Topic Modelling in Knowledge Management Documents BPS Statistics Indonesia

M Y Hendrawan[1], **N W K Projo**[1]

[1]Statistical Computing Department, Politeknik Statistika STIS, Jakarta, Indonesia

*Corresponding author's e-mail: 221709869@stis.ac.id

Abstract. Knowledge management is an important activity in improving the performance an organization. BPS Statistics Indonesia has recently implemented such a system to improve the quality and efficiency of business processes. The purposes of this research are: 1) implementing topic modelling on BPS Knowledge Management System to identify groups of document topics; 2) providing recommendations on which the best topic modelling; 3) building a web service function of topic modelling for BPS that includes data preprocessing function and topic group recommendation function. This study applies the Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) topic modelling methods to determine the best grouping techniques for knowledge management systems in BPS Statistics Indonesia. The results show that the LDA model using Mallet is the best model with 25 topic groups and a coherence score of 0.4803. The performance result suggest that the best modelling method is the LDA. The LDA model is then successfully implemented in RESTful web service to provide services in the preprocessing function and topic recommendations on documents entered into the Knowledge Management System BPS.

1. Introduction

Knowledge is essential for accelerating the development and improvement of the quality of an organization. Knowledge and information from the organization need to be documented and managed through the knowledge management. Knowledge management can help organizations prepare an environment where members can create, share, and use explicit and tacit knowledge [1]. An organization can share knowledge between individuals through good knowledge management to take policies and solve problems appropriately.

BPS Statistics Indonesia (BPS) is a Non-Ministerial Government Institution whose task is to provide complete, accurate, and up-to-date statistical data to realize a reliable, effective, and efficient National Statistics System (NSS) to support national development [2]. As a government organization providing official statistics data, BPS Statistics Indonesia has implemented knowledge management. The development and implementation of knowledge management are pillars of institutional empowerment development activities in the Statistical Capacity Building Change and Reform for the Development of Statistic (STATCAP-CERDAS) BPS program to improve the effectiveness and efficiency of BPS's business processes. The implementation of knowledge management in BPS can be seen through the knowledge management system (KMS) as a BPS bureaucratic reform change program in institutional governance [3]. So every employee or party in need can access necessary knowledge and information related to the implementation of activities in BPS.

The recent development of the knowledge management system in BPS supports the Indonesia Data Hub (INDAH) program. INDAH is an integrated data platform to improve data literacy and the value of statistics. Besides that, it supports data interoperability and data exploration collaboration in Indonesia. These facts show that Knowledge Management System in BPS is essential and supports many BPS programs and activities. Therefore, excellent and efficient management of the system is needed.

With the increasing number of activities such as surveys, censuses, and other data collection carried out by BPS, the number of knowledge and information managed in the BPS Knowledge Management System is increasing. The process of tracing knowledge is needed to help employees access the necessary knowledge and information. One way to do this is to search by category or group of available documents. So far, the documentation of knowledge and information in KMS BPS has been grouped by type. The grouping uses hard clustering, in which documents can only be grouped into one particular kind of group, even though these documents can contain several topics discussed so that they can be part of several groups. Such grouping is still done manually, and the number of categories is limited. For example, a document titled "ICS Application For CAPI PK SP2020" is only classified in the CAPI category, even though it can also be classified in the SP2020 category.

Several methods can be used to group documents. One approach that can be taken is to use a text mining approach based on the topic of the document, called topic modelling. Topic modelling can be used to identify words from various papers and link documents with the same pattern based on the distribution of each word of the document [4]. Topic modelling is considered a suitable solution to this problem because it is included in fuzzy clustering [5], which can identify groups or related topics in the BPS Knowledge Management System that will facilitate the use of such knowledge.

Several algorithms of the topic modelling method are used in text analysis, including Latent Semantic Analysis (LSA), Non-Negative Matrix Factorization, Probabilistic Latent Semantic Analysis (PLSA), and Latent Dirichlet Allocation (LDA) [6]. This study focuses on implementing and comparing Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) methods. Both methods are chosen because they are most commonly used in previous studies and have represented two groups of topic modelling methods, the non-probabilistic group for LSA and the probabilistic group for LDA. Furthermore, these two methods have been compared to model the text data for short to long text documents which have similarities with knowledge management documents, such as films review, abstracts of health documents [6], and railroad accident text [7]. The implementation of both approaches will be evaluated to get the best document topic grouping for the BPS Knowledge Management System. The best method will then be implemented in web service functions to provide easy access and implementation for BPS.

