SimSearch: A Similarity Search Tool for Research Paper Abstracts using NLP
20170717 Seungwoo Han, 20180162 Jisu Kim

1. Introduction/Motivation

The rapid advancement of research in various domains has led to an exponential increase in the number of scholarly papers being published every year [1]. Consequently, researchers face the challenging task of staying up-to-date with the latest findings and identifying the most relevant and significant research papers within their field of interest. The constant struggle with information overload and the difficulty in locating relevant papers quickly and efficiently highlights the need for a more effective method to help the research process.

This project aims to utilize the power of natural language processing (NLP) techniques to develop a similarity search tool, SimSearch, that assists researchers in finding similar research papers based on their abstracts. Such a tool would save time and effort for researchers and contribute to more efficient knowledge discovery in scientific literature.

2. Related Work

Several related similarity search systems have been proposed in the past to improve searching capabilities for traditional keyword search engines or use other NLP techniques such as topic modeling and clustering [2]. In recent years, embedding-based approaches, particularly the Word2Vec model, have been used for text representation and semantic search [3]. However, the problem remains in capturing the semantic meaning and context information at the document level for efficient similarity search [4].

3. System Description

3.1. Data

We used the Kaggle arXiv dataset, which consists of over 1.7 million scholarly papers across the STEM fields. The data itself includes metadata such as titles, categories, abstracts, and update dates, which can be used for train and test data in SimSearch [5].

3.2. Preprocessing

3.2.1. Generating dataframe and filtering selected features

An essential step in the development of SimSearch is data preprocessing, which involves the selection of relevant features and the removal of unnecessary words from the abstracts. A data frame is generated to filter out the required features, such as title, categories, abstract, and update_date.

	title	categories	abstract	update_date
0	Probing FSR star cluster candidates in bulge/d...	astro-ph	We analyse 20 star cluster candidates projec...	2012-04-02
1	Further Development of the Tetron Model	hep-ph	After a prologue which clarifies some issues...	2014-11-18
2	Characterisation of a three-dimensional Browni...	physics.atom-ph	We present here a detailed study of the beha...	2007-09-18
3	Nonmagnetic Impurity Resonances as a Signature...	cond-mat.supr-con	The low energy band structure of the FeAs ba...	2013-05-29
4	Analysis of optical properties of strained sem...	cond-mat.mes-hall	Using multiband k*p theory we study the size...	2010-02-11

Figure 1. Filtered selected features dataframe

3.2.2. Removing unnecessary words

As a part of preprocessing abstract, we removed certain types of words that carry less semantic information, using Part-of-Speech (POS) tagging. We removed the following POS tags:

- CC: Coordinating conjunction (e.g., and, but, or)
- DT: Determiner (e.g., the, a, an)
- IN: Preposition or subordinating conjunction (e.g., in, on, with)
- TO: The preposition ‘to’
- PRP: Personal pronoun (e.g., I, you, he, she)

- PRP$: Possessive pronoun (e.g., my, you, his, her)
- MD: Modal verb (e.g., can, may, should)
- WP: Wh-pronoun (e,g., who, what, which)
- WP$: Possessive wh-pronoun (e.g., whose)
- WRB: Wh-adverb (e.g., when, where, why) [6]

3.3. Implementation

3.3.1. TF-IDF with N-grams

We implemented our first model using TF-IDF with n-grams and cosine similarity. TF-IDF (Term Frequency - Inverse Document Frequency) works by determining the relative frequency of words in a specific document compared to the inverse proportion of that over the entire document corpus. Intuitively, this calculation determines how relevant a given word is in a document [7]. We included n-grams in TF-IDF calculation to provide a way to capture phrases and local context, where the meaning is determined by a combination of words rather than a single word. We included both 1-grams (individual words) and 2-grams (two consecutive words) in the TF-IDF calculation. Cosine similarity measures the similarity between two vectors of an inner product space [8]. When applied to TF-IDF vectors, which are non-negative, the cosine similarity ranges from 0 to 1. A value of 0 indicates no similarity, with a value of 1 indicating that the vectors are identical. In our implementation, the cosine similarity is calculated among the TF-IDF vector representations of abstracts, giving a measure of how similar the content of different abstracts is.

