Conservation Encyclopedia: Large Language Models for Text Summarization and Question Answering

Rosa Wu
American Heritage School
MehtA+

Naomi Morato
Skyline High School
MehtA+

Jayani Mannam
Northview High School
MehtA+

Noah Jacob
Carlmont High School
MehtA+

July 28, 2023

Abstract

With the exponential rise in research being published each day, environmental conservationists around the world are facing an issue with their ability to keep up. For those working on the ground, reading through dense research papers is an inhibiting factor upon their ability to conduct conservation work efficiently. Thus, a decision support tool is needed to quickly gather best practices and next steps. Our goal is to create a text summarizer and question-answering system for The Integrative Conservation Clinic, a non-profit online conservation encyclopedia. Fortunately, Large Language Models, an artificial intelligence model that can process language data and perform tasks such as text comprehension, generation, and summarization, can be used to accomplish this goal. In our study, Machine Learning is used to offer answers to common questions in terms of creating article summaries and providing responses to user-prompted questions.

1 Introduction

There are roughly 32,310 conservationist scientists and foresters in the United States. Those in this field of work manage and check the quality of the forest, soil, and environment around them either commercially or locally. This number does not include the number of other conservationists working to protect the environment and ecosystems as well. One of the main problems for these conservationists is the inaccessibility of information. This huge challenge was taken on by The Integrative Conservation Clinic at The College of William and Mary.

This non-profit project is a website designed to connect conservationists with the tools, information, and people needed to support informed actions. With this spread of new and critical information, we can expect more effective conservation. To set this plan into action and at a faster speed, we used Large Language Models and Natural Language Processing to help create this database for conservationist knowledge.

Natural Language Processing, or NLP, is a field of artificial intelligence and computational linguistics that focuses on the interaction between computers and human language. The primary goal of NLP is to enable computers to understand, interpret, generate, and respond to human language in a way that is meaningful and contextually relevant.

NLP involves a wide range of tasks and applications. This includes tasks like sentiment analysis, topic modeling, and named entity recognition. However, two of the most notable uses are text summarization and question answering.

Large Language Models, or LLMs, are a specific type of artificial intelligence model that is trained on a massive amount of text data to generate human-like language. These models use deep learning techniques and neural networks to process language data and can perform various NLP tasks. In summary, LLMs are a type of model used within the broader field of NLP to achieve tasks related to language understanding and generation.

This is useful as it allows computers to analyze vast amounts of text that would be unmanageable for humans to handle manually. It enables the automation of tasks that involve text processing, leading to increased productivity and reduced human effort. Overall, NLP plays a crucial role in bridging the gap between human language and computers, opening up a wide array of practical applications and significantly impacting various industries, including healthcare, finance, customer service, education, and more.

By summarizing dense research articles and finding answers in those summaries to specific questions using Machine Learning, we can ensure that the long search for relevant information in the field of conservation is shortened.

2 Related Work

There have been other publications that have attempted to use machine learning for conservation efforts. Fernandes (2020) [**?**] investigated how machine learning, specifically the use of RStudio, could extract predictors' values where the wildlife data points were located and merged this list with the existing wildlife data frame. The final model was able to predict sustainable future sites that animals could relocate to if the animals' ecosystem ever became inhabitable. Lapeyrolerie (2022) [**?**] conducted a similar encyclopedia effort, researching whether deep reinforcement learning could help humans make better conservation decisions. Our model expands on these concepts and introduces an innovative approach to creating an encyclopedia with an integrated question-answering system. Alongside concise conservation articles, this system aims to offer conservationists enhanced access to valuable information while encouraging them to contemplate essential issues pertaining to their studies. By combining these elements, our model empowers conservationists with a comprehensive resource for their research and decision-making processes.

3 Methodology

3.1 Dataset

We used two datasets: one containing 400 articles on various conservation topics, and the other containing a set of 97 smaller, mini-articles. In addition, 7 articles on Yellowstone conservation were used to train the document similarity model.

Preprocessing To preprocess the data, we converted all article PDFs to text files. Then, we lemmatized and tokenized the summaries to input into the cosine similarity vectorizer model.

3.2 Model

The text summarization model we employed is called Pegasus-X-Large, known for its abstractive summarization capabilities. It features a distinctive encoder-decoder architecture, enabling it to identify crucial sentences within an input document and generate a single output sequence by masking those sentences. Pegasus is based on the transformer model, a type of neural network that excels at understanding context and meaning by analyzing relationships within sequential data, such as the words in a sentence.

