Prompt-Based Annotated Bibliography Generator

Nick Selvitelli, Eoin Flynn, Sam Phillippo

ABSTRACT:

As the volume of scholarly literature continues to grow, finding relevant research articles for specific research topics becomes increasingly difficult. In this paper, we propose a potential solution to this problem, with our novel Natural Language Processing solution. We propose the Prompt-Based Annotated Bibliography Generator, that given a research topic of any length and complexity, will return an annotated bibliography containing relevant research articles. For this purpose, we built a large vector database, which contains SciBERT word embeddings of 55,000 documents retrieved from the CORE research paper database. Using this database, we are able to quickly retrieve the most similar research paper to a given user prompt. Our results indicate that this algorithm could be drastically improved, but are encouraging for potential future work.

1 INTRODUCTION:

The goal of our project is to allow users to enter a search prompt and return a bibliography of research articles citations related to the search. Our approach involved vectorizing a large dataset of research article titles, topics, and abstracts. We then store these vectorized articles along with their citation in a database. To use our application, we vectorize the users prompt and do a vector similarity search on the database to find the articles most similar to the users desired topic. While we ran into multiple issues throughout development, we were able to create a working demo which depending on the search, was able to produce an auto-generated bibliography of related articles.

2 METHODS:

The core system of the annotated bibliography generator works using word embeddings. The CORE dataset [1] contains the largest repository of freely available research articles and was used as the source of information to recommend. Our approach processed research article abstracts into vector embeddings that are then stored in a vector database for quick retrieval. The user provides a problem statement which is also converted into an embedding. The problem statement embedding is then used as a query within the vector database to find articles with similar embeddings.

The entire process of building the annotated bibliography generator is split into four main chunks: data preprocessing, generating embeddings, storing embeddings, and building the generator. Each topic is split into the following sections 2.1 through 2.4.

2.1 Data Preprocessing

The 2018 Metadata-only dataset from CORE was used as the source for articles used to build the annotated bibliographies. Details of the dataset are discussed in **Section 3**. The structure of the dataset proved to be unwieldy so a significant amount of preprocessing was needed. The dataset consists of over two-thousand compressed documents with widely varying sizes. Some files were too large to process on most machines so each file needed to be split into multiple smaller files before using for batch embedding generation. This chunking process was executed as a way to make the dataset more manageable before processing the data inside.

After chunking, each file was processed on a per document level. For each document in a file, key features were extracted. These features were then normalized and tokenized using the Python library NLTK [4].

2.2 Generating Embeddings

After the data preprocessing is complete for a single document, the normalized features are then used to generate a vector embedding to represent the article. The pretrained SciBert [2] word embeddings model was used to process each word. Each word embedding for a given document is then averaged together to produce a single embedding. This process proved to be unreliable as described in **Section 5**. These embeddings were processed with an NVIDIA Tesla V100 SXM2 on the Northeastern University Research Computing team Discovery cluster.

2.3 Storing Embeddings

After computing each document embedding from a file in the dataset, each embedding was inserted into a vector database paired with the corresponding article's citation and abstract. The "vectordb" database library [3] was used to store the embeddings locally. The database uses the Hierarchical Navigable Small World (HNSW) algorithm to compute Approximate Nearest Neighbor search between vectors using Euclidean distance (L2).

2.4 Building the Annotated Bibliography Generator

The actual bibliography generator works by generating an embedding for the given user input using the same process as described in **Section 2.2**. This embedding is then used as a query vector in the vector database. The top six nearest vector embeddings are returned and then formatted into an annotated bibliography response.

3 DATA

For this project, we used the CORE 2018 metadata only dataset. This dataset contained a comprehensive array of metadata on roughly 123 million scientific articles, books, and other documents. For our purposes, we only considered documents containing an abstract which reduced our dataset to 85 million articles. This data was stored across 5572 compressed json files and once uncompressed, each file ranged in size from a gigabyte to over 100 gigabytes.

