J. Intell. Comput. Health Inform. Vol. 3, No. 1, March 2022
ISSN: 2715-6923, pp.1~8 | DOI: 10.26714/jichi.v3i1.9205

Implementation of Named Entity Recognition with a Developing Question Answering System: A Case Study in the Merapi Volcano Museum

Arfiani Nur Khusna*, Okhy Kharisma Putri

Department of Informatics, Universitas Ahmad Dahlan, Jl. Ringroad Selatan, Kragilan, Tamanan, Kec. Banguntapan, Kabupaten Bantul, Daerah Istimewa Yogyakarta 55191, Indonesia
*Corresponding author: arfiani.khusna@tif.uad.ac.id

Dimas Chaerul Ekty Saputra

Department of Biomedical Engineering, Universitas Gadjah Mada, Jl. Teknika Utara, Pogung Kidul, Sinduadi, Kec. Mlati, Kabupaten Sleman, Daerah Istimewa Yogyakarta 55284, Indonesia

Article histroy :
Received: 27 Feb 2022
Accepted: 1 Nov 2022
Avalaible online: 7 Nov 2022

Research article

Abstract: Merapi volcano museum is a place to get some information about active mountain activities, the general public can access the website page at mgm.slemankab.go.id. Indeed, visitors are given easy access, but the information provided by the website is not fully complete, causing visitors to feel dissatisfied. Based on the results of a questionnaire from 40 respondents, it was found that 50.55% of website visitors did not get the information they wanted. Therefore, in this research, we built a Question Answering System (QAS) using the Named Entity Recognition (NER) method that has been implemented into Telegram. To improve the performance of the QAS system, testing and analysis has been carried out with a "white box" approach. The results show that the QAS system has 3 regions and 3 independent paths, with path 1 being 1-2-3-4-11, path 2 being 1-2-3-4-5-6-7-8-11, and path 3 being 1-2-3-4-5-6-7-9-10-11. Based on the results of this study, all three paths can produce the correct answer.

Keywords: NER; QAS; MERAPI VOLCANO MUSEUM; WHITE-BOX TESTING; DISSATISFIED INFORMATION

1. Introduction

Museum is a place that is used to store or accommodate all kinds of historical relics that can be known and studied (Brown & Mairesse, 2018). Mount Merapi Museum is one of the places used as a means of knowledge about volcanoes. Indonesia is located in the Ring of Fire area, making Indonesia has 129 active volcanoes. This museum provides information related to Mount Merapi in the form of eruption relics, documentation, display of Mount Merapi's miniature artists, earthquake simulation rooms and so on. Along with the development of technology, the Mount Merapi Museum currently has a website with the name Merapi Volcano Museum (mgm.slemankab.go.id). The Merapi volcano museum website provides general information related to the museum and visitors can obtain this information by accessing the website, opening each menu tab on the website to get the information they are looking for, then visitors reading the information. The information available on the website cannot be obtained by visitors quickly and accurately, because visitors need to access one by one page on the website and find the desired information.

Therefore, to obtain information that is concise and precise with what visitors want, a question-and-answer system is made that is able to provide answers in the form of information to visitors without having to open each menu tab page, so that visitors can quickly get information and understand the information (Gusmita et al., 2014). The system is expected to be able to display answers in the form of short text according to the questions that visitors have asked. One way is to build a Question Answering System with the object of the Merapi Volcano Museum so that it can make it easier for visitors to find answers to the information they are looking for quickly and efficiently.

Question Answering System (QAS) or question and answer system is a system that can provide direct answers to users (Sapitri & Al-faraby, 2018). Question Answering

System provides correct and precise answers. In general, the question-and-answer system consists of three components, namely Question Analysis, Passage Retrieval and Answer Extraction. Named Entity Recognition (NER) method is the core of information extraction and involves processing of structured and unstructured documents (Yadav & Bethard, 2019). Identification expressions refer to people, places, organizations, companies, dates and times. Where the purpose of the Question Answering System (QAS) is to get and return the correct and appropriate answers to answer questions from users (Yu et al., 2020).