The LSA modelling techniques are superior and appropriate to LDA in providing a choice of film recommendations to be watched by users [8]. The LDA model is better than the LSA model in grouping topics in electronic books [9]. The comparative study of LSA and LDA research for the analysis of railroad accident text also found that the two methods complement each other because each method produced several topics that were not identified by the other methods [7].

Research on the development of the LDA model using the MAchine Learning for LanguagE Toolkit (Mallet) is also a reference in this study. The topic modelling using LDA with Mallet provided a higher coherence score evaluation value than LDA methods with standard Gensim packages in grouping topics in documents in the form of related articles "multi-tier supply chain in Industry 4.0" [10]. The implementation of topic modelling using LDA with Mallet successfully grouped ten topics related to job trends in the information technology sector based on the information available on "LinkedIn"[11].

2. Methods

2.1. Scope of Research

This research focuses on topic modelling using Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) methods. Implementation of topic modelling with the LSA method in this study will use the Gensim package. In contrast, implementation with the LDA method will use two package,

namely Gensim and Mallet. They are implemented in Python programming language. The results will be evaluated, and the method judged to be superior will be applied in a web service for data preprocessing and topic recommendation functions to provide easy access and implementation for BPS.

2.2. Source and Data Collecting Methods

This research uses secondary data of all documents based on Bahasa Indonesia as knowledge and information available in the BPS Knowledge Management System from the beginning of the system implementation on December 4, 2019, until February 1, 2021. Two hundred seven documents are obtained through the web scraping method to extract text data from the BPS Knowledge Management System website. All documents are in Indonesian. The web scraping method in collecting this data has received permission from the Statistical Information System Integration Function BPS as the BPS Knowledge Management System manager. The data collection flow using the web scraping method carried out in this study can be seen in Figure 1.

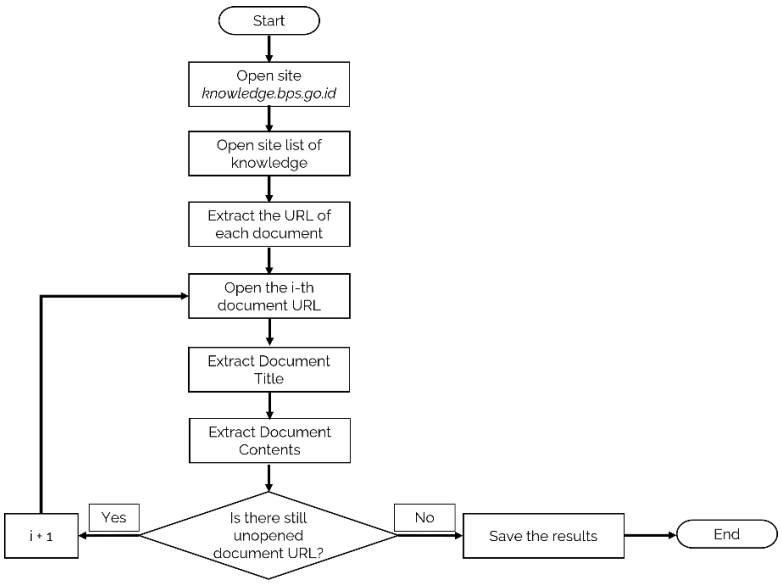

Figure 1. Flowchart Data Collection Using Web Scraping Method.

In this research, the web scraping method uses the Beautiful Soup package. It starts with opening the knowledge.bps.go.id page followed by opening a page containing a list of knowledge to extract each document's Uniform Resource Locator (URL). After all the document URL lists are collected, the list is opened to perform an extraction of the title and content of the knowledge management document. An example of web scraping data can be seen in Table 1. The title and content are successfully extracted without including the HTML tag.

Table 1. Example of data from web scraping.

