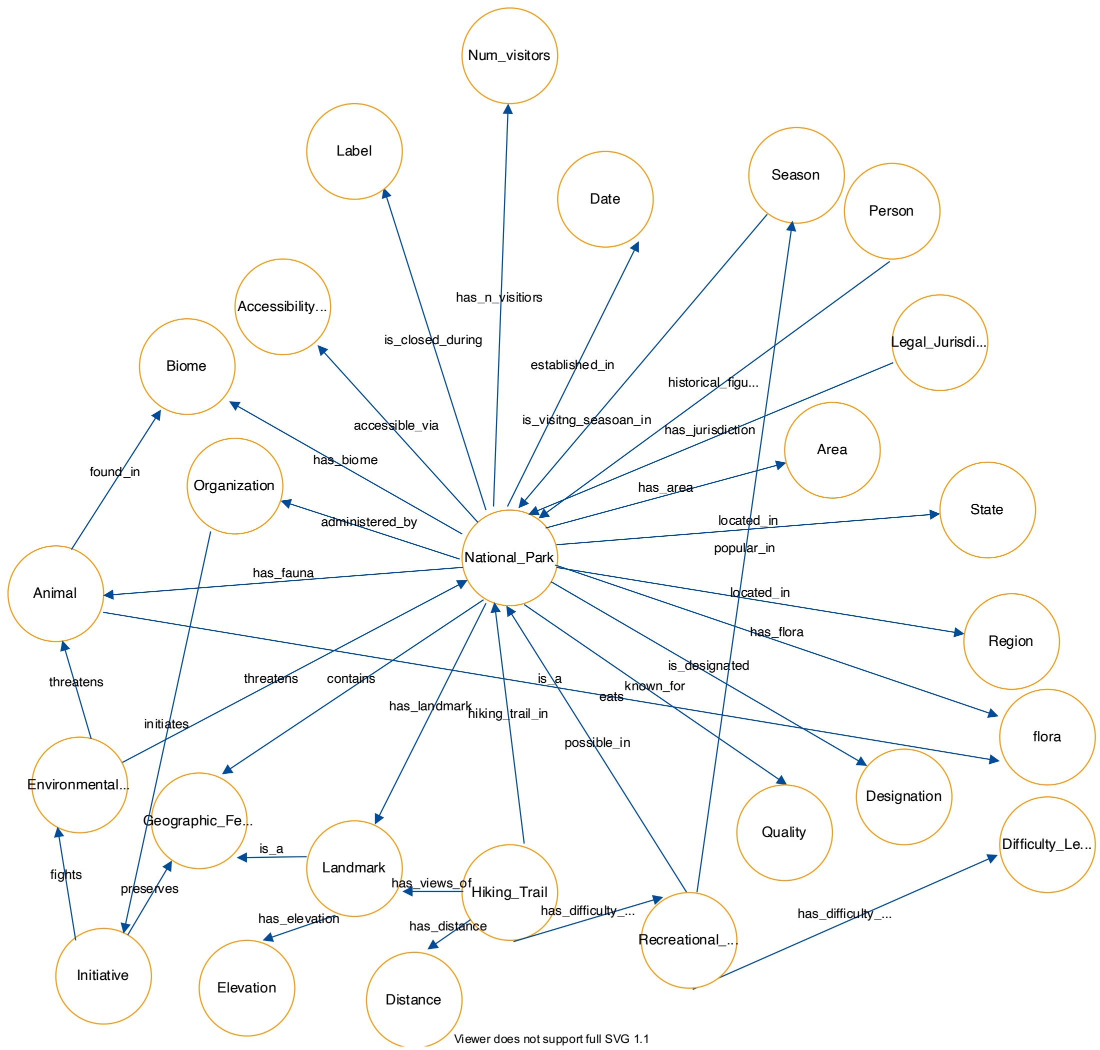
Jerry Liu. (2022). *Vector Store index*. Vector Store Index - LlamaIndex 🦙 0.9.11. https://docs.llamaindex.ai/en/stable/module_guides/indexing/vector_store_guide.html

OpenAI. (n.d.). *API reference - openai API*. API Reference - OpenAI API. https://platform.openai.com/docs/api-reference

Wikimedia Foundation. (2023, November 27). *List of national parks of the United States*. Wikipedia. https://en.wikipedia.org/wiki/List_of_national_parks_of_the_United_States