Based on the description of the background above, a study was proposed by applying the Question Answering System with the Named Entity Recognition method in answering questions and getting answers directly and precisely with the title "Question Answering System with Named Entity Recognition at the Merapi Volcano Museum".

2. Methods

2.1 Information retrieval

Information retrieval system according to another opinion is a process related to the representation, storage, retrieval, and retrieval of information that is relevant to the information needs of users (Khusna & Agustina, 2018). This statement indicates that the information retrieval system contains the process of storing, providing, representing, identifying, and searching or searching for relevant documents in a database, in order to meet the information needs of users.

2.2 Question answering system

Question Answering System (QAS) is a computer system with the ability to answer questions posed using natural language by utilizing natural language processing techniques, such as Information Retrieval and Information Extraction (Luan et al., 2019). The question-and-answer system aims to find the accuracy and consistency of answers to a question that is in accordance with the answer document.

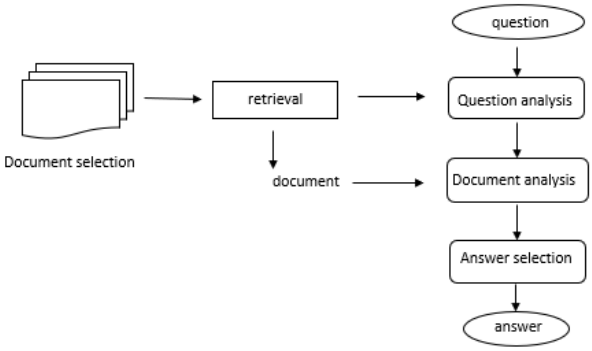

Fig 1. Flowchart of the proposed QAS

QAS assist users in expressing information needs in a more specific and natural form, for example users want to know the price of admission to the Merapi volcano museum, users can provide input questions such as "When is the inauguration of the Merapi volcano museum?" on the QAS and the user gets *output* in the form of a text "Museum was inaugurated on October 1, 2009 by the Minister of Energy and Mineral Resources Purnomo Yusgiantoro" as an answer. It is different with Search Engines, if the user inputs the question above as a query, the search engine will return output in the form of a list of documents containing the query keyword, and to find the desired information or answer, the user must open, read, and filter every information in the document by carefully. The our proposed QAS framework can be seen in Fig. 1.

2.3 Question analysis

The stage for processing user questions according to the type of answer expected and to get keywords from the question. The questions given are related to the discussion related to the Merapi volcano museum in Indonesia. Question analysis is performed to determine whether the question sentence began with a question word and contains the question words what, when, where, and how much. The question pattern produces the target, context, and property, so that the question pattern and the answer pattern will be interrelated. Target answers based on what question words (object), when (time), where (location), and how much (count) they contain. The questions analysis can be seen in Table 1.

Table 1. Analysis of question

FEATURE	DESCRIPTION
Question mark	When ?
Question	When is the inauguration of the Merapi Volcano Museum ?
Question Pattern	(When) <T> <C> ?

When <C> is Context, <T> is Target and <P> is Property

2.4 Preprocessing

In this research, data preprocessing is applied as cleaning unnecessary words by separating each word, removing affixes, and removing words that are not important in the document. The preprocessing stages are divided into four main parts including one, two three, and four.

a. Text Normalization

Text normalization is used to convert text from uppercase to lowercase, removing symbols and characters other than the alphabet except those that have been specified (Pramanik & Hussain, 2019).

b. Tokenization

Tokenization is the process of converting document text into words or tokens. The tokens are separated by a space parameter (Prasad, 2021).

c. Stop-word Removal

Stopword is a filtering process, a removal approach namely the removal of words that often appear in documents, such as words and, yang, or, and others (Ladani & Desai, 2020).

d. Stemming

Please provide 4 to 6 keywords which can be used for indexing purposes

Stemming is a process to get the meaning of the word by changing the token into the root word first. Stemming is done to change the form of an affixed word into a basic word to match the good and correct word structure in Indonesian (Bougar & Ziyati, 2019). The preprocessing of data text can be seen in Tabel 2.

