2021

Topic Identification & Text Summarization for Computer Science Journals

Group 4

Nestor Molina

Ngoc Phan

William Baker

GitHub Repos

https://github.com/nphan20181/nlp_project

Motivation & Significance

When working on a research paper or article, most researchers often have difficulty in writing an effective abstract. According to the Writing Center at the University of Wisconsin-Madison, an abstract is a short summary of a research paper that includes one paragraph of 6-7 sentences or 150-250 words [1]. The Writing Center at the University of Wisconsin-Madison also indicates that “search engines and bibliographic databases use abstracts to identify key terms for indexing research articles” [1]. Therefore, an effective abstract would enable identification of relevant key terms and potentially increase the article’s access rate as it increases the chance of the article showing up on the search page when the users search for related topics. This research applies text summarization techniques to generate a short summary of Computer Science research articles, which could further serve as an abstract of the article. Relevant topics would also be extracted from the article’s texts to enable researchers to quickly identify relevant keywords for their paper. This could further enable researchers to come up with an effective title for their paper and attract more readers to access their article.

Objectives

The objectives of this project can be split into several distinct parts.

- Data collection
- Data pre-processing
- Topic extraction
- Article summarization
- Model evaluation

Using pre-identified keywords and abstracts as targets, we will develop two different models. The first model will be used to identify the topic of the relevant text and the second model will be used to summarize that same text. After developing both working models, the models would be evaluated by comparing the summarized text with the abstract and the generated topics with the article’s keywords. Additionally, a pre-trained text summarization model would serve as a baseline model for comparison across models.

Dataset

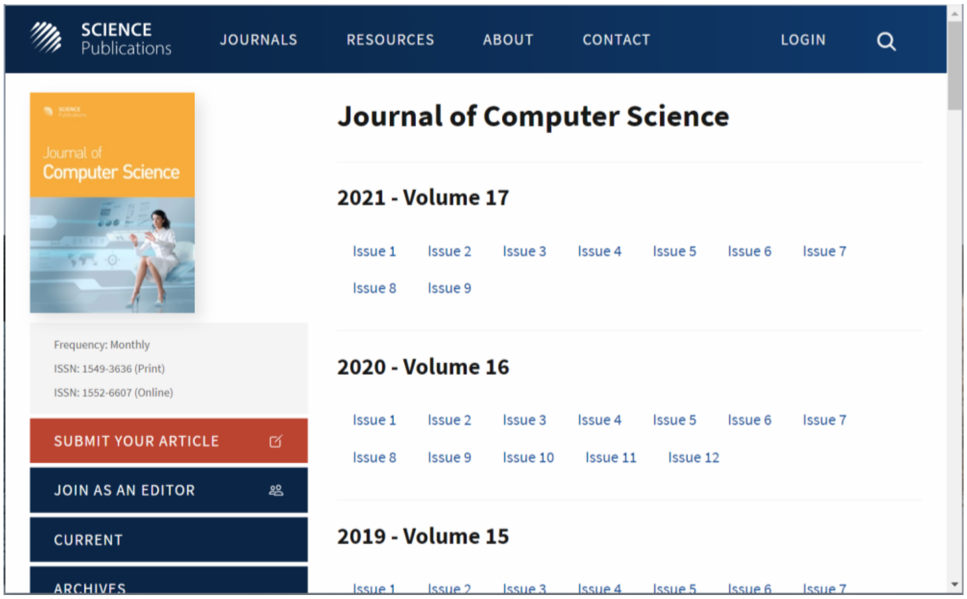
Data Collection

A collection of published articles from 2005 to 2021 in the Journal of Computer Science has been retrieved from the Science Publications website [2]. Web scraping has been performed on the following web pages to retrieve article's information and download the article's PDF file.

- Archive Page of Journal of Computer Science [2]

  This page contains the URLs for all journal issues from 2005 to current. On this page, web scraping has been performed to retrieve the URLs of journal issues along with year of publication, volume number, and issue number.