To prepare our files for input into Pegasus, we developed a code that facilitates the conversion of PDF files to text files. Our Pegasus model first had a layer that tokenized the text, which sets up the model for sequence-to-sequence generation tasks using the "transformers" library, which prepares it for future tasks like text summarization.

Once we imported the text files, we created and configured a summarization pipeline. We then processed the text from the PDF and joined the lines together in a string. Finally, we defined the parameters for the text generation model and split the CUDA memory into smaller chunks to avoid memory errors. Following all of these steps, the model generated a short paragraph summary of the inputted article.

A database containing 7 articles with a focus on Yellowstone conservation was compiled. To assess the relevance of each document to two specific themes, namely "species relocation" and "plastic pollution," we utilized a TF-IDF vectorizer and a cosine similarity matrix. This process allowed us to calculate the similarity of each article to the chosen themes. By assigning similarity values to each article for the two themes, we formulated a set of targeted questions based on these themes.

To generate answers to these questions, we used a Large Language Model, or LLM, called DistilBERT. DistilBERT is a sophisticated language model capable of generating contextual responses and providing detailed information in response to specific inquiries related to the topics of interest, "species relocation" and "plastic pollution." This approach enables us to gain valuable insights from the articles.

4 Results

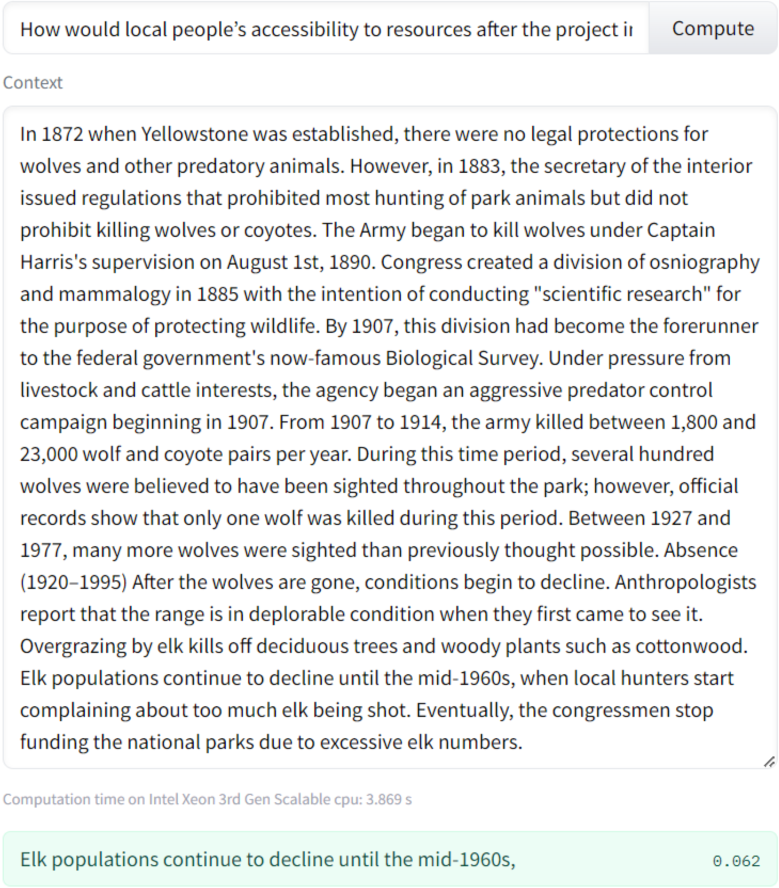
We tried to make the question as broad as possible so that the model could give us multiple answers to our question After analyzing the article, we found that the DistilBERT delivered the most substantial and accurate answers to the questions. It's also a smaller model, and thus is more inexpensive to use. use.

This project presented unique challenges that required innovative solutions to overcome. The abundance of information and the nature of vague questions hindered the LLMs' ability to provide accurate responses.

To address this, we devised a sophisticated approach involving the creation of a document similarity matrix. By implementing this matrix, we could effectively assess an article's relevance to a particular theme, enabling us to craft more targeted and precise questions based on the article's content.

Furthermore, we recognized the importance of addressing computational constraints. Working with large datasets and complex algorithms in Google Collab often led to CUDA memory errors, impeding our progress. We had to use memory management techniques to limit the amount of data on the GPU.

To tackle this issue, the clinic may want to potentially upgrade to Collab Pro, which will significantly alleviate these computational limitations. This is still better than GPT as it's more inexpensive.