2.5 Answer extraction

Answer extraction is the process of finding documents that contain candidate answers from a given pattern of questions.

a. Document Segmentation

Document segmentation is a process of converting candidate answer documents into sentences (Zhang et al., 2021). This process uses a period as a marker between one sentence and another. The document segmentation can be seen in Table 3.

Table 2. Text preprocessing process

TOKENIZATION	FILTERING	STEMMING
When	When	When
Inauguration	Inauguration	Official
Museum	Museum	museum
Mountain	Mountain	Mountain
Fire	Fire	Fire
Merapi	Merapi	Merapi

Table 3. Document segmentation.

KEYWORDS QUESTION	COUNT	KEYWORDS ANSWER	COUNT
Official	1	Official	1
Museum	1 2	Museum	1 2

Table 4. Named entity recognition

QUESTIONS	WHEN IS INAUGURATION OF THE MERAPI VOLCANO MUSEUM?
Pattern Question	(when) <T> <C> <P> ?
Answer Pattern on DB	<QT> <E> <K> <J>
Candidate Answer	The inauguration of the museum was held on October 1, 2009 by the Minister of Energy and Mineral Resources Purnomo Yusgiantoro
NER Pattern Answer	[Inauguration-museum=object]held on[date-1-October-2009=time]by[Minister of Energy and Mineral Resources=object][Purnomo-Yusgiantoro=person]

b. Number of Keywords

After being a sentence separation process, then to the number of keywords process, which is a process of calculating the number of keywords that exist in each sentence (Wick & Puppe, 2018). The number of keywords process can be seen in notation (1).

$$Kj \geq \left|\sqrt{Kp-1}\right| + 1 \quad (1)$$
$$2 \geq \left|\sqrt{2-1}\right| + 1$$
$$2 \geq 0.414 + 1$$
$$2 \geq 1.414$$

where *Kp* is number of keywords in the question and *Kj* is number of keywords in the answer candidate.

2.6 Named entity recognition

Named Entity Recognition used to find and identify a number of entities from each candidate sentence such as person, location, time, object and count. Entities are based on existing question words such as "who" looks for a person entity, "where" looks for a location entity, "when" looks for a time entity (time), "what" looks for an object entity and "how" looks for a count entity. After the Named Entity Recognition process is carried out, the results of candidate answers are obtained, and if there are entities that do not meet the requirements based on the question words, they will not be processed. Analysis of NER can be seen in Table 4, where <T> is a Target, <C> is a Context, <P> is a Property, <QT> is a Question Tag, <E> is a Entity, <K> is a keyword, and <J> is a Answer.

2.7 White-box testing

The evaluation method is used to test how well the system finds and displays answers. This stage uses the white box testing method. White box testing is an

approach between testing derived from structural knowledge and software implementation. The white box uses the control structure of the procedural design (Syaikhuddin et al., 2018). There are four main stages of songho including flowchart notation, regions, independent path, cyclomatic complexity, test case design.

a. Flowchart Notation

Flowchart Notation (FN) is a program graph generated from mapping program flowcharts to represent the control flow of existing program logic (Ramos-Merino et al., 2018).

b. Regions

Regions (R) an area bounded by an edge and nodes. The area outside the graph is also a region (Azmi et al., 2018).

c. Independent Path

Independent Path (IP) is a path that traverses at least one new set of statements or a new condition in a program (Qiu et al., 2020).

d. Cyclomatic Complexity

Cyclomatic Complexity (CC) used to determine the number of independent paths which are the basic paths (Project, 2019).

d. Test Case Design

Test Case Design (TCD) is a set of scenarios that have been prepared so that the system to be tested can meet the conditions according to the test case (Garousi et al., 2020).

