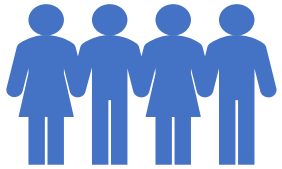
Topic Identification & Text Summarization for Computer Science Journals



Group 4

Ngoc Phan

Nestor Molina

William Baker

GitHub Repository

https://github.com/nphan20181/nlp_project

Agenda

- Introduction
- Work Process Flow
- Dataset
- Data Preprocessing
- Topic Identification Model
- Text Summarization Model
- Web Application Demo

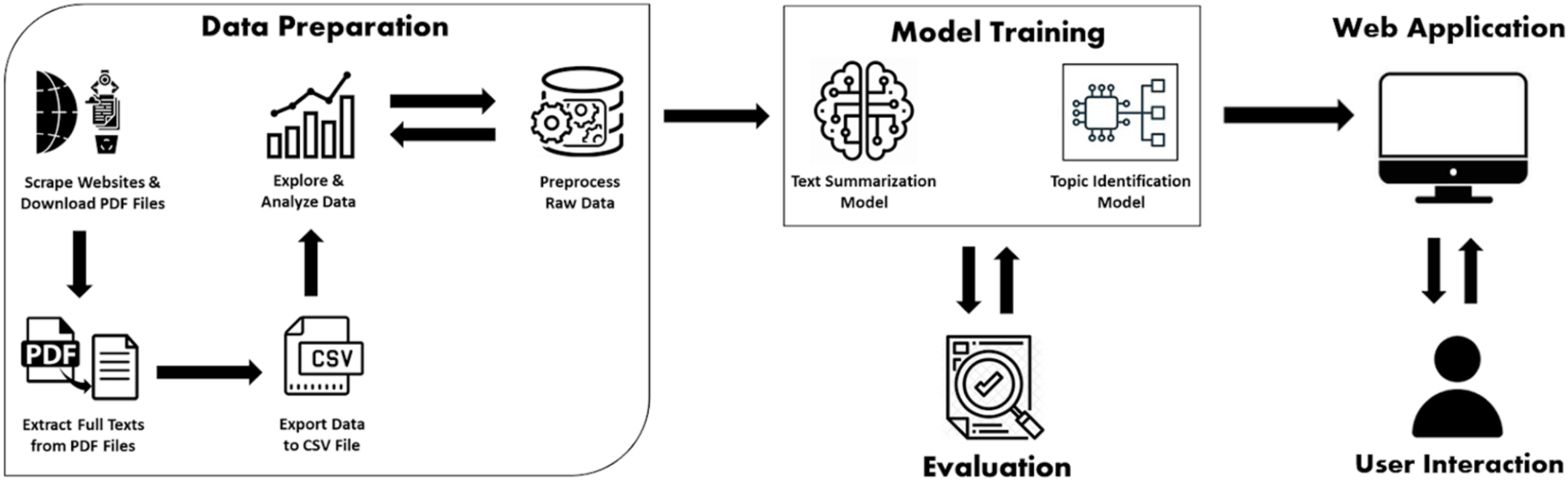
Introduction: Problem Statement

- Most researchers often have difficulty
  - Writing an effective abstract
  - Identifying appropriate keywords that reflect the main ideas
- The Writing Center at the University of Wisconsin-Madison
  - an abstract
    - short summary of a research paper
    - 6-7 sentences or 150-250 words
  - *“search engines and bibliographic databases use abstracts to identify key terms for indexing research articles”*

Introduction: Proposed Solution

- A web application that performs text summarization and topics extraction
  - User provides the input text
  - Application provides
    - summarized text
    - relevant topics along with recognized keywords or terms

