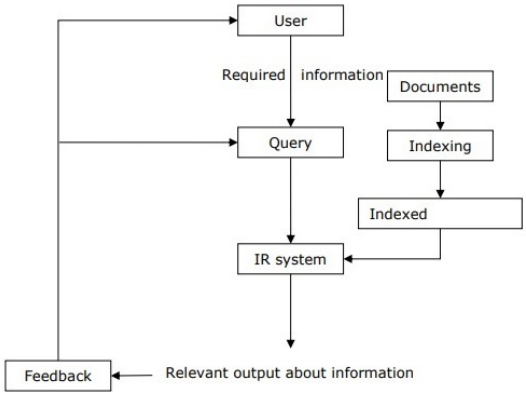
Abstractive-Question-Answering

Project Overview

In this project, we will utilize **Pinecone** to build an *abstractive question-answering system*. This system works by searching massive document stores for relevant information and then using this information to synthetically generate answers. We need three main components:



1. A vector index to store and run semantic search

Think of a **vector index** as a specialized filing cabinet designed to store and efficiently retrieve multidimensional data called vectors. These vectors are created by *embedding models*.

2. Retriever Models for Embedding Context Passages

A **retriever model** is a crucial component in question-answering systems and information retrieval tasks. Its primary function is to efficiently identify and retrieve the most relevant context passages from a large corpus of text, given a specific query.

A) Embedding:

- **Context Passages:** Each context passage is converted into a dense vector representation, or embedding. This embedding captures the semantic and syntactic meaning of the text.
- **Query:** The query is also transformed into a similar vector representation.

B) Similarity Search:

The retriever model calculates the similarity between the query vector and the embeddings of the context passages. This is typically done using cosine similarity.

C) Ranking and Retrieval:

The passages with the highest similarity scores are ranked and returned as the most relevant results.

3. A generator model to generate answers

A generator model is designed to generate text-like answers to questions. The model takes a prompt, such as a question or a partial answer, as input. The retriever retrieves relevant data by matching the prompt and vector embedding in the index. Then that data, along with the prompt, is used to generate an answer. If it doesn't have any information related to our prompt, it just generates some random text as output.

Load and Prepare Dataset

* Our raw source data will be taken from the Wiki Snippets dataset, which contains over 17 million passages from Wikipedia. However, since indexing the entire dataset may take some time, we will only utilize 50,000 passages that include "History" in the "section title" column.

* We are loading the dataset in streaming mode so that we don't have to wait for the whole dataset to download (which is over 9GB). Instead, we iteratively download records one at a time.


```
{'wiki_id': 'Q7649565',
 'start_paragraph': 20,
 'start_character': 272,
 'end_paragraph': 24,
 'end_character': 380,
 'article_title': 'Sustainable Agriculture Research and Education',
 'section_title': "2000s & Evaluation of the program's effectiveness",
 'passage_text': "preserving the surrounding prairies. It ran until March 31, 2001.\nIn 2008, SARE celebrated its 20th anniversary. To that date, the program had funded 3,700 projects
and was operating with an annual budget of approximately $19 million. Evaluation of the program's effectiveness As of 2008, 64% of farmers who had received SARE grants stated that
they had been able to earn increased profits as a result of the funding they received and utilization of sustainable agriculture methods. Additionally, 79% of grantees said that they
had experienced a significant improvement in soil quality though the environmentally friendly, sustainable methods that they were"}
```


* We will extract 'article_title', 'section_title', and 'passage_text' from each document and create a dataframe.

	article_title	section_title	passage_text
0	Taupo District	History	was not until the 1950s that the region starte...
1	Sutarfeni	History & Western asian analogues	Sutarfeni History strand-like pheni were Phena...
2	The Bishop Wand Church of England School	History	The Bishop Wand Church of England School Histo...
3	Teufelsmoor	History & Situation today	made to preserve the original landscape, altho...
4	Surface Hill Uniting Church	History	in perpetual reminder that work and worship go...