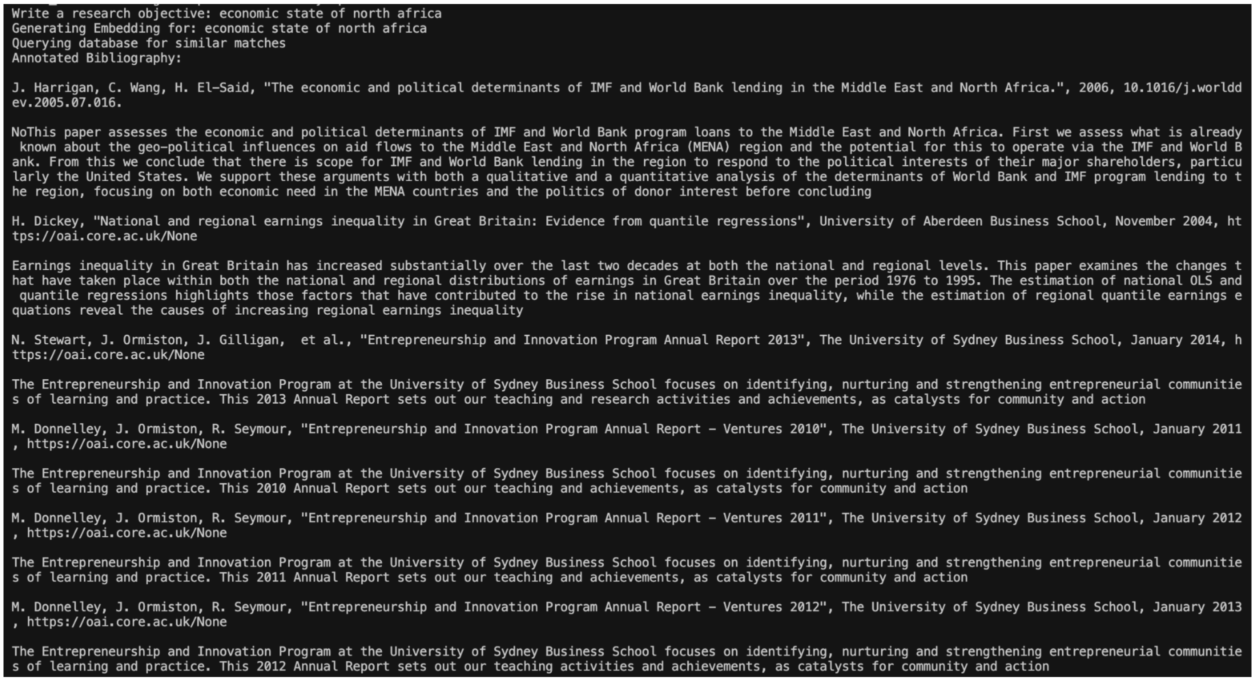
Since our objective only required the metadata necessary for citations and understanding the contents of the article, we worked with a small subset of each article's metadata. In each citation we used doi, title, authors, data_published, abstract, publishers, topic, and journals. If these features were not included we dropped the article from our search data. We also found situations where features were included but without any useful information. For example, we found multiple articles with abstracts that were only a couple words long. To correct for this, we tested each article's metadata to ensure features appear in useful formats before including it in the search data.

Due to time and resource constraints, we were only able to process 55 of the over 5000 json files. This equated to around 55,000 research articles being used in our final search dataset.

4 RESULTS

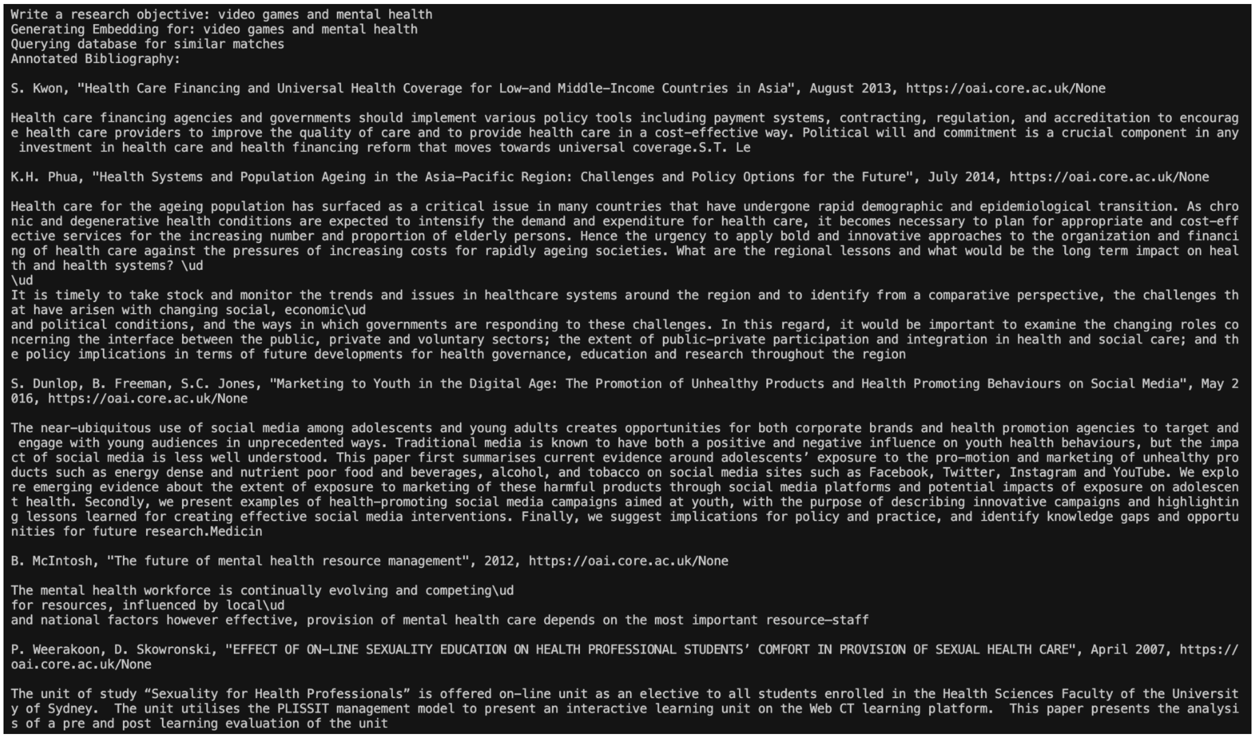
Due to the size of the dataset we used, a large portion of our efforts involved working around constraints we had in processing such a large quantity of data. While SciBERT only took a few seconds to vectorize each article, having this process any significant portion of our data would require either more computing resources or much faster vectorization. Even with using the Discovery GPU cluster to parallelize the vectorization process, only being able to vectorize 55,000 articles was not ideal. This would need to be scaled up in order to get more accurate citation results for users. We also believe the vectorization process itself did not allow for enough differentiation between articles leading to vectors of unrelated topics being seen as similar. This was apparent during testing where certain search terms would provide multiple articles directly related to the search topic while other searches would return article citations with no apparent relation to the topic or eachother.