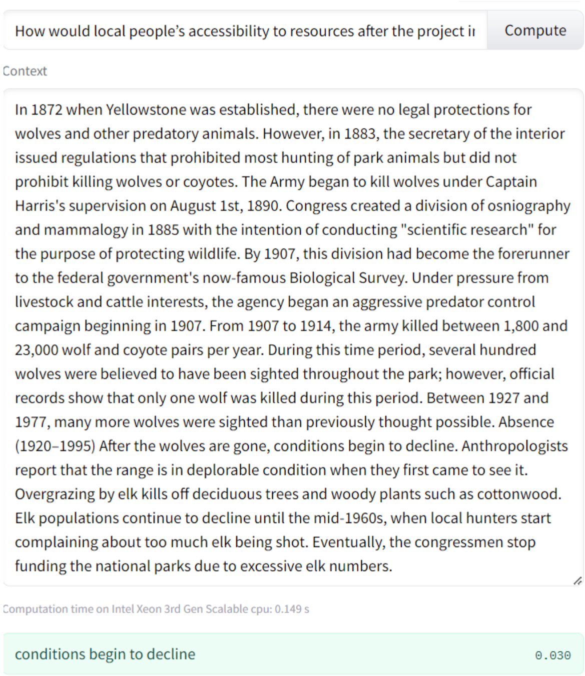
How would local people’s accessibility to resources after the project i

In 1872 when Yellowstone was established, there were no legal protections for wolves and other predatory animals. However, in 1883, the secretary of the interior issued regulations that prohibited most hunting of park animals but did not prohibit killing wolves or coyotes. The Army began to kill wolves under Captain Harris's supervision on August 1st, 1890. Congress created a division of osniography and mammalogy in 1885 with the intention of conducting "scientific research" for the purpose of protecting wildlife. By 1907, this division had become the forerunner to the federal government's now-famous Biological Survey. Under pressure from livestock and cattle interests, the agency began an aggressive predator control campaign beginning in 1907. From 1907 to 1914, the army killed between 1,800 and 23,000 wolf and coyote pairs per year. During this time period, several hundred wolves were believed to have been sighted throughout the park; however, official records show that only one wolf was killed during this period. Between 1927 and 1977, many more wolves were sighted than previously thought possible. Absence (1920–1995) After the wolves are gone, conditions begin to decline. Anthropologists report that the range is in deplorable condition when they first came to see it. Overgrazing by elk kills off deciduous trees and woody plants such as cottonwood. Elk populations continue to decline until the mid-1960s, when local hunters start complaining about too much elk being shot. Eventually, the congressmen stop funding the national parks due to excessive elk numbers.

Computation time on Intel Xeon 3rd Gen Scalable cpu: 3.869 s

Elk populations continue to decline until the mid-1960s, 0.062

Figure 1: Results of the Flan T5 model



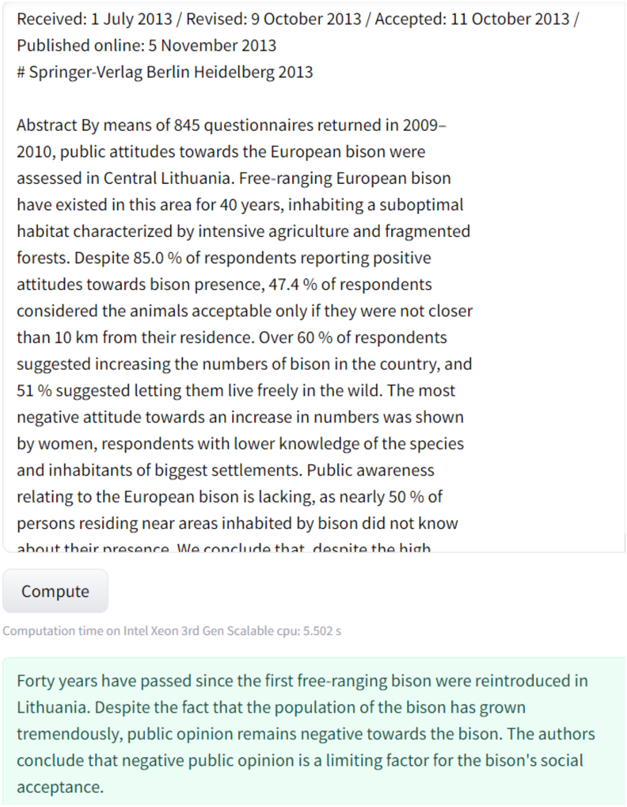
How would local people’s accessibility to resources after the project i