URL	Title	Content
https://qasp2020.bps.go.id/posts/003d3cb419b0446996deefd5d96a1a7d/capi/akses-coolsis.bps.go.id	*AKSES COOLSIS.BPS.GO.ID [CAPI]*	*Untuk mengakses coolsis.bps.go.id jaringan internet harus terhubung melalui VPN BPS.*
https://qasp2020.bps.go.id/posts/6cf90b90-18c3-411f-8780-09cbbb569b01/instrumen/sp2020	*SP2020 [INSTRUMEN]*	*C1 hanya untuk 6 orang anggota rumah tangga, jika lebih dari itu, dapat diberikan C1 tambahan*

2.3. Data Preprocessing

Data preprocessing is an early stage in text mining that aims to convert text from human language into a machine-managed format, compose unstructured text, and maintain keywords useful for representing topics [12]. The stages of preprocessing this research data, as seen in Figure 2, begin with performing text normalization, which includes converting text to non-capital letters/case folding, removing special characters and punctuation, and removing white space. After that, the text will go through a process of removing words that often appear (stopwords), converting words into essential words (stemming), and tokenization (tokenization). Furthermore, the text transformation stage converts text data to the appropriate format by forming bigram models and trigrams to group frequently simultaneous words, including the corpus and data dictionary through the Bag of Words (BOW) method. The word weighting method is BOW because both LSA and LDA generative processes use the BOW word weighting approach regardless of the word sequence. Until now, the dominant method is still based on Bag of Words, where the corpus is converted into a word-document matrix, and the order of terms can be ignored [6].

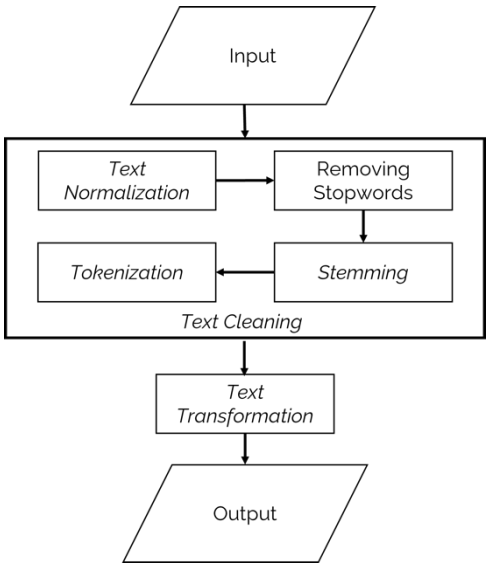

Figure 2. Flowchart of Data Preprocessing.

2.4. Topic Modelling

The topic modelling stage is carried out after the data has gone through the preprocessing phase and is ready for further processing. We apply two types of modelling methods: LSA and LDA. The first method, LSA, is an algebraic statistical method based on the application of Single Value Decomposition (SVD), which presents the semantic space of the document [6]. By applying this method, LSA can extract the structure and relationship of hidden words in the document through text vector representation to calculate the similarity between the texts and find similar words. The implementation of LSA topic modelling in this study will use the Gensim package.

The LDA is a probabilistic generative modelling topic model designed to extract topics from text [12]. It represents the documents as a random mixture of hidden topics characterized by a set of probabilities that define the word included in a topic. The implementation of topic modelling with the LDA can be implemented through a variational Bayes approach or a Gibbs sampling approach [13]. Therefore, the implementation of the LDA method in this study will be carried out using these two approaches using the Gensim for the variational Bayes approach and the Mallet for the Gibbs sampling approach. Even though Mallet and Gensim are two different package, the implementation of Mallet in this study will use a link function, namely "gensim.models.wrappers.LdaMallet" provided by the Gensim package.

Before implementing topic modelling using LSA and LDA methods, each method will determine the best parameters for modelling through the parameter tuning process. After a series of topic

modelling processes are carried out, the resulting topic models need to be evaluated to see the level of effectiveness in grouping topics by calculating the coherence score. The coherence score can measure the degree of semantic similarity between high-scoring words in the topic by distinguishing semantically interpretable topics and artifacts of statistical inference [14]. The coherence score can be evaluated by several methods. We apply the C_V algorithm [15], which found that C_V provides the strongest correlation between the value of coherence evaluation and the results of human interpretation related to the resulting topic. In addition, the distribution of topics generated by topic modelling using the LDA method will be evaluated through visualization using the pyLDAvis module. The evaluation is not carried out on the modelling results using the LSA modelling because a similar module is not available.

2.5. Web Service

The web service function development is the last stage in classifying documents. The services developed and provided in the web service function in this study include the function of the text data preprocessing process and the function to recommend topic groups in new documents. The web service is developed using the Representational State Transfer (REST) architecture.