In **figure 1** and **figure 2**, we have examples of citation results for the search topics “economic state of North Africa” and “video games and mental health”. These searches give promising results which are either directly or tangentially related to the topic. This could be further improved by simply increasing our citation search space. In **Figure 3**, we used the search term “health of hotdog”. This search failed and produced results with no relation. We believe this is because the vector created from the search term was most similar to the vectors classified as unknown in our database. This could be due to either no articles with similar context being available in the database or more likely the vectorization not being accurate enough to identify similarities between the search and articles available. In **Figure 4,** we used the search term “business in Australia”. While the resulting citations are related to business, this shows an issue that will grow worse as the database grows. We found that the CORE dataset contains many iterations of articles that are updated yearly. This causes search results to get clogged with articles that have virtually identical metadata. Fixing this would involve creating a metric that stops articles with vectors that are too similar from being shown all together.



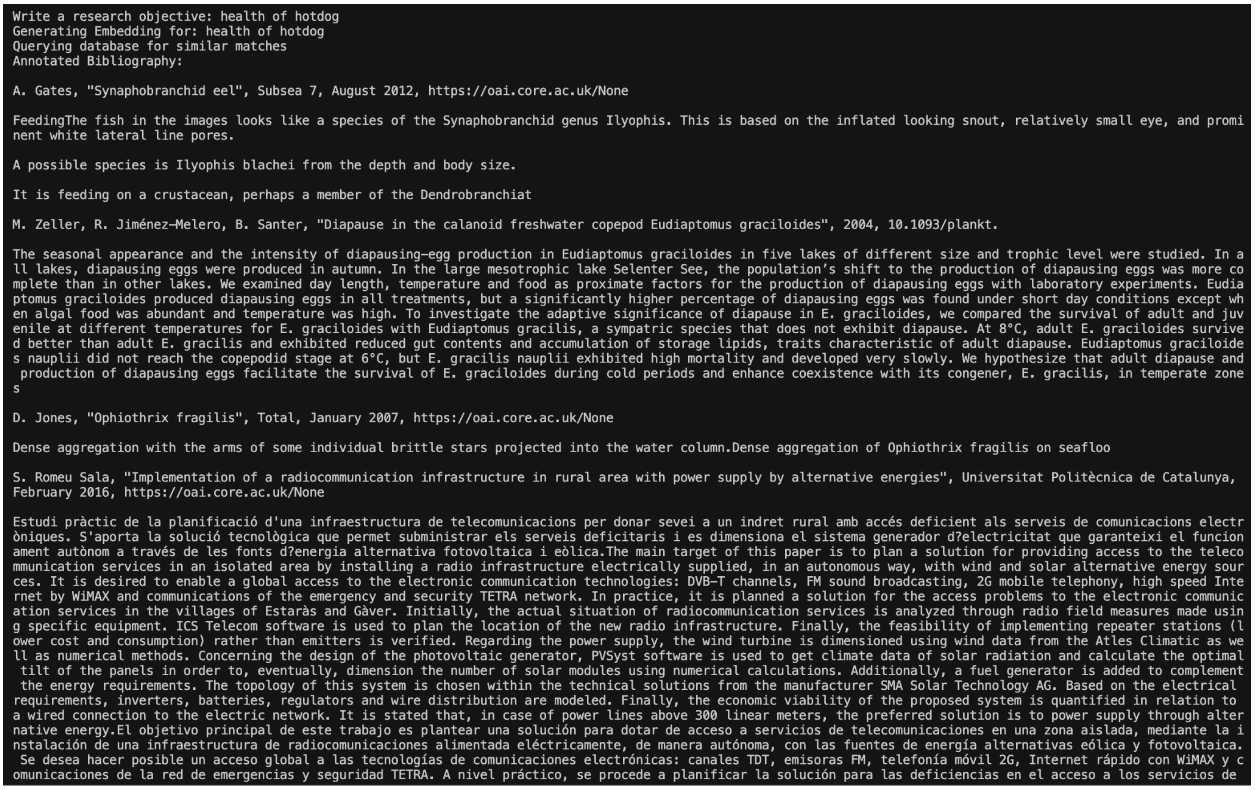
```
Write a research objective: economic state of north africa
Generating Embedding for: economic state of north africa
Querying database for similar matches
Annotated Bibliography:

J. Harrigan, C. Wang, H. El-Said, "The economic and political determinants of IMF and World Bank lending in the Middle East and North Africa.", 2006, 10.1016/j.worldd
ev.2005.07.016.

NoThis paper assesses the economic and political determinants of IMF and World Bank program loans to the Middle East and North Africa. First we assess what is already
 known about the geo-political influences on aid flows to the Middle East and North Africa (MENA) region and the potential for this to operate via the IMF and World B
ank. From this we conclude that there is scope for IMF and World Bank lending in the region to respond to the political interests of their major shareholders, particu
larly the United States. We support these arguments with both a qualitative and a quantitative analysis of the determinants of World Bank and IMF program lending to t
he region, focusing on both economic need in the MENA countries and the politics of donor interest before concluding

H. Dickey, "National and regional earnings inequality in Great Britain: Evidence from quantile regressions", University of Aberdeen Business School, November 2004, ht
tps://oai.core.ac.uk/None

Earnings inequality in Great Britain has increased substantially over the last two decades at both the national and regional levels. This paper examines the changes t
hat have taken place within both the national and regional distributions of earnings in Great Britain over the period 1976 to 1995. The estimation of national OLS and
 quantile regressions highlights those factors that have contributed to the rise in national earnings inequality, while the estimation of regional quantile earnings e
quations reveal the causes of increasing regional earnings inequality

N. Stewart, J. Ormiston, J. Gilligan,  et al., "Entrepreneurship and Innovation Program Annual Report 2013", The University of Sydney Business School, January 2014, h
ttps://oai.core.ac.uk/None

The Entrepreneurship and Innovation Program at the University of Sydney Business School focuses on identifying, nurturing and strengthening entrepreneurial communitie
s of learning and practice. This 2013 Annual Report sets out our teaching and research activities and achievements, as catalysts for community and action

M. Donnelley, J. Ormiston, R. Seymour, "Entrepreneurship and Innovation Program Annual Report - Ventures 2010", The University of Sydney Business School, January 2011
, https://oai.core.ac.uk/None

The Entrepreneurship and Innovation Program at the University of Sydney Business School focuses on identifying, nurturing and strengthening entrepreneurial communitie
s of learning and practice. This 2010 Annual Report sets out our teaching and achievements, as catalysts for community and action

M. Donnelley, J. Ormiston, R. Seymour, "Entrepreneurship and Innovation Program Annual Report - Ventures 2011", The University of Sydney Business School, January 2012
, https://oai.core.ac.uk/None

The Entrepreneurship and Innovation Program at the University of Sydney Business School focuses on identifying, nurturing and strengthening entrepreneurial communitie
s of learning and practice. This 2011 Annual Report sets out our teaching and achievements, as catalysts for community and action

M. Donnelley, J. Ormiston, R. Seymour, "Entrepreneurship and Innovation Program Annual Report - Ventures 2012", The University of Sydney Business School, January 2013
, https://oai.core.ac.uk/None

The Entrepreneurship and Innovation Program at the University of Sydney Business School focuses on identifying, nurturing and strengthening entrepreneurial communitie
s of learning and practice. This 2012 Annual Report sets out our teaching activities and achievements, as catalysts for community and action
```