3. Results and Discussion

In this research, the dataset used is sourced from https://mgm.slemankab.go.id/. The dataset is stored in a question and answer system folder. The folder contains 4 kinds of files, includes of object, time, location, and number folders. The data system architecture and flowchart of QAS system can be seen in Fig. 2 and 3.

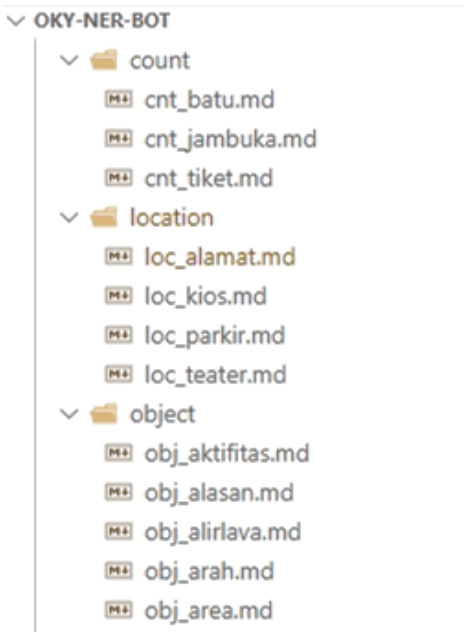

Fig 2. The architecture files dataset of QAS system

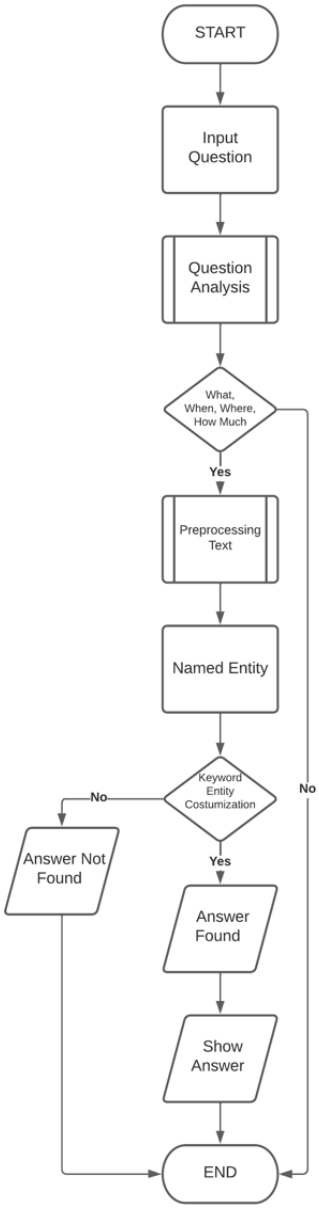

Fig 3. Flowchart of the proposed QAS system

As shown in Fig. 3, the testing begins in input or enter in the form of a question sentence. The question sentence begins with a question word which includes the words what, where, when, and how much. Furthermore, the system analyzes the questions that have been asked by the user. The question analysis stage is carried out to find out whether the question sentence contains a question word at the beginning of the sentence, if it has not started with a question word, the system will end and the user must re-enter the question starting with a question word (what, when, where and how much). The question analysis process is used to determine the pattern of existing questions, namely the existence of targets, contexts and properties.

The next stage is text preprocessing, this stage is done to process the question sentences using the tokenization, stemming and filtering processes. Then proceed to the named entity recognition stage, this stage is carried out to identify and adjust the entities contained in the question pattern with the existing entities in the database and the engine for the named entity method. The entity in question

is what word looks for object entities, when to look for time entities, where to look for location entities, and how much to look for count entities.

Show answers, at this stage the system will display answers that match the pattern of questions and entities that exist in the named entity. The answers displayed are in the form of museum-related information that has been stored in the system database. However, if the entity in the question sentence matches the entity in the named entity, the system will display the output "No answer found".

a. Trial Question

Question testing is done through a system where the user inputs questions to the telegram bot. The list of questions inputted into the telegram bot along with the accuracy of the answers generated by the system can be seen in Table 5, 6, 7, and 8.