Figure 1: Screenshot of archive page of Journal of Computer Science



Frequency: Monthly
ISSN: 1549-3636 (Print)
ISSN: 1552-6607 (Online)

Journal of Computer Science

2021 - Volume 17

Issue 1 Issue 2 Issue 3 Issue 4 Issue 5 Issue 6 Issue 7 Issue 8 Issue 9

2020 - Volume 16

Issue 1 Issue 2 Issue 3 Issue 4 Issue 5 Issue 6 Issue 7 Issue 8 Issue 9 Issue 10 Issue 11 Issue 12

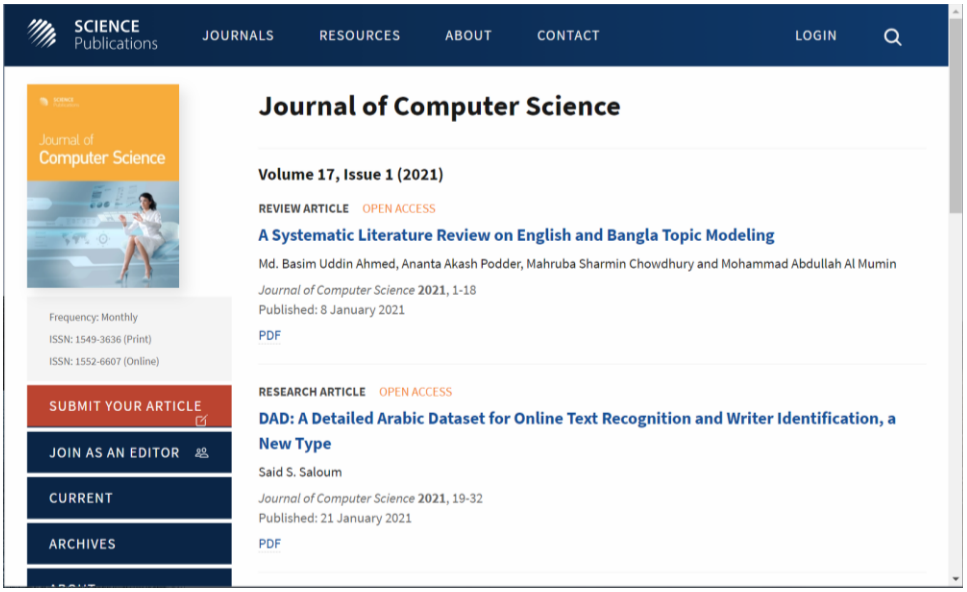
2019 - Volume 15

Issue 1 Issue 2 Issue 3 Issue 4 Issue 5 Issue 6 Issue 7

- Issue Page of Journal of Computer Science

  This page shows all articles belonging to the selected journal issue. The web page can be accessed by clicking on the hyperlink for a specific journal issue on the Archive Page of Journal of Computer Science. Web scraping has been performed on this page to retrieve the article's title, URL of the article's PDF file, and URL of the article's abstract page.

Figure 2: Screenshot of Issue Page of Journal of Computer Science



Journal of Computer Science

Frequency: Monthly
ISSN: 1549-3636 (Print)
ISSN: 1552-6607 (Online)

Volume 17, Issue 1 (2021)

REVIEW ARTICLE OPEN ACCESS

A Systematic Literature Review on English and Bangla Topic Modeling

Md. Basim Uddin Ahmed, Ananta Akash Podder, Mahruba Sharmin Chowdhury and Mohammad Abdullah Al Mumin

Journal of Computer Science **2021**, 1-18
Published: 8 January 2021
PDF

RESEARCH ARTICLE OPEN ACCESS

DAD: A Detailed Arabic Dataset for Online Text Recognition and Writer Identification, a New Type