Appendix

A1. Initial Ontology



A2. Chatbot Responses

Chatbot Response	Questions					
	“When was Yellowstone National Park established?”	**“What factors make the Grand Canyon unique?”**	**“What activities are available in Shenandoah National Park?”**	**“Which national parks are located in California?”**	**“Explain the potential effects of climate change on Glacier National Park's glaciers and ecosystems. How can the park adapt to these changes?”**	**Average Response Time in Seconds**
TF-IDF Sentence Cosine Similarity	the yellowstone national park archives maintain collections of historical records of yellowstone and the national park service.	grand canyon association.	the results of long's residency were featured in the photography exhibit "wild beauty: the artful nature of shenandoah national park" held at the looking glass art gallery in the historic hawley silk mill, in hawley, pennsylvania.list of amphibians of shenandoah national parklist of birds of shenandoah national parklist of national parks of the united statesofficial website of the national park serviceshenandoah national park— when past is present: archaeology of the displaced in shenandoah national parknasa earth observatory satellite images of shenandoah national park and park's vicinityunited states geological survey: geologic map of the shenandoah national park region, virginiahenry heatwole's guide to shenandoah national park and skyline driveinteractive virtual tours of shenandoah national parkthe ground beneath our feet online exhibit of the virginia historical society regarding creation of the shenandoah national parka guide to the shenandoah valley oral history project, 2005–2006	he redwood national and state parks (rnsp) are a complex of one national park and three california state parks located in the united states along the coast of northern california.	the park's largest glacier is bear glacier.	7.57
Sentence Transformer Cosine Similarity	it was established by the 42nd u.s. congress with the yellowstone national park protection act and signed into law by president ulysses s. grant on march 1, 1872. yellowstone was the first national park in the u.s. and is also widely held to be the first national park in the world.	the national parks: america's best idea: parks-grand canyon.	it is also designated as a national scenic byway.shenandoah national park offers 196,000 acres (306.2 sq mi; 793.2 km2) of backcountry and wilderness camping.	the national parks.	the research compiled contributes to a broader understanding of climate changes in the park.	6.53
Transformers Question Answering Pipeline	1947	increased visitation and traffic congestion	farming or mining	everglades	negatively affected food supplies throughout the continent	26.75
OpenAI GPT	Yellowstone National Park was established by the 42nd U.S. Congress and signed into law by President Ulysses S. Grant on March 1, 1872.	The Grand Canyon is unique due to several factors. Firstly, its size and depth set it apart from other canyons in the world. It is a massive gorge of the Colorado River, stretching for miles and reaching depths of over a mile. Additionally, the exposed layers of colorful rocks dating back to Precambrian times make it a geological wonder. The combination of size, depth, and the variety of rock formations make the Grand Canyon a truly unique natural feature.	In Shenandoah National Park, there are several activities available for visitors. These include hiking on over 500 miles of trails, including a section of the Appalachian Trail. Camping is also popular, with five major campgrounds open from April to October-November. Horseback riding, bicycling, and fly fishing in the streams and rivers for native brook trout are other activities that can be enjoyed in the park. Additionally, there are lodges and cabins available for accommodation, and visitors can explore the park's waterfalls and participate in ranger-led programs to learn about the park's history, flora, and fauna.	Pinnacles National Park and Yosemite National Park are located in California.	The potential effects of climate change on Glacier National Park's glaciers and ecosystems include the loss of habitat for plant and animal species that rely on cold water, reduced stream flow during dry seasons, lower water table levels leading to an increased risk of forest fires, and a decrease in the aesthetic appeal that glaciers provide to visitors. These changes can have a significant impact on the park's ecosystems and overall biodiversity. To adapt to these changes, Glacier National Park can take several measures. One approach is to continue monitoring and researching the impacts of climate change on the park's glaciers and ecosystems. This can help in understanding the specific vulnerabilities and developing targeted conservation strategies. Additionally, the park can focus on promoting sustainable practices and raising awareness about climate change among visitors. This can include educational programs, interpretive signage, and outreach initiatives to inform visitors about the importance of preserving the park's ecosystems and reducing their own carbon footprint. Furthermore, the park can collaborate with other organizations and agencies to implement climate change mitigation strategies. This can involve reducing greenhouse gas emissions, promoting renewable energy sources, and advocating for policies that support climate action. Overall, adapting to the effects of climate change on Glacier National Park's glaciers and ecosystems requires a combination of research, conservation efforts, and public engagement to ensure the long-term preservation of this unique and valuable natural resource.	8.13