This research uses the REST architecture because, in its implementation, it runs through a simple Hypertext Transfer Protocol (HTTP). It also allows support for several data formats, including the JSON data format. The REST architecture on web services based on the Flask microframework can produce better performance than SOAP in terms of requests and responses for web services [16]. The RESTful web service architecture implementation in this research will use Flask microframework. Flask can appropriately handle HTTP request functions, simple and light to run compared to other python-based web frameworks.

3. Results

3.1. Preprocessing Results

After the data collection process is complete, the document with the title and the content must be preprocessed. The input in this preprocessing stage is the title text and content of the knowledge management document. The input of the title and content text that has been preprocessed will be combined into one part. The first stage includes text cleaning, and the sample results are in Table 2.

Table 2. Example of data before and after text cleaning.

Before text cleaning	After text cleaning
Judul: AKSES COOLSIS.BPS.GO.ID [CAPI] Konten: Untuk mengakses coolsis.bps.go.id jaringan internet harus terhubung melalui VPN BPS.	*Judul: akses coolsis bps go id cap Konten: akses coolsis bps go id jaring internet hubung vpn bps*

The text cleaning stage is followed by preprocessing and transformation of the text by bigram and trigram models and creating a data corpus using the Bag of Words method. The data that has been preprocessed can be used to model the topic. The data structure of the preprocessed corpus that is ready for modelling can be seen in the bold characters in Table 3. While the characters that are not in bold in Table 3 represent the dictionary of words and identities in the resulting corpus.

Table 3. The structure of the corpus after preprocessing.

Corpus
[(0, 2), (1, 3), (2, 1), (3, 2), (4, 2), (5, 1), (6, 1), (7, 1), (8, 1)] *[('akses', 2), ('bps', 3), ('cap', 1), ('coolsis', 2), ('go_id', 2), ('hubung', 1), ('internet', 1), ('jaring', 1), ('vpn', 1)]*

3.2. Topic Modelling Using LSA Methods

In implementing LSA topic modelling with the Gensim package, one of the crucial parameters to determine is the number of topics (k). It is necessary to perform tuning parameters through coherence score calculation.

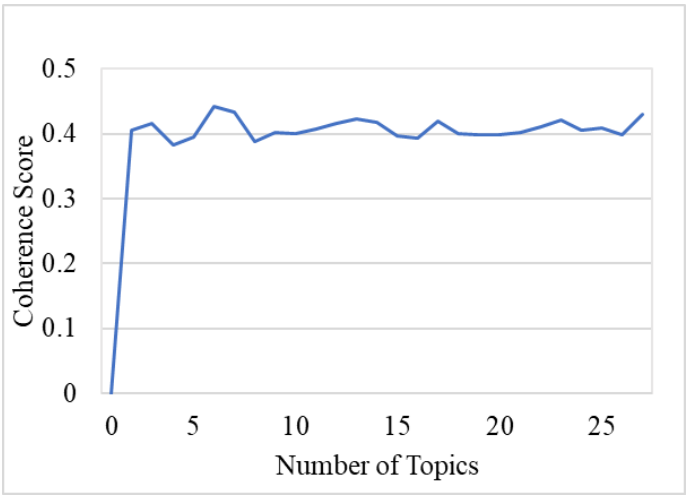

Figure 3. Result of LSA tuning parameter.

The number of topics in the parameter tuning process is carried out by iterations and calculating coherence scores from 2 to 27 topics. Figure 3 shows the coherence score of the tuning parameter results applied to the LSA model. As shown in the Figure 3, the model with six topics applied in the LSA has the highest coherence score of 0.4412. The distribution of the top five words that compose the six topics and independent interpretations related to these topics can be seen in Table 4. The results of the LSA modelling still have word distributions in related topics, such as in topic one with topic three and topic two with topic five.

Table 4. List of topics and top five word distributions from the best LSA topic modelling.