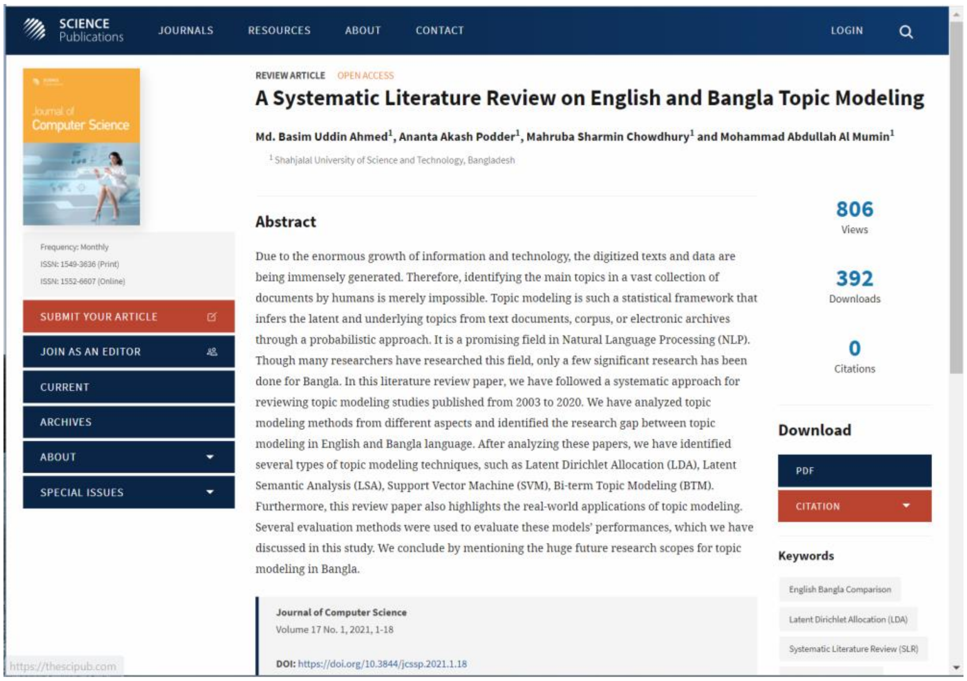
Said S. Saloum

Journal of Computer Science **2021**, 19-32
Published: 21 January 2021
PDF

- Article's Abstract Page

  This page contains general information for one specific article and can be accessed from Issue Page of Journal of Computer Science. Web scraping has been performed on this page to retrieve the abstract and the keywords of the selected article.

Figure 3: Screenshot of article's Abstract Page



Frequency: Monthly
ISSN: 1549-3636 (Print)
ISSN: 1552-6607 (Online)

REVIEW ARTICLE OPEN ACCESS

A Systematic Literature Review on English and Bangla Topic Modeling

Md. Basim Uddin Ahmed[1], Ananta Akash Podder[1], Mahruba Sharmin Chowdhury[1] and Mohammad Abdullah Al Mumin[1]

[1] Shahjalal University of Science and Technology, Bangladesh

Abstract

Due to the enormous growth of information and technology, the digitized texts and data are being immensely generated. Therefore, identifying the main topics in a vast collection of documents by humans is merely impossible. Topic modeling is such a statistical framework that infers the latent and underlying topics from text documents, corpus, or electronic archives through a probabilistic approach. It is a promising field in Natural Language Processing (NLP). Though many researchers have researched this field, only a few significant research has been done for Bangla. In this literature review paper, we have followed a systematic approach for reviewing topic modeling studies published from 2003 to 2020. We have analyzed topic modeling methods from different aspects and identified the research gap between topic modeling in English and Bangla language. After analyzing these papers, we have identified several types of topic modeling techniques, such as Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), Support Vector Machine (SVM), Bi-term Topic Modeling (BTM). Furthermore, this review paper also highlights the real-world applications of topic modeling. Several evaluation methods were used to evaluate these models' performances, which we have discussed in this study. We conclude by mentioning the huge future research scopes for topic modeling in Bangla.