Work Process Flow

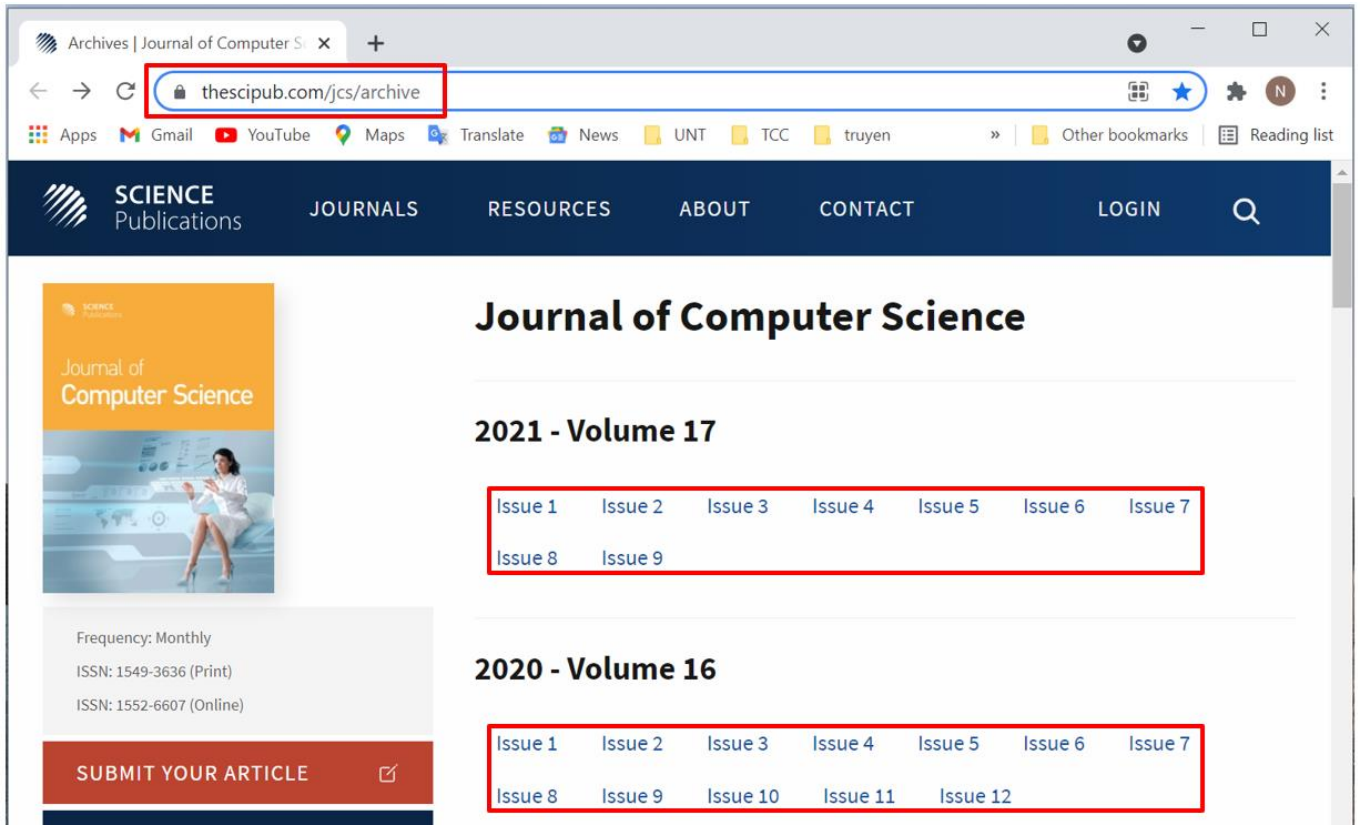


Dataset

Data Collection

Step 1

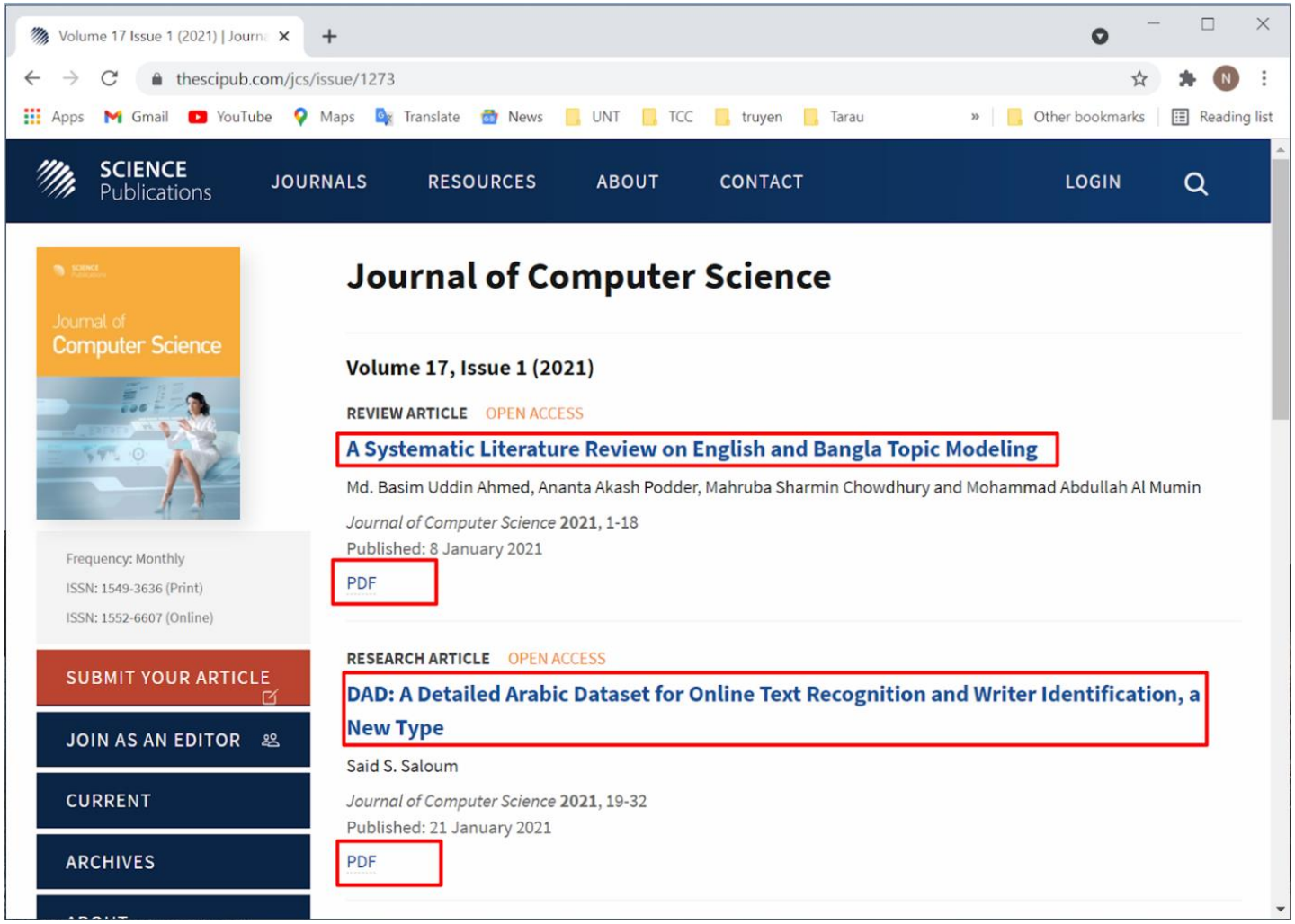
- Access the archive page of Journal of Computer Science at https://thescipub.com/jcs/archive
- Retrieve URLs of all journal issues



Data Collection (cont.)