3.3.2. Doc2Vec

To create document-level embeddings capable of capturing semantic meaning and context information, the Doc2Vec model is employed. This model, an extension of the Word2Vec model, provides two algorithms for creating document embeddings: Paragraph Vector-Distributed Memory (PV-DM) and Paragraph Vector-Distributed Bag of Words (PV-DBOW). PV-DM preserves word order within documents, while PV-DBOW ignores it in favor of a more simplistic approach [9]. In the implementation phase of SimSearch, the Doc2Vec model is trained using the preprocessed data. Given a research paper's abstract as input, the model infers a vector representing that document's semantic meaning. Then, the most similar abstracts are retrieved based on the inferred vector, providing users with a list of relevant research papers.

4. User Interaction

4.1. Deployment

It is deployed using Gradio and Hugging Face. Users can try out the demo at the following URL: https://ddiddu-simsearch.hf.space/

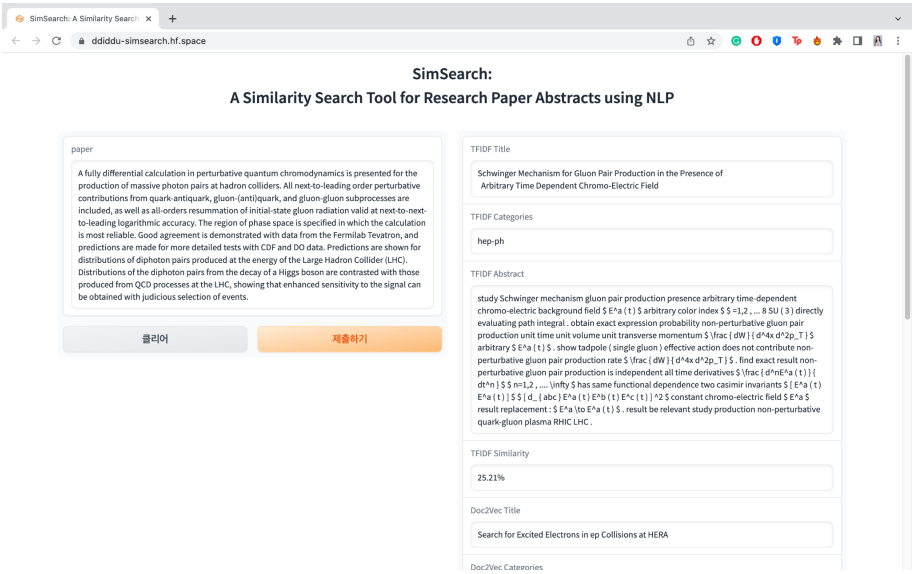

Figure 2. Interface of the demo

4.2. Working Scenario

The users input the abstract of a paper they are interested in. It then processes the abstract and searches for similar papers via TF-IDF and Doc2Vec models. The output received by the user includes: paper title, category, abstract, similarity. The user can compare the suggestions from both the TF-IDF and Doc2Vec models to find the most relevant paper.

5. Discussion

5.1. Results

In our test case, we observed the following results.

Result 1. For the input paper titled ‘Calculation of prompt diphoton production cross sections at Tevatron and LHC energies’ [10], the TF-IDF model recommended ‘Schwinger Mechanism for Gluon Pair Production in the Presence of Arbitrary Time Dependent Chromo-Electric Field’ [11]. They are relevant because both papers focus on pair production. The Doc2Vec model recommended ‘Search for Excited Electrons in ep Collisions at HERA’ [12]. While it still is in the particle physics field, it focuses on a different topic, excited electrons.