Topic	5 Top Words	Topic Discussion
T1	*keluarga, isi, anggota, duduk. tinggal*	Related to family data collection in the census
T2	*task, keluarga, kerja, klik, tombol,*	instructions for operating the device
T3	*duduk, sensus, sp, keluarga, anggota*	population census activities in general
T4	*properti, analisis, harga, jual, commerce*	Sales-related activities
T5	*klik, task, action, isi, duduk*	instructions for operating the device
T6	*air, rumah, lantai, tempat, listrik*	the house and its components

3.3. Topic modelling Using LDA Methods

The implementation of topic modelling using the LDA method uses the Gensim. The tuning parameters in LDA Gensim with α=0.9, β=0.3, and 16 topics give the highest coherence score of 0.4825. These parameters are implemented in the LDA model. Still, after visualizing the distribution of topics using pyLDAvis, the resulting topics tend to cluster and overlap in one quadrant, as shown in

Figure 4. To get better distribution, this LDA model is developed by applying the LDA model using Mallet.

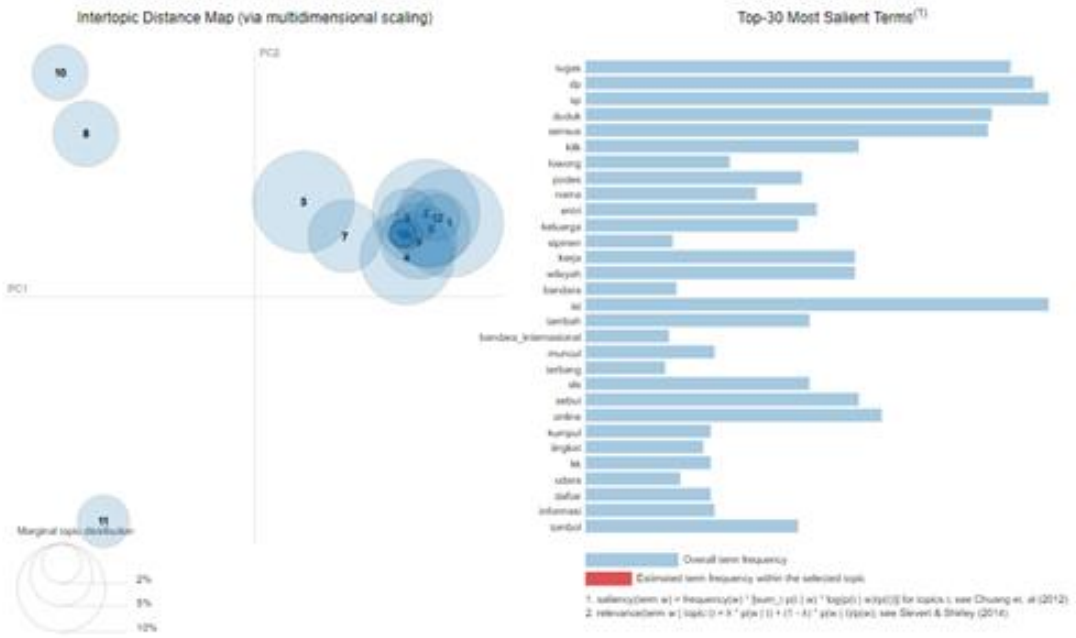

Figure 4. Visualization of 16 topics distribution by LDA Gensim.

Implementing the LDA method with the Mallet model uses tuning parameters to find optimum topics based on coherence score. The best model is as many as 25 topics with a coherence score of 0.4803, not much different from the previous LDA model. However, the distribution and separation of topics produced are much better with a more diverse distribution, although some topics intersect, as seen in Figure 5.

Figure 5. Visualization of 25 topics distributed by LDA Mallet.

The distribution of the top five words with the most frequency in arranging the 25 topics and interpretations of those topics can be seen in Table 5.

Table 5. List of topics and top five word distributions from the best LDA topic modelling (all documents are Indonesian)