Step 2

- Access each journal issue page
  - Retrieve URLs of all articles
  - Download pdf files of all published articles



Data Collection (cont.)

Step 2 (cont.)

- Extract article's full text from pdf file
  - Ignore all texts before ***Introduction*** header
  - Ignore all texts after ***References*** header
  - Keep only texts in between ***Introduction*** and ***Reference*** headers

Journal of Computer Sciences

Original Research Paper

Design and Implementation of Security in Healthcare Cloud Computing

Molamoganyi Gorata, Adamu Murtala Zungeru, Mmoloki Mangwala and Joseph Chuma

Department of Electrical, Computer and Telecommunication Engineering, College of Engineering and Technology, Botswana International University of Science and Technology, Private Bag 16, Palapye, Botswana

Article history
Received: 27-12-2016
Revised: 15-03-2017
Accepted: 10-04-2017

Corresponding Author:
Adamu Murtala Zungeru
Department of Electrical, Computer and Telecommunication Engineering, College of Engineering and Technology, Botswana International University of Science and Technology, Private Bag 16, Palapye, Botswana
Email: zungerum@biust.ac.bw

Abstract: As technology keeps on evolving, different organisations make use of the recent trends in technology and the health sector is no exception. As the cost of healthcare services is increasing, healthcare professionals are becoming scarce. Healthcare organisations have also adopted the latest technology of cloud computing. The introduction of cloud computing has proved to be a feasible idea on the information technology community. Rather than keeping the patient's information in a file in a health facility he/she was treated in, the information is stored in a cloud so that it can be shared amongst all health organisations and health professionals. Information is stored in a central place where it can be easily accessed, thus saving time and avoiding repetition of always writing the information each time a patient is attended to in a different facility. However, there are issues with sharing such information on the cloud since it is sensitive information. Ensuring this sensitive information security, availability and scalability are a major factor in the cloud computing environment. In this study, we proposed a mathematical model for measuring the availability of data and machines (nodes). We also present the current state-of-the-art research in this field by focusing on several shortcomings of current healthcare solutions and standards and we further proposed a system that will encrypt data before it is being sent to the cloud. The system is intended to be linked to the cloud in such a way that, before the client submits the data to the cloud and, the data will go through that system for encryption. The paper presents the steps to achieve the proposed system and also a sample encrypted and decrypted file.

Keywords: Cloud Computing, Healthcare Services, Information Security, Availability

Introduction

In the olden days, information about patients was stored in files in different health facilities where there have been treated. The problem with this method is when the file gets lost, or there is fire, the patient's medical history cannot be retrieved from anywhere and each time the patient is attended at a different facility, a file has to be created which is time-consuming and wastage of resources. With the introduction of cloud computing, there are new possibilities in health sector such as easy and flexible access to medical data, opportunities for new business models (Lohr *et al*., 2010).

Cloud computing is the type of computing used for sharing resources over the internet using virtual machines rather than physical machines resources like servers, storage applications and services can be rapidly provided and released with minimal management effort or service provider interaction (Gavrilov and Trajkovik, 2012).

The application of cloud computing in the health sector is called health cloud. However, as much as the health cloud brings many benefits there are also some challenges. As we know that the health care deals with very sensitive information, there is aneed for an increased security and privacy levels so that this information does not fall into the wrong hands. The availability of the information to the users is also very important. Security needs to be implemented on both the cloud and the client side. On the work of previous

Science Publications

DOI: 10.3844/jcssp.2017.34.47

Fig. 14. Shows the file after decryption

Results after Decryption

The image in Fig. 14 shows the decrypted file. This is basically what the user will be able to see after granted access to the secured information via random key generation. The result is produced in the CloudSim software.

Conclusion

In recent years, cloud computing has become the buzzword in the information technology community. Cloud computing has not only benefited the information technology community, even the health care organisations are benefiting from this technology. The sharing of health information on the cloud had made it easy for health professionals to access patient's information from anywhere around the world. Treatment of patients has now been made easy because the patient history is now available on the cloud. Although the whole world is greatly benefiting from this technology, there are also some issues with the security of data in the cloud. Strong security measures should be implemented so that data does not fall on the wrong hands. A mathematical model is proposed in this study for computing availability of information and servers on cloud computing. This model considers both data and node availability parameters which some of them are static during the cloud lifetime and others have dynamic nature. That means they can change based on the resource capacity or can be varying based on the different algorithm that the systems use. The paper also present the current state-of-the-art research in this field by focusing on several shortcomings of current healthcare solutions and standards and we further proposed a system that will encrypt data before it is being sent to the cloud. The system is intended to be linked to the cloud in such a way that, before the client submits the data to the cloud the data will go through that system for encryption. The paper presents the steps to achieve the proposed system and a sample of the encrypted and decrypted file using our proposed method is given.As a future work the input function for the graph and weigh of graph need more attention for achieving the complete availability model, also, we have intended to work on security model which will be responsible for securing data on the client side. Outside these, there is need to compare our proposed security solution with other existing ones, which is another step to see the performance of our design.