Result 2. For the input paper titled ‘Time Series, Stochastic Process and Completeness of Quantum Theory’ [13], both the TF-IDF model and Doc2Vec model recommended ‘Forget time’ [14]. Both papers seem to question the fundamental aspects of quantum theory, but their focus differs: the input paper focuses on statistical interpretation of quantum mechanics and experimental data analysis, while the recommended paper focuses on the theoretical and conceptual aspects of quantum gravity. Both papers share the field of quantum theory, but the specific topics they focus on differ.

5.2. Advantages and Limitations

Our system has the following advantages:

- Improves efficiency and effectiveness of finding relevant research papers
- Improves the accuracy of identifying similar papers based on abstracts
- Reduces the time and effort required to conduct literature reviews
- Facilitates the potential for enhanced collaboration and knowledge sharing

Despite its advantages, our system has limitations:

- Delimited to papers in the dataset: the accuracy of our recommendation system may be lower for papers belonging to categories not covered by the dataset. If a paper's subject is not related to our dataset, our system may not be able to provide highly relevant recommendations.

5.3. Further Improvement

For further improvement, we can consider adding more datasets to the recommendation system. It could lead to improved accuracy and coverage across a wider range of categories. And, we can use web crawlers to automatically collect papers from sources like Google Scholar.

6. Conclusion

Our work with this tool and the methods it employs, such as text preprocessing, POS tagging, N-grams, and key information extraction, are directly related to the concepts discussed in Natural Language Processing with Python. It provides a practical application of these concepts and demonstrates the potential impact of NLP in improving the efficiency and effectiveness of the field of scholarly search.

7. References

[1] Fire, Michael & Guestrin, Carlos. (2019). Over-optimization of academic publishing metrics: Observing Goodhart's Law in action. GigaScience. 8. 10.1093/gigascience/giz053.

[2] Xie, Pengtao & Xing, Eric. (2013). Integrating Document Clustering and Topic Modeling.

[3] Jatnika, Derry & Bijaksana, Moch & Ardiyanti, Arie. (2019). Word2Vec Model Analysis for Semantic Similarities in English Words. Procedia Computer Science. 157. 160-167. 10.1016/j.procs.2019.08.153.

[4] Ostendorff, Malte. (2020). Contextual Document Similarity for Content-based Literature Recommender Systems.

[5]University, C. (2023, June 10). ArXiv dataset. Kaggle. https://www.kaggle.com/datasets/Cornell-University/arxiv

[6] Universal pos tags. (n.d.-b). https://universaldependencies.org/u/pos/

[7] Ramos, J. (2003, December). Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning (Vol. 242, No. 1, pp. 29-48).

[8] *Cosine similarity*. Cosine Similarity - an overview | ScienceDirect Topics. (n.d.). https://www.sciencedirect.com/topics/computer-science/cosine-similarity

[9] Le, Quoc & Mikolov, Tomas. (2014). Distributed Representations of Sentences and Documents. 31st International Conference on Machine Learning, ICML 2014. 4.

[10] Balazs, C., Berger, E. L., Nadolsky, P., & Yuan, C. P. (2007). Calculation of prompt diphoton production cross sections at Fermilab Tevatron and CERN LHC energies. *Physical Review D*, *76*(1), 013009.

[11] Nayak, G. C. (2009). Schwinger mechanism for gluon-pair production in the presence of arbitrary time-dependent chromo-electric fields. *The European Physical Journal C*, *59*, 715-722.

[12] Aaron, F. D., Alexa, C., Andreev, V., Antunovic, B., Aplin, S., Asmone, A., ... & Meyer, H. (2008). Search for excited electrons in ep collisions at HERA. *Physics Letters B*, *666*(2), 131-139.

[13] Kupczynski, M. (2011, March). Time series, stochastic processes and completeness of quantum theory. In *AIP Conference Proceedings* (Vol. 1327, No. 1, pp. 394-400). American Institute of Physics.

[14] Rovelli, C. (2009). Forget time. *arXiv preprint arXiv:0903.3832*.