806
Views

392
Downloads

0
Citations

Keywords

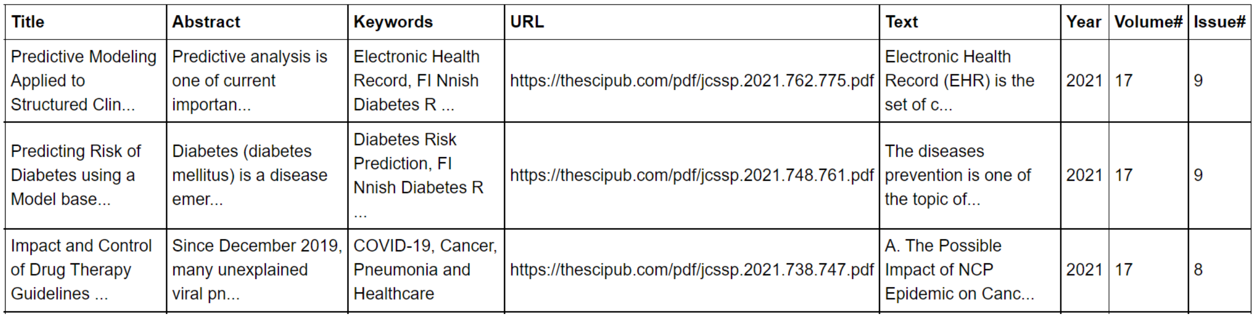
English Bangla Comparison
Latent Dirichlet Allocation (LDA)
Systematic Literature Review (SLR)

Journal of Computer Science
Volume 17 No. 1, 2021, 1-18

DOI: https://doi.org/10.3844/jcssp.2021.1.18

After retrieving the URL of all published articles and relevant information, a Python code has been written to download the articles and extract the article's texts from PDF files. Once the texts have been extracted, the data is saved to a csv file. The result dataset contains 2,631 articles and 8 columns. Table 1 displays the first three rows of the data.

Table 1: First three rows of dataset

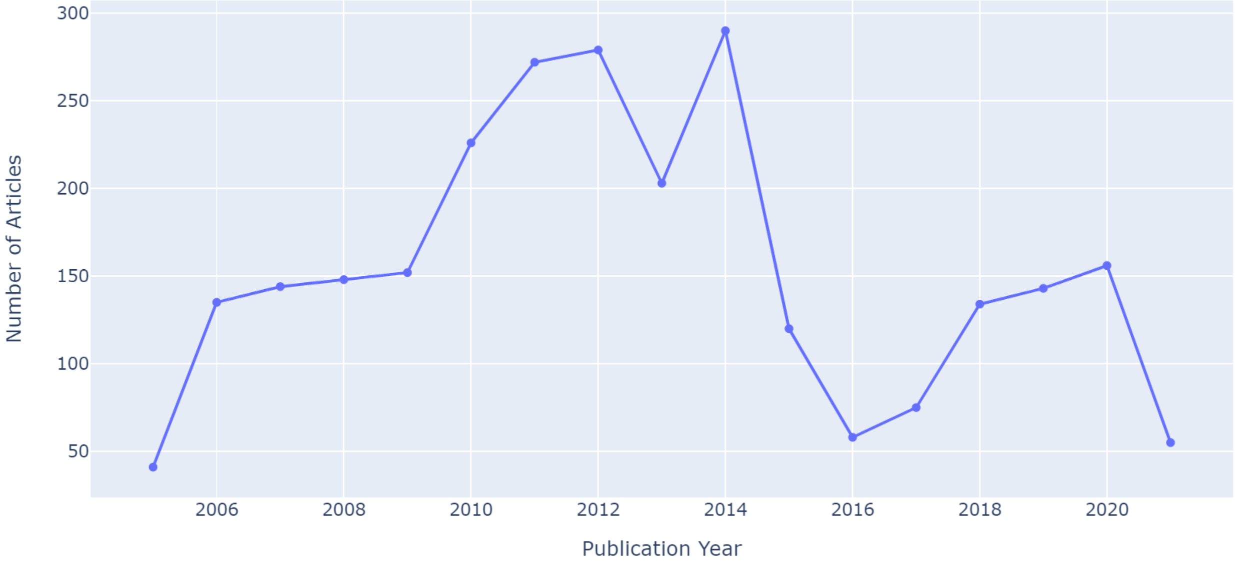


Title	Abstract	Keywords	URL	Text	Year	Volume#	Issue#
Predictive Modeling Applied to Structured Clin...	Predictive analysis is one of current importan...	Electronic Health Record, FI Nnish Diabetes R ...	https://thescipub.com/pdf/jcssp.2021.762.775.pdf	Electronic Health Record (EHR) is the set of c...	2021	17	9
Predicting Risk of Diabetes using a Model base...	Diabetes (diabetes mellitus) is a disease emer...	Diabetes Risk Prediction, FI Nnish Diabetes R ...	https://thescipub.com/pdf/jcssp.2021.748.761.pdf	The diseases prevention is one of the topic of...	2021	17	9
Impact and Control of Drug Therapy Guidelines ...	Since December 2019, many unexplained viral pn...	COVID-19, Cancer, Pneumonia and Healthcare	https://thescipub.com/pdf/jcssp.2021.738.747.pdf	A. The Possible Impact of NCP Epidemic on Canc...	2021	17	8