Initialize Pinecone Index

The Pinecone index stores vector representations of our historical passages which we can retrieve later using another vector (query vector). To build our vector index establish a connection with an API. Now we create a new index. We will name it "abstractive-question-answering". We specify the metric type as "cosine" and dimension as 768 because the retriever we use to generate context embeddings is optimized for cosine similarity and outputs 768-dimension vectors.

Initialize Retriever

The retriever will create embeddings such that the questions and passages that hold the answers to our queries are close to one another in the vector space. We will use a SentenceTransformer model based on Microsoft's MPNet as our retriever. This model performs quite well for comparing the similarity between queries and documents. We can use Cosine Similarity to compute the similarity between query and context vectors generated by this model (Pinecone automatically does this for us).

Generate Embeddings and Upsert

Next, we need to generate embeddings for the context passages. We will do this in batches to help us more quickly generate embeddings and upload them to the Pinecone index. When passing the documents to Pinecone, we need an id (a unique value), context embedding, and metadata for each document representing context passages in the dataset. The metadata is a dictionary containing data relevant to our embeddings, such as the article title, section title, passage text, etc.

```
{'dimension': 768,
 'index_fullness': 0.0,
 'namespaces': {'': {'vector_count': 49984}},
 'total_vector_count': 49984}
```

Initialize Generator

We will use ELI5 BART for the generator which is a Sequence-To-Sequence model trained using the 'Explain Like I'm 5' (ELI5) dataset. Sequence-To-Sequence models can take a text sequence as input and produce a different text sequence as output.

The input to the ELI5 BART model is a single string which is a concatenation of the query and the relevant documents providing the context for the answer. The documents are separated by a special token `<p>`, so the input string will look as follows:

question: What is a sonic boom? context: `<p>` A sonic boom is a sound associated with shock waves created when an object travels through the air faster than the speed of sound. `<p>` Sonic booms generate enormous amounts of sound energy, sounding similar to an explosion or a thunderclap to the human ear. `<p>` Sonic booms due to large supersonic aircraft can be particularly loud and startling, tend to awaken people, and may cause minor damage to some structures. This led to prohibition of routine supersonic flight overland.

```
query = "when was the first electric power system built?"
result = query_pinecone(query, top_k=1)
result
```

```
{'matches': [{'id': '3593',
              'metadata': {'article_title': 'Electric power system',
                           'passage_text': 'Electric power system History In '
                                           '1881, two electricians built the '
                                           "world's first power system at "
                                           'Godalming in England. It was '
                                           'powered by two waterwheels and '
                                           'produced an alternating current '
                                           'that in turn supplied seven '
                                           'Siemens arc lamps at 250 volts and '
                                           '34 incandescent lamps at 40 volts. '
                                           'However, supply to the lamps was '
                                           'intermittent and in 1882 Thomas '
                                           'Edison and his company, The Edison '
                                           'Electric Light Company, developed '
                                           'the first steam-powered electric '
                                           'power station on Pearl Street in '
                                           'New York City. The Pearl Street '
                                           'Station initially powered around '
                                           '3,000 lamps for 59 customers. The '
                                           'power station generated direct '
                                           'current and',
                           'section_title': 'History'},
              'score': 0.690402865,
```

Results

```
query = "what was the war of currents?"
context = query_pinecone(query, top_k=5)
query = format_query(query, context["matches"])
generate_answer(query)
```

```
('The War of Currents was a series of events in the early 1900s between Edison '
 'and Westinghouse. The two companies were competing for the market share of '
 'electric power in the United States')
```

```
query = "who was the first person on the moon?"
context = query_pinecone(query, top_k=10)
query = format_query(query, context["matches"])
generate_answer(query)
```

```
('The first person to walk on the moon was Neil Armstrong in 1969. He walked '
 'on the moon in 1969. He was the first person to walk on the moon.')
```

```
query = "what was NASAs most expensive project?"
context = query_pinecone(query, top_k=3)
query = format_query(query, context["matches"])
generate_answer(query)
```

```
('The Space Shuttle was the most expensive project in the history of NASA. It '
 'cost about $10 billion to build.')
```