Author's Contributions

All authors equally contributed in this work.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

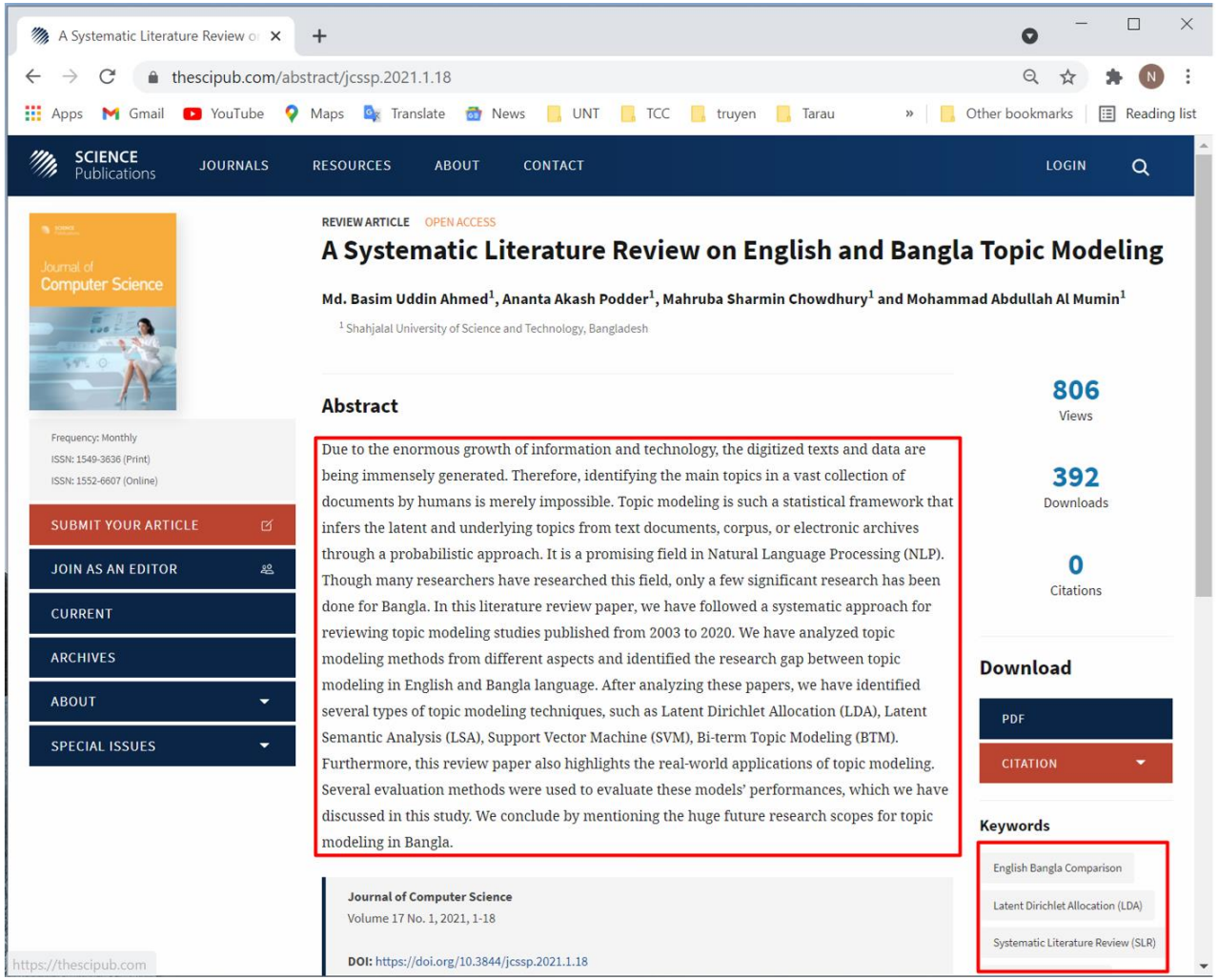
References

Al-Khanjari, Z., A. Al-Ani and S. Al-Hermizy, 2014. A proposed security architecture for establishing privacy domains in e-health cloud. Eur. Scientific J., 2: 322-330.

Data Collection (cont.)

Step 3

- Access each article page
  - Retrieve article's abstract and keywords



Data Collection (cont.)

Step 4

- Store collected data in a csv file
- Result dataset
  - 7 columns: date, title, abstract, keywords, file name, URL of pdf file, article's text.
  - 2,691 examples