Exploratory Data Analysis

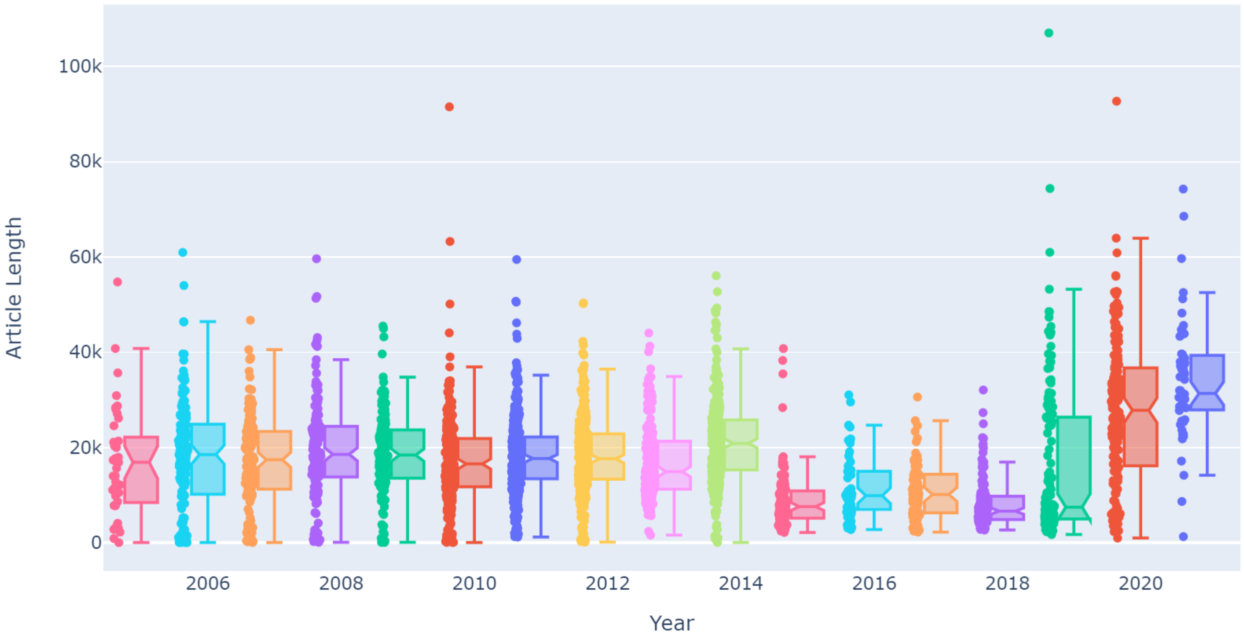
Figure 4 shows the number of published articles per year. Year 2014 has the greatest number of published articles while year 2005 has the least number of published articles. There seems to be an increasing trend in the number of published articles from 2005 to 2014. After 2014, the trend seems to be decreasing.

Figure 4: Number of published articles per year



As shown in the Figure 5, the article's length did not change much before 2014. However, most articles published between 2105 and 2018 seem to have shorter length compared to that of articles published before 2015 and after 2018. Furthermore, most articles published after 2019 seem to have longer text compared to articles published in other time periods.

Figure 5: Distribution of article's length per year



References

1. *Writing an Abstract for Your Research Paper*. (n.d.). The Writing Center | University of Wisconsin – Madison. https://writing.wisc.edu/handbook/assignments/writing-an-abstract-for-your-research-paper/

2. *Journal of Computer Science*. (n.d.). Science Publications. https://thescipub.com/jcs/archive