- For ***objects***. Kinds of question is "*what*" can be seen in Table 5
- For ***times***. Kinds of question is "*when*" can be seen in Table 6
- For ***locations***. Kinds of question is "*where*" can be seen in Table 7
- For ***counts***. Kinds of question is "*how much*" can be seen in Table 8

Table 5. Sample of users question of objects "*what*"

QUESTIONS	ANSWERS APPROPRIATE	NO
What is the history of the museum ?	v	-
What is the profile of the museum?	v	-
What are the museum collections?	v	-
Is there a museum guide available?	v	-
What are the museum facilities?	v	-
What is the museum email?	v	-
What information is there in the museum?	v	-
What is the vision and mission of the museum?		v
What's on the first floor of the museum?	v	-
What's on the second floor of the museum?	v	-
What are the areas in the museum?	v	-
Is there a hotspot at the museum?	v	-
What is the use of a museum home theater?	v	-
What is the use of the museum auditorium?	v	-
What souvenirs are available in the museum?	v	-
What is the function of the open theater museum?	v	-
What are the directions to the museum?	-	v

Table 6. Sample of users question of times "*when*"

QUESTIONS	ANSWERS APPROPRIATE	NO
When o'clock open museums ?	v	-
When is the museum open?	v	-
When was the museum officially inaugurated?	v	v

Table 7. Sample of users question of locations "*where*"

QUESTIONS	ANSWERS APPROPRIATE	NO
Where is the address for the museum?	v	-
Where is the museum parking?	-	v
Where is the home theater museum?	-	v
Where is the museum's auditorium?	-	v
Where is the museum kiosk?	-	v
Where is the musholla of the museum?	-	v
Where is the open theater museum?	-	v

Table 8. Sample of users question of counts "*how much*"

QUESTIONS	ANSWERS APPROPRIATE	NO
How much do museum tickets cost?	v	-
What are the opening hours of the museum?	v	-
How many areas are there in the museum?	v	-
How big is the museum?	v	-
How many years has the museum been operating?	v	-
What is the phone number for the museum?	v	-
What is the capacity of the museum's home theater?	v	-

b. White-Box Testing

System testing using white box testing is done by making test cases obtained from independent paths. The independent path is obtained by calculating the cyclomatic complexity value of the flowgraph. The flowgraph is based on the flowchart of the question-and-answer system. Fig. 3 for white box testing flowgraph.

- Number of regions of the flow chart.
- *Cyclomatic complexity* V(G) for flow chart
  V(G) = E – N + 2
  V(G) = P + 1

Then:

- *Flowgraph* has 3 regions.
- V(G) = 10 edges – 9 nodes + 2 = 3.
- V(G) = 2 predicate nodes + 1 = 3.

So, the cyclomatic complexity for the flowgraph in Fig. 2 is 3. A high cyclomatic complexity value indicates a complex procedure that is difficult to understand, test and maintain. The cyclomatic complexity risk value can be seen in Table 9.

Table 9. Cyclomatic complexity risk value

CC	PROCEDURE	RISK
1 – 4	A simple procedure	Low
5 – 10	A well-structured and stable procedure	Low
11 – 20	A more complex procedure	Currently
21 – 50	A complex procedure, alarming	Tall
>50	An error-prone, extremely troublesome, untestable procedure	Very high

Question Answering System here according to the relationship of cyclomatic complexity and risk in a fairly complex procedure with a low level of risk. Based on the previous cyclomatic complexity calculation, an independent path will be determined. In this question answering system, there are 3 independent paths, including:

Path 1 = 1-2-3-4-11. Description: The user inputs a question, then the question is analyzed. If no question word is found, the system stops and the user inputs the question again. The question must begin with a question word.