Figure 1



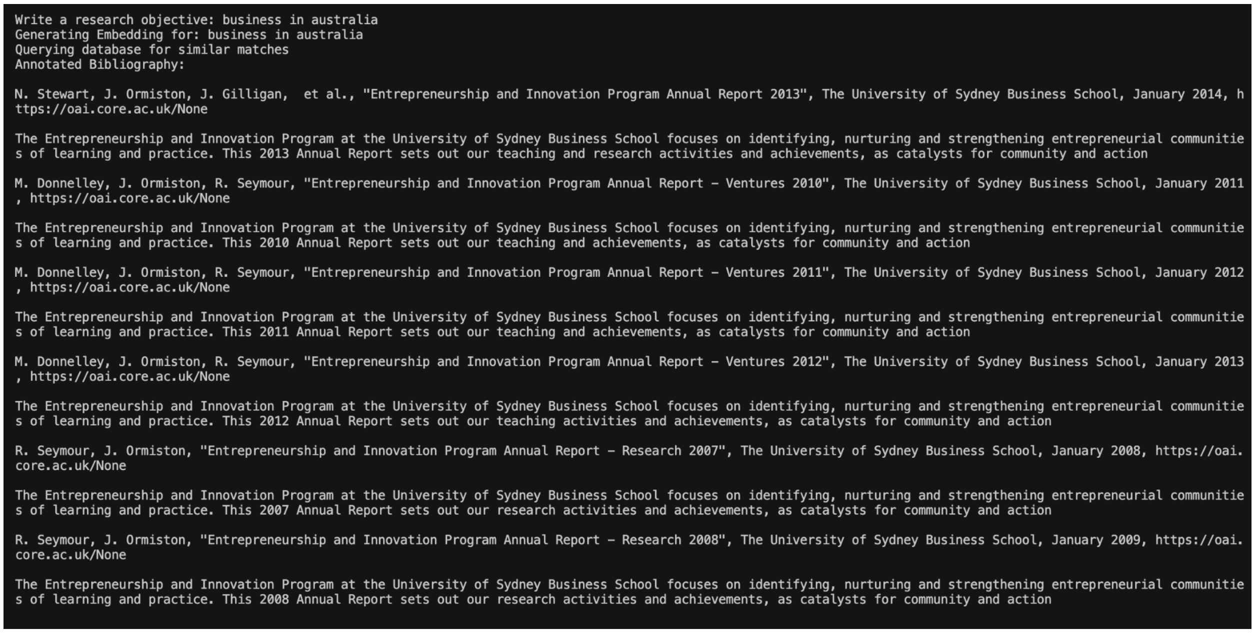
```
Write a research objective: video games and mental health
Generating Embedding for: video games and mental health
Querying database for similar matches
Annotated Bibliography:

S. Kwon, "Health Care Financing and Universal Health Coverage for Low–and Middle–Income Countries in Asia", August 2013, https://oai.core.ac.uk/None

Health care financing agencies and governments should implement various policy tools including payment systems, contracting, regulation, and accreditation to encourag
e health care providers to improve the quality of care and to provide health care in a cost–effective way. Political will and commitment is a crucial component in any
 investment in health care and health financing reform that moves towards universal coverage.S.T. Le

K.H. Phua, "Health Systems and Population Ageing in the Asia–Pacific Region: Challenges and Policy Options for the Future", July 2014, https://oai.core.ac.uk/None

Health care for the ageing population has surfaced as a critical issue in many countries that have undergone rapid demographic and epidemiological transition. As chro
nic and degenerative health conditions are expected to intensify the demand and expenditure for health care, it becomes necessary to plan for appropriate and cost–eff
ective services for the increasing number and proportion of elderly persons. Hence the urgency to apply bold and innovative approaches to the organization and financi
ng of health care against the pressures of increasing costs for rapidly ageing societies. What are the regional lessons and what would be the long term impact on heal
th and health systems? \ud
\ud
It is timely to take stock and monitor the trends and issues in healthcare systems around the region and to identify from a comparative perspective, the challenges th
at have arisen with changing social, economic\ud
and political conditions, and the ways in which governments are responding to these challenges. In this regard, it would be important to examine the changing roles co
ncerning the interface between the public, private and voluntary sectors; the extent of public–private participation and integration in health and social care; and th
e policy implications in terms of future developments for health governance, education and research throughout the region

S. Dunlop, B. Freeman, S.C. Jones, "Marketing to Youth in the Digital Age: The Promotion of Unhealthy Products and Health Promoting Behaviours on Social Media", May 2
016, https://oai.core.ac.uk/None

The near–ubiquitous use of social media among adolescents and young adults creates opportunities for both corporate brands and health promotion agencies to target and
 engage with young audiences in unprecedented ways. Traditional media is known to have both a positive and negative influence on youth health behaviours, but the impa
ct of social media is less well understood. This paper first summarises current evidence around adolescents’ exposure to the pro–motion and marketing of unhealthy pro
ducts such as energy dense and nutrient poor food and beverages, alcohol, and tobacco on social media sites such as Facebook, Twitter, Instagram and YouTube. We explo
re emerging evidence about the extent of exposure to marketing of these harmful products through social media platforms and potential impacts of exposure on adolescen
t health. Secondly, we present examples of health–promoting social media campaigns aimed at youth, with the purpose of describing innovative campaigns and highlightin
g lessons learned for creating effective social media interventions. Finally, we suggest implications for policy and practice, and identify knowledge gaps and opportu
nities for future research.Medicin

B. McIntosh, "The future of mental health resource management", 2012, https://oai.core.ac.uk/None

The mental health workforce is continually evolving and competing\ud
for resources, influenced by local\ud
and national factors however effective, provision of mental health care depends on the most important resource–staff

P. Weerakoon, D. Skowronski, "EFFECT OF ON–LINE SEXUALITY EDUCATION ON HEALTH PROFESSIONAL STUDENTS’ COMFORT IN PROVISION OF SEXUAL HEALTH CARE", April 2007, https://
oai.core.ac.uk/None

The unit of study “Sexuality for Health Professionals” is offered on–line unit as an elective to all students enrolled in the Health Sciences Faculty of the Universit
y of Sydney.  The unit utilises the PLISSIT management model to present an interactive learning unit on the Web CT learning platform.  This paper presents the analysi
s of a pre and post learning evaluation of the unit
```