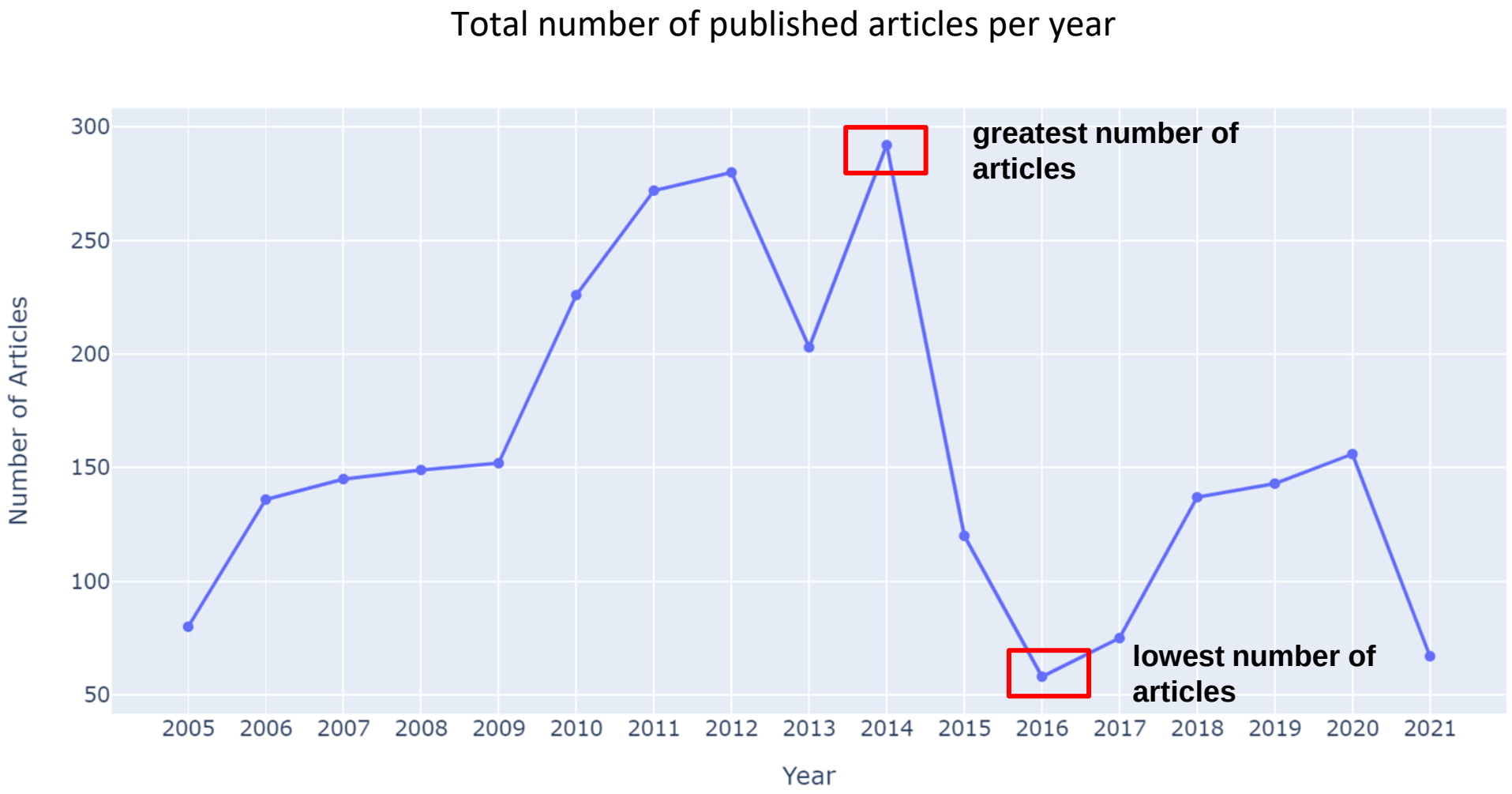
Date	Title	Abstract	Keywords	File Name	URL	Text
Published: 8 January 2021	A Systematic Literature Review on English and ...	Due to the enormous growth of information and ...	English Bangla Comparison, Latent Dirichlet Al...	2021_17_1_jcssp.2021.1.18.pdf	https://thescipub.com/pdf/jcssp.2021.1.18.pdf	Because of the rapid development of Informatio...
Published: 21 January 2021	DAD: A Detailed Arabic Dataset for Online Text...	This paper presents a novel Arabic dataset tha...	Arabic Dataset, Arabic Benchmark, Arabic Recog...	2021_17_1_jcssp.2021.19.32.pdf	https://thescipub.com/pdf/jcssp.2021.19.32.pdf	In the literature, many papers that focus on A...
Published: 20 January 2021	Collision Avoidance Modelling in Airline Traff...	An Air Traffic Controller (ATC) system aims to...	Air Traffic Control, Collision Avoidance, Conf...	2021_17_1_jcssp.2021.33.43.pdf	https://thescipub.com/pdf/jcssp.2021.33.43.pdf	Collision avoidance on air traffic becomes ver...

Data Analysis

- Summary statistics of un-processed data

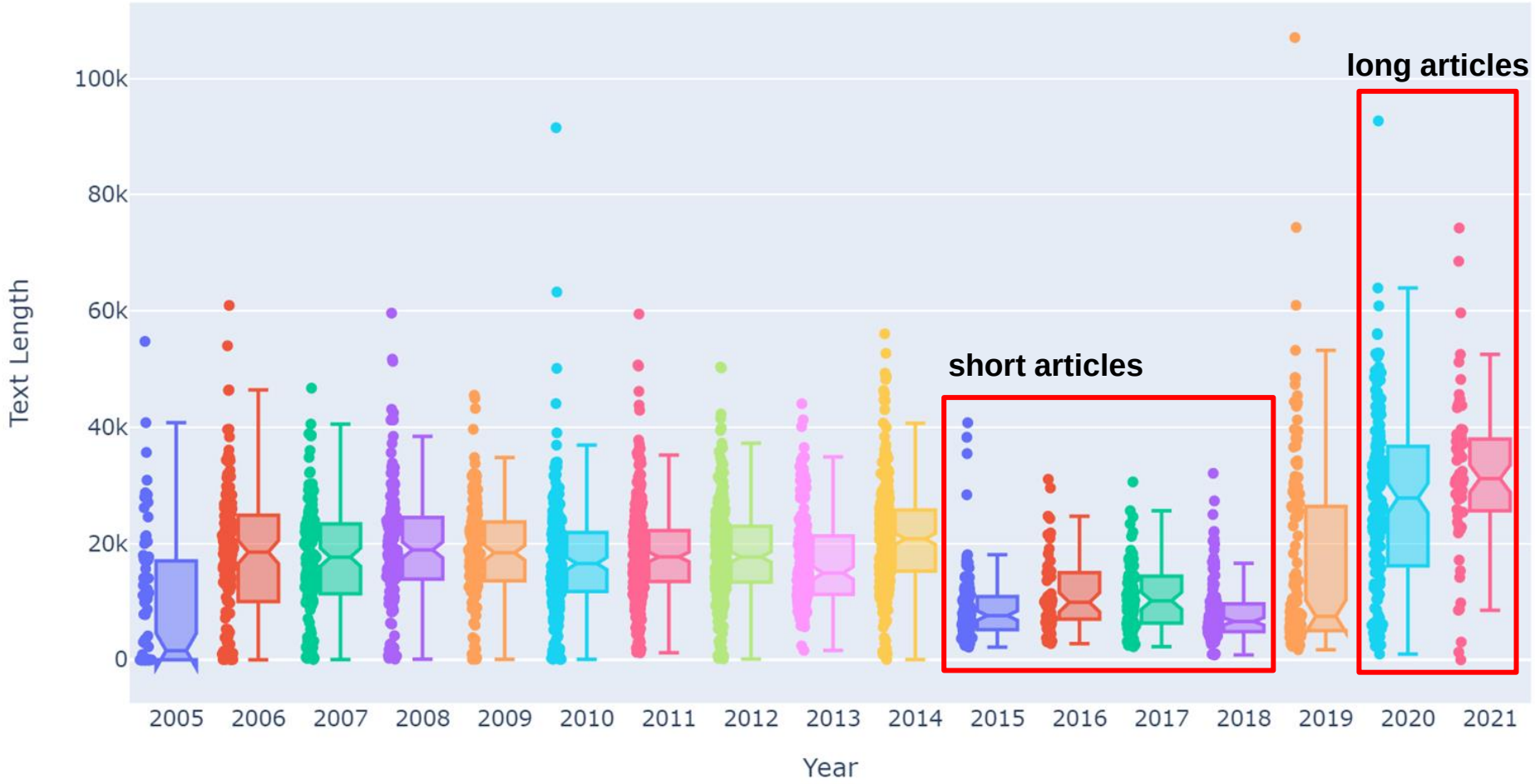
	mean	median	std	min	max
Title Length	83	81	26	14	260
Abstract Length	1437	1270	717	59	7295
Number of Keywords	5	5	2	1	16
Text Length	17686	16546	11257	3	107028