Path 2 = 1-2-3-4-5-6-7-8-11. Description: The user inputs a question, the analysis of the question sentence contains a question word at the beginning of the sentence, the question is processed, the system successfully finds the answer and is displayed.

Path 3 = 1-2-3-4-5-6-7-9-10-11. Description: The user inputs a question, the analysis of the question sentence contains a question word at the beginning of the sentence, the question is processed, the system does not find an answer because the question pattern does not match the NER keyword entity. The 3 independent paths above will be tested with data, so that at least each independent path will be passed at least 1 time.

c. Test Case

Test Case Validation will be carried out by testing using the program. So, it is known that the expected results are in accordance with the results of the application. Testing case scenario can be seen in Table 10.

Table 10. Test case scenario

CODE	INFORMATION
V (Valid)	indicates the component in the scenario has a true or valid value, so the system is successful.
I (Invalid)	indicates the scenario component has an incorrect or invalid value.

- Independent path 1

Test of scenario:

Table 11. Independent test scenario 1

SCENARIO NAME	CONDITION
Enter the question sentence	Input the question sentence without starting with the question word

Test of data:

Table 12. Independent trial scenario 1.

SCENARIO NAME	ENTER THE QUESTION SENTENCE
Data	Since when was the museum inaugurated?
Expected results	The system displays "the sentence does not start with a question word, please repeat your question..."
Results	Show "the sentence does not start with a question word, repeat your question..."
Validation	v

- Independent path 2

Test of scenario:

Table 13. Independent test scenario 2.

SCENARIO NAME	CONDITION
Enter the question sentence	Input the question sentence starting with the question word

Test of data:

Table 14. Independent trial scenario 2

SCENARIO NAME	ENTER THE QUESTION SENTENCE
Data	When was the museum inaugurated?
Expected results	The system performs an answer search and displays the answer
Results	The system displays the answer
Validation	v

- Independent Path 3

Test of scenario:

Table 15. Independent test scenario 3

SCENARIO NAME	CONDITION
Enter the question sentence	The question sentence input begins with a question word, but the question pattern does not match the NER keyword entity.

Test of data:

Table 16. Independent trial scenario 3

SCENARIO NAME	ENTER THE QUESTION SENTENCE
Data	Since when was the museum inaugurated?
Expected results	The system displays “No answer”
Results	Show "No answer"
Validation	v

d. Test Case Results

For the case results can be seen in Table 17.

Table17. Test case results.

SCENARIO	SCENARIO RESULTS	TEST RESULT	IN ACCORDANCE
Enter the question sentence	You're welcome	You're welcome	v
Enter the question sentence	You're welcome	You're welcome	v
Enter the question sentence	You're welcome	You're welcome	v

Based on the Table 17, from the test results on the system, after being tested using data in the form of question sentence input with various possibilities, 3 test tests are valid or in Path with expectations.

4. Conclusion

A question and answer system has been built using the named entity recognition approach in the system. Named entity recognition is a process to find answers. Preprocessing is done for tokenization, filtering and stemming. The extraction results are then displayed as answers to the questions that have been asked, then tested using the white box testing method. From the test, 3 regions and 3 independent paths were obtained, with Path 1 = 1-2-3-4-11, Path 2 = 1-2-3-4-5-6-7-8-11, and Path 3 = 1-2-3-4-5-6-7-9-10-11.

The results of this study show that there are three paths that can return the correct answer to each user who asks. Based on the results of this study, it can be concluded that this method can provide the information needed by the user.