Topic	5 Top Words	Topic Discussion	Topic	5 Top Words	Topic Discussion
T1	*tombol, task, tekan, nama, tambah*	The instructions for duties of officers	T14	*sls, kode, periksa, dokumen, bila*	The problem of checking SLS documents
T2	*isi, kirim, baik, perlu, konfirmasi,*	The Instructions for sending tasks	T15	*cap, ics, update, status, aplikasi*	The use of capi and ics
T3	*jumlah, indonesia, besar, bandara, statistik*	The use of statistics in Indonesia	T16	*dp, entri, entry, urut, nomor*	About Resident entries
T4	*keluarga, anggota, sebut, nama, kepala*	The entry of family members in the population census	T17	*milik, rumah, air, sewa, sumber*	The characteristics of the residents' residence
T5	*online, kk, isi, nik, lengkap*	Filling in the identity in the Online Population Census	T18	*assign, sampel, bahasa, mungkin, proses*	Sample problems in field activities
T6	*duduk, sensus, online, sp, informasi*	Related to the Online Population Census	T19	*klik, browser, tombol, baru, muncul*	Instructions for using web-based applications
T7	*tinggal, alamat, isi, tempat, sama*	Filling in the resident's address	T20	*giat, bantu, tetap, apa, terima*	Related to activities in general
T8	*tanggal, mulai, situs, covid, jumlah*	Activities in the Covid pandemic	T21	*perlu, beda, rupa, salah, aplikasi*	Related to app issues
T9	*bps, go_id, akses, dashboard, online*	Related on applications and online sites	T22	*kerja, pilih, beri, sesuai, hasil*	Related to work in general
T10	*klik, grup, kaizala, action, admin*	Instructions for using the Kaizala app	T23	*buka, aplikasi, buat, meeting, minta*	the online meeting mechanism instructions
T11	*wilayah, satu, tingkat, baik, sebut*	Activities that use the regional level	T24	*indonesia, surat, orang, nikah, nomor*	Content of marriage in the Population Census Question Questionnaire
T12	*analisis, properti, kondisi, commerce, bulan*	Related activities related to sales	T25	*tugas, sebut, login, mitra, password*	Instructions about officer application login
T13	*sp, tugas, pk, koseka, laksana*	Implementation of Koseka's duties in the population census			

3.4. Evaluation of LSA and LDA Methods

Implementing KMS BPS document topic modelling using the LSA method produces the best model with six topics and a coherence score of 0.4412. While the implementation of topic modelling using the LDA model gets the best model with Mallet model as many as 25 topics with a coherence score of 0.4803 and a better spread of topic groups. Table 6 is the summary of the topic modelling results of this study.

Table 6. Summary of topic modelling evaluation.

Variable	LSA	LDA-Gensim	LDA-Mallet
Number of Topics	6	16	25
Coherence Score	0.4412	0.4825	0.4803
Visualization	Not Available to Visualize	Tend to Cluster in One Quadrant and Overlap	More Widespread Topic

The evaluation based on the coherence scores of the three implementations of the topic modelling method, topic modelling using the LDA method, both LDA with Gensim and LDA with Mallet, have better scores than the LSA method. The coherence score from the LDA topic modelling generated using the Mallet package has a smaller value of 0.0022 points. It is not much different from the coherence score from the LDA topic modelling developed using the Gensim package. However, the LDA topic modelling with Mallet resulted in a better-distributed topic group characterized by a wider distribution of topics in the visualization results with pyLDAvis. The results are in line with the results of research by Zhou, Awasthi, and Cardinal which found that modeling the LDA topic with Mallet gave better grouping results with the LDA Gensim method in article documents related to "Multi-tier supply chain in Industry 4.0". Therefore, topic modelling for 25 topic groups using Mallet is the best model for grouping document topics in the BPS Knowledge Management System.

3.5. RESTful Web Service Development Modelling Topics

Having found the LDA method with Mallet as the best, we recommend it for the BPS Knowledge Management System. The model is implemented in the form of a RESTful Web Service function through the Flask microframework. The services developed in this research include text data preprocessing functions and topic group recommendation functions.

The service to provide text data preprocessing functions can receive input from client data in title text and content text from knowledge management documents. The service function for preprocessing this data can be accessed by calling the preprocess method via the GET method on the URL and entering the required document title and content input parameters. The output generated from this service function is a text data corpus that is ready to be used for further processing in text data processing and the status of the success of the request for the function in JSON file format. An example of a request for a preprocessing function, input parameters entered through the URL, and the output generated from this preprocessing service can be seen in Figure 6.

Request URL

```
http://127.0.0.1:5000/Topik_Model/preprocess?judul=AKSES%20COOLSIS.BPS.GO.ID%20%5BC
API%5D&konten=Untuk%20mengakses%20coolsis.bps.go.id%20jaringan%20internet%20harus%2
0terhubung%20melalui%20VPN%20BPS.
```

Response body

```
{
  "corpus hasil": [
    [
      "akses",
      "coolsis",
      "bps",
      "go_id",
      "cap",
      "akses",
      "coolsis",
      "bps",
      "go_id",
      "jaring",
      "internet",
      "hubung",
      "vpn",
      "bps"
    ]
  ],
  "status": "Dokumen Berhasil Dilakukan Preprocessing"
}
```

Figure 6. Example of request and output of service text data preprocessing.