Data Analysis (cont.)



Data Analysis (cont.)

Distribution of article's text length



Data Preprocessing

Missing Values & Invalid Data Handling

- Missing values
  - 40 examples with missing article’s full text
  - 5 examples with missing keywords
- Invalid data
  - Found garbage in several article’s full texts
- Result dataset contains 1,811 articles

Text Processing

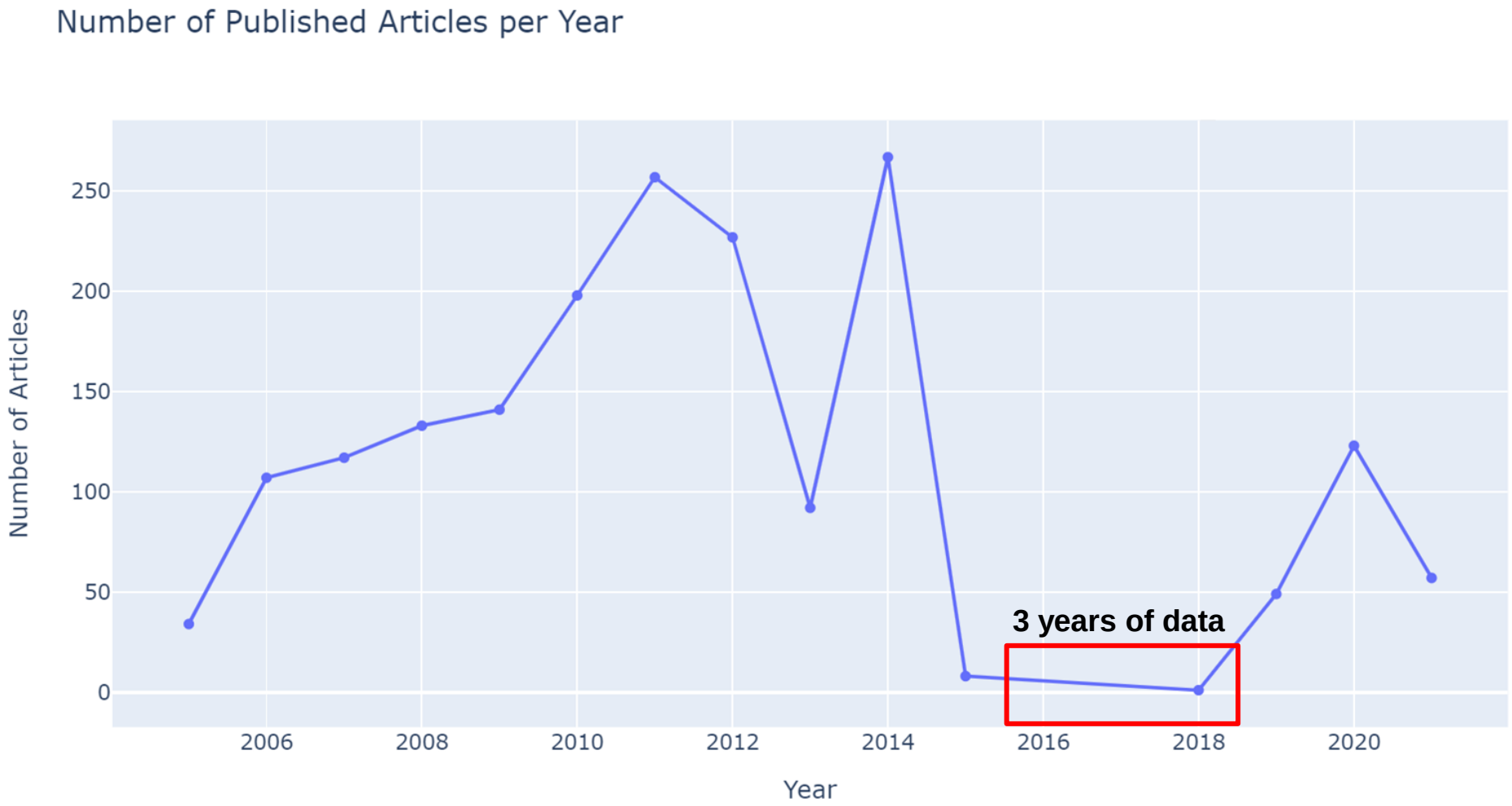
- Remove hyperlinks, footnote texts, and figure captions
- Convert texts to lowercase
- Remove stop words
- Apply lemmatization on article’s full text and abstract
- Apply stemming on article’s keywords