Figure 2



```
Write a research objective: health of hotdog
Generating Embedding for: health of hotdog
Querying database for similar matches
Annotated Bibliography:

A. Gates, "Synaphobranchid eel", Subsea 7, August 2012, https://oai.core.ac.uk/None

FeedingThe fish in the images looks like a species of the Synaphobranchid genus Ilyophis. This is based on the inflated looking snout, relatively small eye, and promi
nent white lateral line pores.

A possible species is Ilyophis blachei from the depth and body size.

It is feeding on a crustacean, perhaps a member of the Dendrobranchiat

M. Zeller, R. Jiménez-Melero, B. Santer, "Diapause in the calanoid freshwater copepod Eudiaptomus graciloides", 2004, 10.1093/plankt.

The seasonal appearance and the intensity of diapausing-egg production in Eudiaptomus graciloides in five lakes of different size and trophic level were studied. In a
ll lakes, diapausing eggs were produced in autumn. In the large mesotrophic lake Selenter See, the population’s shift to the production of diapausing eggs was more co
mplete than in other lakes. We examined day length, temperature and food as proximate factors for the production of diapausing eggs with laboratory experiments. Eudia
ptomus graciloides produced diapausing eggs in all treatments, but a significantly higher percentage of diapausing eggs was found under short day conditions except wh
en algal food was abundant and temperature was high. To investigate the adaptive significance of diapause in E. graciloides, we compared the survival of adult and juv
enile at different temperatures for E. graciloides with Eudiaptomus gracilis, a sympatric species that does not exhibit diapause. At 8°C, adult E. graciloides survive
d better than adult E. gracilis and exhibited reduced gut contents and accumulation of storage lipids, traits characteristic of adult diapause. Eudiaptomus graciloide
s nauplii did not reach the copepodid stage at 6°C, but E. gracilis nauplii exhibited high mortality and developed very slowly. We hypothesize that adult diapause and
 production of diapausing eggs facilitate the survival of E. graciloides during cold periods and enhance coexistence with its congener, E. gracilis, in temperate zone
s

D. Jones, "Ophiothrix fragilis", Total, January 2007, https://oai.core.ac.uk/None

Dense aggregation with the arms of some individual brittle stars projected into the water column.Dense aggregation of Ophiothrix fragilis on seafloo

S. Romeu Sala, "Implementation of a radiocommunication infrastructure in rural area with power supply by alternative energies", Universitat Politècnica de Catalunya,
February 2016, https://oai.core.ac.uk/None

Estudi pràctic de la planificació d'una infraestructura de telecomunicacions per donar sevei a un indret rural amb accés deficient als serveis de comunicacions electr
òniques. S'aporta la solució tecnològica que permet subministrar els serveis deficitaris i es dimensiona el sistema generador d?electricitat que garanteixi el funcion
ament autònom a través de les fonts d?energia alternativa fotovoltaica i eòlica.The main target of this paper is to plan a solution for providing access to the teleco
mmunication services in an isolated area by installing a radio infrastructure electrically supplied, in an autonomous way, with wind and solar alternative energy sour
ces. It is desired to enable a global access to the electronic communication technologies: DVB-T channels, FM sound broadcasting, 2G mobile telephony, high speed Inte
rnet by WiMAX and communications of the emergency and security TETRA network. In practice, it is planned a solution for the access problems to the electronic communic
ation services in the villages of Estaràs and Gàver. Initially, the actual situation of radiocommunication services is analyzed through radio field measures made usin
g specific equipment. ICS Telecom software is used to plan the location of the new radio infrastructure. Finally, the feasibility of implementing repeater stations (l
ower cost and consumption) rather than emitters is verified. Regarding the power supply, the wind turbine is dimensioned using wind data from the Atles Climatic as we
ll as numerical methods. Concerning the design of the photovoltaic generator, PVSyst software is used to get climate data of solar radiation and calculate the optimal
 tilt of the panels in order to, eventually, dimension the number of solar modules using numerical calculations. Additionally, a fuel generator is added to complement