Service functions to provide recommendations for topic groups that can be categorized as document groups also require input parameters in titles and knowledge management content. This function can be executed by requesting the method with the name *suggests* via the GET method on the URL. This function service provides output in JSON format to show the function status that is successfully executed. In addition, it includes a list of recommended topics based on the three highest probability topic groups, with topic one recommendation being the dominant topic included in the knowledge management document as input. The recommendation output of this topic can be the basis for managing and grouping documents recorded in the BPS Knowledge Management System. Examples related to requests and the output display generated by the topic group recommendation function can be seen in Figure 7.

Request URL

```
http://127.0.0.1:5000/Topik_Model/suggest?judul=ISIAN%20NOMOR%20SURAT%20ATAU%20AKTA%20PERNIKAHAN%20%5BSP2020-ONLINE%5D&konten=Nomor%20yang%20manakah%20yang%20diisikan%20pada%20pertanyaan%20%22Nomor%20Surat%2FAkta%20Pernikahan%22%20di%20web%20sensus%20online%3F%20Isikan%20nomor%20yang%20dilingkari%20pada%20bagian%20bawah%20seperti%20pada%20Gambar%201%20untuk%20yang%20memiliki%20akta%20perkawinan%20dari%20Catatan%20Sipil.%20%20Sementara%20itu%2C%20isikan%20nomor%20yang%20berada%20di%20bawah%20kata-kata%20%22Kutipan%20Akta%20Nikah%22%20jika%20surat%20nikah%20berupa%20buku%20nikah%20yang%20didapatkan%20dari%20Kantor%20Urusan%20Agama%20(KUA)%2C%20biasanya%20penduduk%20yang%20beragama%20Islam.
```

Response body


```
{
  "hasil rekomendasi topik": [
    {
      "Topik 1": "Isian Terkait Pernikahan pada Kuesioner"
    },
    {
      "Topik 2": "Pengisian Identitas SP Online"
    },
    {
      "Topik 3": "Sensus Penduduk Online"
    }
  ],
  "status": "Dokumen Berhasil Dilakukan Preprocessing"
}
```


Figure 7. Example of request and output service topic of topic group recommendation.

The implementation of web service functions related to preprocessing text data and topic recommendations built into this research has been accessible online. In addition, to provide convenience for users in accessing web services, this research has provided documentation related to preprocessing functions and topic recommendation functions created using swagger UI. The documentation provides information related to the description of web service functions, the parameters needed, up to the type of data used, as shown in Figure 8.

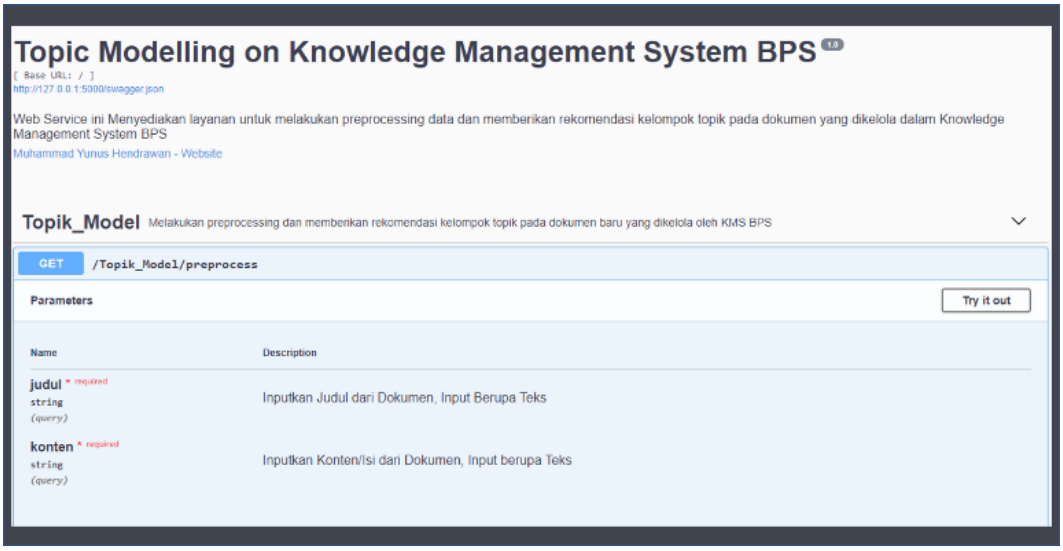

Figure 8. Web service documentation with Swagger

4. Conclusion

The BPS Knowledge Management System documents can be grouped using topic modelling methods. Based on the evaluation of modelling methods, the LDA model using Mallet is the best model with a coherence score of 0.4803 for 25 better topic groups distribution. This model is recommended to be used in the KMS BPS document.