Preprocessed Dataset

- Summary statistics of preprocessed dataset

	Min	Max	Mean	Median	Standard Deviation
Number of Keywords	1	16	5	5	2
Title Length	14	208	82	80	26
Abstract Length	128	5149	1411	1266	666
Text Length	2033	107028	22214	20505	10104
Year	2005	2021	2012	2011	4

Preprocessed Dataset (cont.)



Topic Identification Model

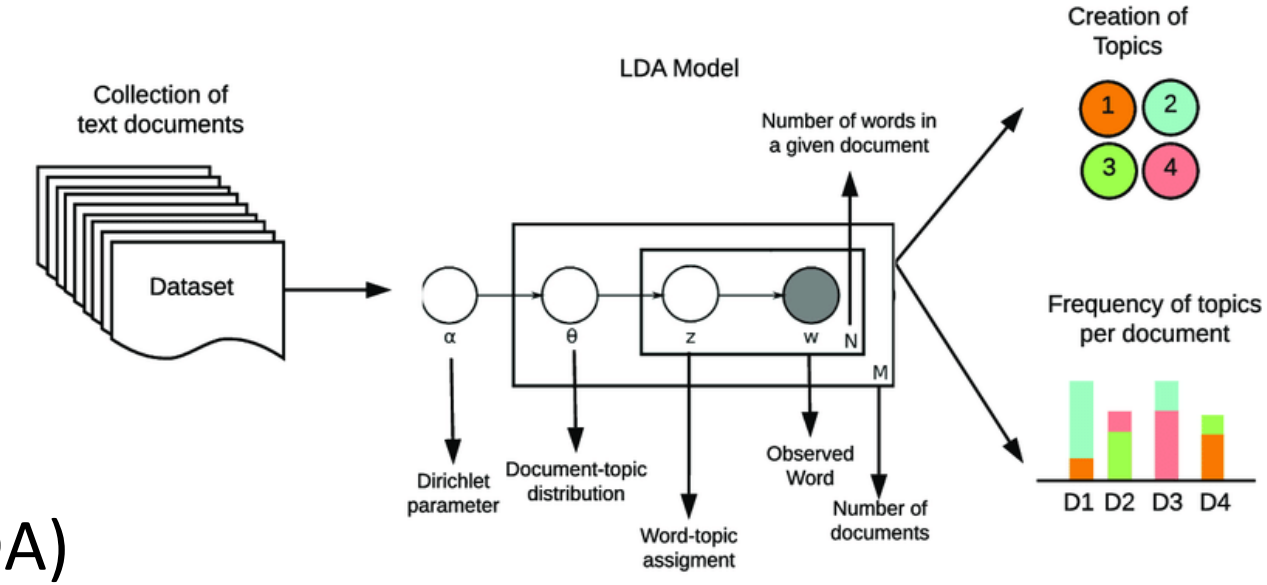
Topic Identification: Model Development

Libraries used:

- gensim (for model)
- pyLDAvis (for visualization)

Underlying model:

- Latent Dirichlet Allocation (LDA)