 the energy requirements. The topology of this system is chosen within the technical solutions from the manufacturer SMA Solar Technology AG. Based on the electrical
requirements, inverters, batteries, regulators and wire distribution are modeled. Finally, the economic viability of the proposed system is quantified in relation to
a wired connection to the electric network. It is stated that, in case of power lines above 300 linear meters, the preferred solution is to power supply through alter
native energy.El objetivo principal de este trabajo es plantear una solución para dotar de acceso a servicios de telecomunicaciones en una zona aislada, mediante la i
nstalación de una infraestructura de radiocomunicaciones alimentada eléctricamente, de manera autónoma, con las fuentes de energía alternativas eólica y fotovoltaica.
 Se desea hacer posible un acceso global a las tecnologías de comunicaciones electrónicas: canales TDT, emisoras FM, telefonía móvil 2G, Internet rápido con WiMAX y c
omunicaciones de la red de emergencias y seguridad TETRA. A nivel práctico, se procede a planificar la solución para las deficiencias en el acceso a los servicios de
```

Figure 3



```
Write a research objective: business in australia
Generating Embedding for: business in australia
Querying database for similar matches
Annotated Bibliography:

N. Stewart, J. Ormiston, J. Gilligan,  et al., "Entrepreneurship and Innovation Program Annual Report 2013", The University of Sydney Business School, January 2014, h
ttps://oai.core.ac.uk/None

The Entrepreneurship and Innovation Program at the University of Sydney Business School focuses on identifying, nurturing and strengthening entrepreneurial communitie
s of learning and practice. This 2013 Annual Report sets out our teaching and research activities and achievements, as catalysts for community and action

M. Donnelley, J. Ormiston, R. Seymour, "Entrepreneurship and Innovation Program Annual Report - Ventures 2010", The University of Sydney Business School, January 2011
, https://oai.core.ac.uk/None

The Entrepreneurship and Innovation Program at the University of Sydney Business School focuses on identifying, nurturing and strengthening entrepreneurial communitie
s of learning and practice. This 2010 Annual Report sets out our teaching and achievements, as catalysts for community and action

M. Donnelley, J. Ormiston, R. Seymour, "Entrepreneurship and Innovation Program Annual Report - Ventures 2011", The University of Sydney Business School, January 2012
, https://oai.core.ac.uk/None

The Entrepreneurship and Innovation Program at the University of Sydney Business School focuses on identifying, nurturing and strengthening entrepreneurial communitie
s of learning and practice. This 2011 Annual Report sets out our teaching and achievements, as catalysts for community and action

M. Donnelley, J. Ormiston, R. Seymour, "Entrepreneurship and Innovation Program Annual Report - Ventures 2012", The University of Sydney Business School, January 2013
, https://oai.core.ac.uk/None

The Entrepreneurship and Innovation Program at the University of Sydney Business School focuses on identifying, nurturing and strengthening entrepreneurial communitie
s of learning and practice. This 2012 Annual Report sets out our teaching activities and achievements, as catalysts for community and action

R. Seymour, J. Ormiston, "Entrepreneurship and Innovation Program Annual Report - Research 2007", The University of Sydney Business School, January 2008, https://oai.
core.ac.uk/None

The Entrepreneurship and Innovation Program at the University of Sydney Business School focuses on identifying, nurturing and strengthening entrepreneurial communitie
s of learning and practice. This 2007 Annual Report sets out our research activities and achievements, as catalysts for community and action

R. Seymour, J. Ormiston, "Entrepreneurship and Innovation Program Annual Report - Research 2008", The University of Sydney Business School, January 2009, https://oai.
core.ac.uk/None

The Entrepreneurship and Innovation Program at the University of Sydney Business School focuses on identifying, nurturing and strengthening entrepreneurial communitie
s of learning and practice. This 2008 Annual Report sets out our research activities and achievements, as catalysts for community and action
```

Figure 4

5 DISCUSSION

There are three major problems that could explain how the performance of our final product was less than ideal. The first two issues focus on how the embeddings were generated and the third deals with processing power and time.