The model is implemented in the form of functions in RESTful Web Service using Flask microframework in python. Service for the preprocessing stage of text data produces a corpus of data ready for the following text data processing. The service for the topic recommendation function will provide recommendations of topic groups loaded by knowledge management documents as the basis for grouping KMS BPS documents. Documentation related to the web service development has also been loaded using Swagger to ease user implementation. Through models and web services that have been built, Knowledge Management documents can be grouped in a fuzzy. The results of topic modelling using web services have been able to recommend a document into three dominant topic groups. The recommendation output of this topic can be the basis for managing and grouping documents recorded in the BPS Knowledge Management System to facilitate the search for related documents.

Acknowledgments

The authors would like to thank Politeknik Statistika STIS and Statistical Information System Integration Function BPS Statistics Indonesia for their full support regarding this research.

References

[1] Anggorowati Y 2018 Kajian reformasi birokrasi di Badan Pusat Statistik sebagai sistem terbuka *J. Ilm. Pem. Wid. Pra.* **44** pp 125–38

[2] Laws of The Republic Indonesia Number 16 1997 (Undang-Undang Republik Indonesia Nomor 16 Tahun 1997)

[3] Badan Pusat Statistik 2019 *Laporan Kemajuan Reformasi Birokrasi Badan Pusat Statistik Tahun 2018* (Jakarta: BPS)

[4] Nagaraja S and Chandrappa K Y 2019 Topic modelling *Int. J. of Inn. Tech. and Expl. Engg.* **9** pp. 482–5

[5] Arianto B W and Anuraga G 2020 Pemodelan topik pengguna twitter mengenai aplikasi "ruangguru" *J. Ilm. Das.* **21** pp 149–154

[6] Kherwa P and Bansal P 2019 Topic modeling: a comprehensive review *EAI End. Trans. on Scal. Inf. Syst.* **7** pp 1–16

[7] T. Williams and J. Betak 2018 A comparison of LSA and LDA for the analysis of railroad accident text *Pro. Comp. Sci.* **130** pp. 98–102
[8] Bergamaschi S, Po L and Sorrentino S 2014 Comparing topic models for a movie recommendation system *Proc. on 10th Int. Conf. on Web Information System and Technologies* pp. 172–83
[9] Mohammed S H and Al-Augby S 2020 LSA & LDA topic modeling classification : comparison study on e-books *Ind. J. of Elec. Eng. And Com. Sci.* **19** pp. 353–62
[10] Zhou R, Awasthi A and Stal-Le Cardinal J 2021 The main trends for multi-tier supply chain in industry 4.0 based on natural language processing *Comp. in Indust.* **125** 103369
[11] Athira M, Bhavya K, Soorya K, Ajeesh R and Anoop V S 2016 A new way of topic modeling using mallet for current job trends *Int. J. of Adv. Res. in Comp. and Comm. Eng.* **5** pp 59–63
[12] Blei D M, Ng A Y and Jordan M I 2003 Latent dirichlet allocation *J. of Mach. Learn. Res.* **3** pp 993–1022
[13] Rehurek R and Sojka P 2010 Software framework for topic modelling with large corpora *Proc. of the LREC 2010 Workshop on New Challenges for NLP Frameworks* pp 45-50
[14] Stevens K, Kegelmeyer P, Andrzejewski D and Buttler D 2012 *Proc. Of the 2012 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* pp 952–961
[15] Röder M, Both A and Hinneburg A 2015 Exploring the space of topic coherence measures *Proc. Of 8th ACM Int. Conf. on Web Search and Data Mining* pp 399–408
[16] Putra M G L and Putera M I A 2019 Analisis perbandingan metode soap dan rest yang digunakan pada framework flask untuk membangun web service *SCAN J. Tek. Inf. dan Kom.* **14** pp 1–7