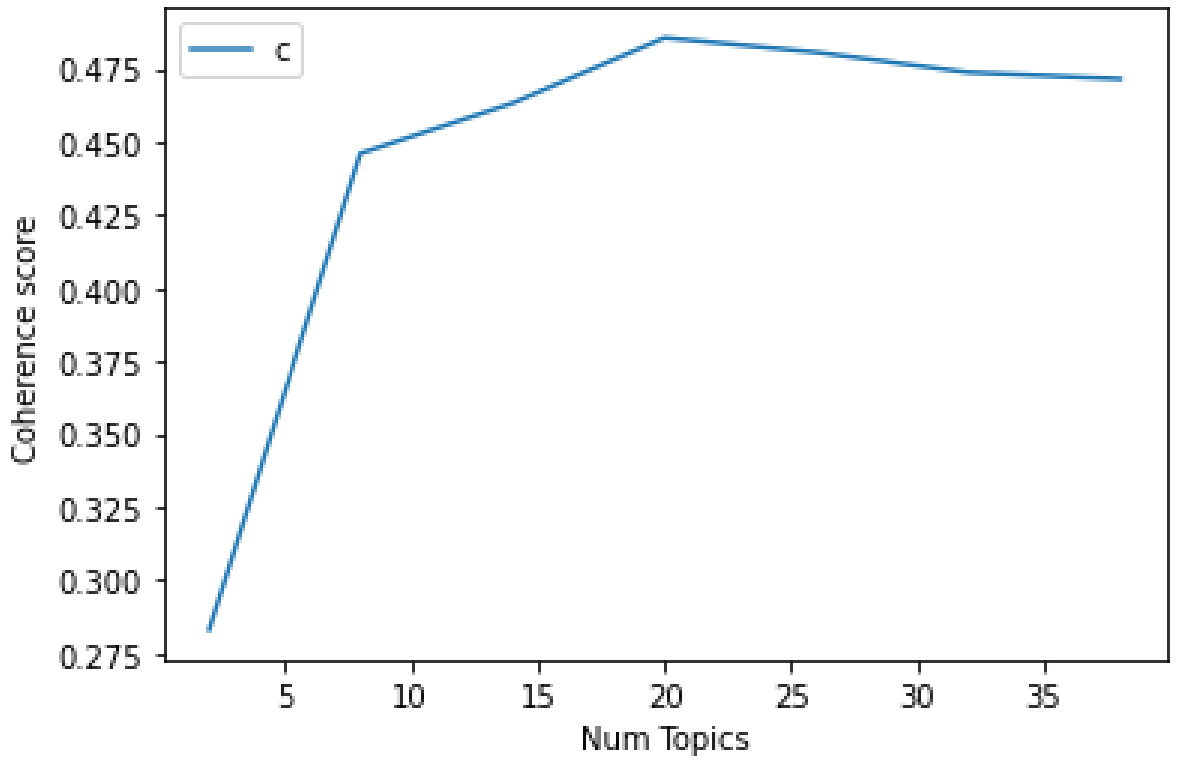


Source

Topic Identification: Model Results

- Initial Scores (num_topics = 35)
  - Perplexity: -7.2397
  - Coherence: 0.45696
- Finding Optimal Num_topics
  - Num Topics = 2 has Coherence value of 0.2831
  - Num Topics = 8 has Coherence value of 0.4461
  - Num Topics = 14 has Coherence value of 0.4632
  - Num Topics = 20 has Coherence value of 0.4856 <- Optimal
  - Num Topics = 26 has Coherence value of 0.4806

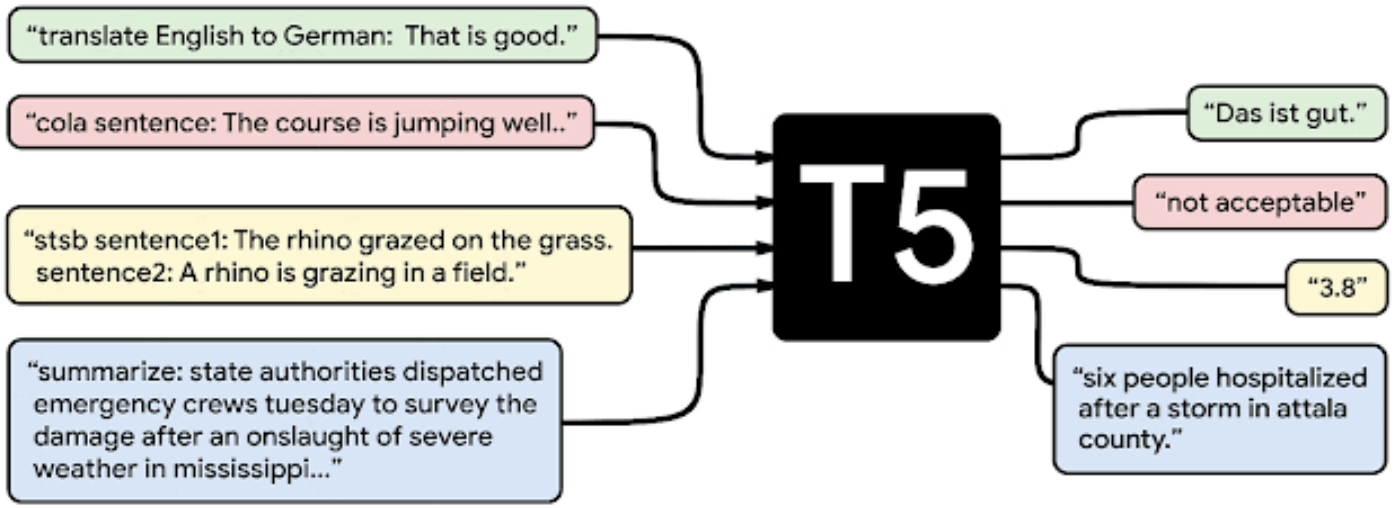
Topic Identification: Finding Optimal Num_topics



Text Summarization Models

Text Summarization: Model Development

- Two models:
  - Statistical model
  - Google's T5



Text Summarization: Statistical Model

- Structure:
  - Count token frequencies
  - Score sentences by term frequency (divide sum by number of words)
  - Determine threshold
  - Sequentially add sentences who score higher than the threshold

Text Summarization: Google’s T5

- Details:
  - Can be accessed through Hugging Face
  - Transfer learning model using transformers
  - Trained on C4 corpus
  - Can be used for classification, translation, question answering, and summarization
  - Input encoded, tokenized text
  - Outputs summary

Text Summarization: Model Results

Output of Statistical model:

```
Pallottino et al. (2002)  Yudhi Purwananto et al. Alonso-Ayuso et al. Alonso-Ayuso et al. (2011)
created groups of airplanes based on altitude levels with 1,000 ft distance between the groups.
Alonso-Ayuso et al. CPLEX is also a Mixed-Integer Programming (MIP) solver. Yudhi Purwananto et al. The
safe distance is measured in NM (Pallottino et al., 2002). 1a. 2a). 1a) is avoided. 1b). 1c). 1d). 2b).
3a). 3a). (a) (b)  Fig. 4a). 4b). 1d).
```

Output of Google's T5:

```
<pad> air traffic control system aims to increase the safety of the airplane passengers. the aim of the
CDR is to create a standard procedure to help the airplane controller. the proposed solution is called
the Velocity and Altitude Change (VAC) model. it uses a mixed-integer linear optimization (MILO)
approach to avoid conflicts.</s>
```

Text Summarization: Model Results

- Statistical Model:
  - Favors common words (in this case - Dates, names)
  - Good length
  - Lots of irrelevant information (because it favors common references, etc)

- Google's T5:
  - Excellent summary
  - Good length
  - Generates unique sentences

Web Application Demo