First, one explanation for less than expected performance could be with the process used to generate each embedding. As explained in **Section 2.2**, each word in a given document was processed separately and then averaged together to produce a single document embedding. This process ignores all semantic information between tokens so that much of the original meaning of each text is lost. Additionally, averaging each vector together reduces the overall usefulness of each individual word vector. One unique word such as “Capillary” would be averaged with many filler words in an abstract so that the meaning is less pronounced. Using sentence-based embedding models could provide better results in the future.

Another reason for loss of accuracy could stem from our choice to use the SciBert word embedding model. According to [2], the SciBert model was trained using a dataset that is composed of 18% papers in the computer science domain and 82% papers in the biomedical domain. Because of this training split, we would expect the SciBert model to excel at generating embeddings for biomedical specific articles, however, we believe the embeddings for papers outside of this field to suffer because of the lack of training. Using a more general case word2vec model might have provided better accuracy for our needs.

Lastly, due to the time constraint imposed, a minimal amount of data processing was used to populate the vector database. The embeddings generation step proved to be the largest bottleneck, severely limiting the efficiency of our system. Provided that we were able to spend more time and resources to completely process the Core dataset, we may have been able to see better results. Another point to make is that SciBert only provides a single pre-trained model with a size of 768 dimensions. Because the model was so large, the computation time needed to generate a single document embedding took over 30 seconds which would take approximately 82 years to process the entire Core dataset using the current computing power at our disposal.

6 FUTURE WORK

Based on the issues addressed in **Section 5**, there are a few follow up actions that could be taken based on our findings. One possible follow up paper could explore building a custom document embedding model using the CORE dataset to specialize in generating embeddings for research articles. Another paper could focus on building off of the work done in [2] and create SciBert models with smaller dimensions to increase computational efficiency.

ACKNOWLEDGEMENTS

This work was completed in part using the Discovery cluster, supported by Northeastern University's Research Computing team.

We would like to thank our professor Amir Tahmasebi for his guidance throughout the completion of this work.

REFERENCES

[1] Knoth, P., Herrmannova, D., Cancellieri, M. et al. 'CORE: A Global Aggregation Service for Open Access Papers', *Nature Scientific Data 10*, 366 (2023). https://doi.org/10.1038/s41597-023-02208-w

[2] I. Beltagy, K. Lo, and A. Cohan, 'SciBERT: A Pretrained Language Model for Scientific Text', in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (EMNLP-IJCNLP), 2019, pp. 3615–3620.

[3] Jina AI, "vectordb," GitHub, March 2024. [Online]. Available: https://github.com/jina-ai/vectordb/

[4] S. Bird, E. Klein, and E. Loper, "Natural language processing with Python: analyzing text with the natural language toolkit," O'Reilly Media, Inc., 2009.

APPENDIX

User Manual:

Demo:

1. Create your conda environment using “bash create_conda_env.script”
2. Navigate to the “src” directory
3. Run the Demo using “python Demo.py”

Creating your own vector database:

1. Register for access and download 2018 Metadata-only dataset from https://core.ac.uk/documentation/dataset#dataset2018
2. (Optional) Split your dataset into multiple folders with “python src/SplitDirectoryFiles.py
3. Reduce file sizes for processing with python src/ChunkLargeFile.py
    a. edit the FOLDERS variable in to point to whatever directories to chunk
4. Create a completed json tracker with touch completed.txt in the NLPProject directory
5. Begin processing your dataset with “python src/ProcessDataBuildDB.py <read_folder_path> <completed_file_path>”

Code Structure:

CitationGenerator.py:

This file contains the function “generate_citation” and its associated helpers. Our pipeline uses this function when generating embeddings to create IEEE citations from given metadata, which are then stored with the associated embedding vectors.

Dependencies:

- parser module from dateutil: parses unformatted dates to IEEE
- datetime: Used by parser module to define date format
- re: Regex module, used for parsing author names

Functions:

- _convert_date: takes in an unformatted date string, and parses it to IEEE format (month year)
- _convert_oai_to_doi: takes in an oai string, and prepends necessary information to retrieve document by link
- _parseFirstName: takes in an unformatted author first/middle name string and converts it to IEEE initials
- _parseAuthor: takes in a single author's name and converts it to IEEE format (first initial(s). Last name)
- _parseAuthors: takes in a list of authors, and produces a single representative string to be used in a citation
- generate_citation: Takes in a dictionary containing a list of authors, a document title, a publisher, a publishing date, an optional doi, and an oai, and uses it to produce an IEEE citation
- test: Tests all of the above functions with assert statements

EmbeddingGenerator.py:

This file contains the function “generate_embedding” which is called with a documents title, abstract, and list of topics, which it uses alongside SciBERT in order to produce a representative vector embedding.