In 1872 when Yellowstone was established, there were no legal protections for wolves and other predatory animals. However, in 1883, the secretary of the interior issued regulations that prohibited most hunting of park animals but did not prohibit killing wolves or coyotes. The Army began to kill wolves under Captain Harris's supervision on August 1st, 1890. Congress created a division of osniography and mammalogy in 1885 with the intention of conducting "scientific research" for the purpose of protecting wildlife. By 1907, this division had become the forerunner to the federal government's now-famous Biological Survey. Under pressure from livestock and cattle interests, the agency began an aggressive predator control campaign beginning in 1907. From 1907 to 1914, the army killed between 1,800 and 23,000 wolf and coyote pairs per year. During this time period, several hundred wolves were believed to have been sighted throughout the park; however, official records show that only one wolf was killed during this period. Between 1927 and 1977, many more wolves were sighted than previously thought possible. Absence (1920–1995) After the wolves are gone, conditions begin to decline. Anthropologists report that the range is in deplorable condition when they first came to see it. Overgrazing by elk kills off deciduous trees and woody plants such as cottonwood. Elk populations continue to decline until the mid-1960s, when local hunters start complaining about too much elk being shot. Eventually, the congressmen stop funding the national parks due to excessive elk numbers.

Computation time on Intel Xeon 3rd Gen Scalable cpu: 0.149 s

conditions begin to decline 0.030

Figure 2: Results of the DistilBERT model



Received: 1 July 2013 / Revised: 9 October 2013 / Accepted: 11 October 2013 / Published online: 5 November 2013
Springer-Verlag Berlin Heidelberg 2013

Abstract By means of 845 questionnaires returned in 2009–2010, public attitudes towards the European bison were assessed in Central Lithuania. Free-ranging European bison have existed in this area for 40 years, inhabiting a suboptimal habitat characterized by intensive agriculture and fragmented forests. Despite 85.0 % of respondents reporting positive attitudes towards bison presence, 47.4 % of respondents considered the animals acceptable only if they were not closer than 10 km from their residence. Over 60 % of respondents suggested increasing the numbers of bison in the country, and 51 % suggested letting them live freely in the wild. The most negative attitude towards an increase in numbers was shown by women, respondents with lower knowledge of the species and inhabitants of biggest settlements. Public awareness relating to the European bison is lacking, as nearly 50 % of persons residing near areas inhabited by bison did not know about their presence. We conclude that, despite the high

Computation time on Intel Xeon 3rd Gen Scalable cpu: 5.502 s

Forty years have passed since the first free-ranging bison were reintroduced in Lithuania. Despite the fact that the population of the bison has grown tremendously, public opinion remains negative towards the bison. The authors conclude that negative public opinion is a limiting factor for the bison's social acceptance.

Figure 3: Results of the Pegasus model

5 Conclusion and Future Work

Our project was motivated by the pressing problem faced by environmental conservationists who often struggle with reading through dense research articles, which are continuously being published at a rapid rate. To address this issue, we set out to create a decision support tool that could efficiently generate article summaries and respond to user-prompted questions. Pegasus consistently provided the most concise summaries of the content, proving to be a valuable asset to our project. Our decision support tool, utilizing DistilBERT and BERT models, delivered substantial and accurate answers to the questions asked.

One significant advantage of our approach allows us to fine-tune the pipeline as needed. This not only provided us with greater control but also made the project more cost-effective for a non-profit organization like The Integrative Conservation Clinic, as we could not afford the expenses of querying ChatGPT's API. The ability to generate multiple answers to broad questions proved to be beneficial in providing comprehensive insights. We believe that this tool will be a valuable asset to environmental conservationists, streamlining the process of gathering best practices and next steps to address critical issues faced in the field.

6 Division of Labor

We divided the work as follows:

- Model: Naomi Morato, Jayani Mannam, Rosa Wu
- Poster: Jayani Mannam, Rosa Wu, Naomi Morato
- Paper: Rosa Wu, Jayani Mannam, Naomi Morato

7 Acknowledgements

We express our gratitude toward our Mentor, Dr. Alli Sabo from the College of William and Mary, for the opportunity to collaborate with The Integrative Conservation Clinic and contribute to their efforts in helping conservationists globally. In addition, we would like to thank our faculty mentors, Ms. Haripriya Mehta and Mr. Bhagirath Mehta, for their guidance and support throughout this project.

8 Bibliography

@articleLapeyrolerie, author = "Marcus Lapeyrolerie", title = "Deep reinforcement learning for conservation decisions, journal = "Research Gate", pages = "15", year = "2021", DOI = "10.1111/2041-210X.13954",

Pioneered in a paper by Lapeyrolerie[**?**]

@articleFernandes, author = "Marcus Lapeyrolerie", title = "Deep reinforcement learning for conservation decisions, journal = "Research Gate", pages = "15", year = "2021", DOI = "10.1111/2041-210X.13954",