Dependencies:

- Pytorch: This is used in order to calculate an average 1D tensor for a 2D tensor
- nltk: We use nltk’s “stopwords” corpus to get a list of stopwords to ignore while embedding

Functions:

- get_stopwords_set: This function is called once upon startup in order to download the nltk stopwords corpus
- generate_embedding: Takes in an embedding model and tokenizer, device string (‘cuda’ or ‘cpu’), and document title, abstract and topics, which it

combines into a single string to be parsed by generate_embedding_from_text

- generate_embedding_from_text: Tokenizes a given string using a given tokenizer, saving the first 300 tokens. For all tokens not in ntlk's stopwords list, it uses SciBERT to generate an embedding for it. All of these vectors are the averaged into a single embedding, which is returned
- test: Contains asserts to test all of the above functions

LocalVectorDB.py:

This file contains the VecDoc class, which uses the "vectordb" library to handle saving our embeddings to and retrieving embeddings from a vector database.

Dependencies:

- docarray: This is a dependency for vectordb, used to create the documents to store in the database
- vectordb: This is the local vector database library used to store and query the vector embeddings

Classes:

- VecDoc: Defines the document structure used in our implemented vector database
- LocalVectorClient: Simple class used to connect to the database to insert and query embeddings
    - insert: Given a list of vector doc embeddings, insert each one into the vector database
    - search: Given a query vector and a response limit n, return the n closest embedding documents to the query vector

PreprocessDataBase.py:

This file contains the "process_file" method, which takes in a filename pointing to a json containing a set of documents to embed. It then applies a given embedding function (SciBERT) to produce a representative embedding for each document, and produces IEEE citations, which are saved together in a local vector database

Dependencies:

- gc: python’s built-in garbage collector, used to save memory after each pass through process_file
- time: python’s timing package, used to track how long each document is taken, and how much more time is expected to take
- pandas: used for the “read_json” function, which decompresses and loads a json file into memory as a pandas dataframe
- VecDoc: from our “LocalVectorDB.py” file, used to insert embeddings into our local database
- generate_citation: from our “CitationGenerator.py” file, used to generate IEEE citations for each document to save in the vector database

Functions:

- process_file: Reads in a json at the given filepath. Filters out all documents that: do not have a title, have abstracts less than 500 characters, or are in a language other than English. For each remaining document, an embedding and citation are generated, and when all have been processed, these are saved to the vector database

ProcessDataBuildDB.py:

This is our main file used to produce embeddings from our dataset. It takes in a path pointing to the dataset folder and a path pointing to a “completed.txt” file that tracks which documents have been embedded. For each file in the dataset directory that is not found in completed.txt, it calls “process_file” to generate embeddings for each document.

Dependencies:

- os: Python operating system import, used to check filepaths and filesizes
- sys: Python system import for reading command line arguments
- collections: Python standard library, used for its Deque class
- Pytorch: Used for cuda parallelization and transformers sub-library

- Transformers: Used to download SciBERT embedding generator and tokenizer
- process_file from: import from our “PreprocessDataBase.py” class, used to process each compressed file in the database
- generate_embedding: import from our “EmbeddingGenerator.py” class, used to generate embeddings for each document
- LocalVectorClient: import from “LocalVectorDB.py” class, used to insert embeddings into a local vector database

Functions:

- shouldChunkFile: function that takes in a filepath and a max processing size, and checks if the file is too large to process
- setupTokenizerAndModel: function called on startup to download SciBERT’s tokenizer and embedding model from huggingface

Demo.py:

This file is called in order to run a demo displaying an interactive version of our project, with a local vector database. It prompts the user to input research objectives, and returns the top 6 annotated bibliographies with the most similar stored vectors.

Dependencies:

- setupTokenizerAndModel: Imported from our “ProcessDataBuildDB.py” file to download SciBERT’s tokenizer and model for user input processing
- generate_embedding_from_text: Imported from our “EmbeddingGenerator.py” file to generate an embedding for the user prompt
- LocalVectorClient: Imported from our “LocalVectorDB.py” file to retrieve information from our local vector database

Classes:

- Demo: class used to store our model, tokenizer, and local vector database
  - getAnnotations: Given a string containing a user request, generates an embedding using its stored model, and parses the stored local

vector database for the 6 most similar vectors, printing the associated annotated bibliographies

SplitDirectoryFiles.py:

This file is used in order to split the dataset equally into a specified number of smaller datasets. We used this file in order to divide the dataset so we could parallel process without potential conflicts

ChunkLargeFile.py:

This file is used to break each file in the dataset that is larger than a specified size into smaller subfiles. This is done to avoid potential memory overloads when processing very large files.

Dependencies:

- os: The python Operating system package used to examine files and filesizes
- concurrent.futures: library used for multiprocessing

Functions:

- utf8len: Gets the utf8 length of a string
- chunk_json_file_write_while_going: given a path to a json file, read in the contents and split to separate files with a specific size, compress new files to xz format
- chunk_xz_file: given a filepath to an xz file, uncompress the file, split the contents into separate files, and recompress the newly created files
- find_xz_files_above_size: given a directory path, return a list of paths to all files within the directory that have a size above a given threshold
- chunk_all_files_in_folder: Given a directory path, find all files over a certain threshold and chunk them into manageable files
- Main: Uses python preprocessing to chunk multiple directories on different threads

Testing.py:

This class can be called to test all of the above files which are testable.