References

Azmi, N. S. A., Singkaravanit-Ogawa, S., Ikeda, K., Kitakura, S., Inoue, Y., Narusaka, Y., Shirasu, K., Kaido, M., Mise, K., & Takano, Y. (2018). Inappropriate expression of an NLP effector in Colletotrichum orbiculare impairs infection on cucurbitaceae cultivars via plant recognition of the C-terminal region. *Molecular Plant-Microbe Interactions*, *31*(1), 101–111. https://doi.org/10.1094/MPMI-04-17-0085-FI

Bougar, M., & Ziyati, E. H. (2019). Stemming algorithm for arabic text using a parallel data processing. *Advances in Intelligent Systems and Computing*, *797*(July), 261–268. https://doi.org/10.1007/978-981-13-1165-9_23

Brown, K., & Mairesse, F. (2018). The definition of the museum through its social role. *Curator: The Museum Journal*, *61*(4), 525–539. https://doi.org/10.1111/cura.12276

Garousi, V., Bauer, S., & Felderer, M. (2020). NLP-assisted software testing: A systematic mapping of the literature. *Information and Software Technology*, *126*, 1–29. https://doi.org/10.1016/j.infsof.2020.106321

Gusmita, R. H., Durachman, Y., Harun, S., Firmansyah, A. F., Sukmana, H. T., & Suhaimi, A. (2014). A rule-based question answering system on relevant documents of Indonesian Quran Translation. *2014 International Conference on Cyber and IT Service Management, CITSM 2014*, 104–107. https://doi.org/10.1109/CITSM.2014.7042185

Khusna, A. N., & Agustina, I. (2018). Implementation of Information Retrieval Using TF-IDF Weighting Method On Detik.Com’s Website. *TSSA-IEEE*.

Ladani, D. J., & Desai, N. P. (2020). Stopword Identification and Removal Techniques on TC and IR applications: A Survey. *2020 6th International Conference on Advanced Computing and Communication Systems, ICACCS 2020*, 466–472.

https://doi.org/10.1109/ICACCS48705.2020.9074166

Luan, Y., Wadden, D., He, L., Shah, A., Ostendorf, M., & Hajishirzi, H. (2019). A general framework for information extraction using dynamic span graphs. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 3036–3046. https://doi.org/10.18653/v1/n19-1308

Pramanik, S., & Hussain, A. (2019). Text normalization using memory augmented neural networks. *Speech Communication*, *109*, 15–23. https://doi.org/10.1016/j.specom.2019.02.003

Prasad, G. N. R. (2021). Identification of Bloom ' s Taxonomy level for the given Question paper using NLP Tokenization technique Turkish Journal of Computer and Mathematics Education Research Article Identification of Cognitive level of Question. *Turkish Journal of Computer and Mathematics Education*, *12*(13), 1872–1875.

Project, M. D. (2019). *N-Grams as a Measure of Naturalness and Complexity*. Department of computer science and media technology (CM), Digitala Vetenskapliga Arkivet.

Qiu, M., Housh, M., & Ostfeld, A. (2020). A two-stage LP-NLP methodology for the least-cost design and operation of water distribution systems. *Water (Switzerland)*, *12*(5), 1–21. https://doi.org/10.3390/W12051364

Ramos-Merino, M., Álvarez-Sabucedo, L. M., Santos-Gago, J. M., & Sanz-Valero, J. (2018). A BPMN Based Notation for the Representation of Workflows in Hospital Protocols. *Journal of Medical Systems*, *42*(10). https://doi.org/10.1007/s10916-018-1034-2

Sapitri, A. I., & Al-faraby, S. (2018). Analisis Metode Pattern Based Approach Question Answering System Pada Dataset Hukum Islam Berbasis Bahasa Indonesia. *Media Informatika Budidarma (MIB)*, *2*(4), 159–164. http://dx.doi.org/10.30865/mib.v2i4.950

Syaikhuddin, M. M., Anam, C., Rinaldi, A. R., & Conoras, M. E. B. (2018). Conventional Software Testing Using White Box Method. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, *3*(1), 65–72. https://doi.org/10.22219/kinetik.v3i1.231

Wick, C., & Puppe, F. (2018). Fully Convolutional Neural Networks for Page Segmentation of Historical Document Images of Historical Document Images. *In 2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, 287–292. https://doi.org/10.1109/DAS.2018.39

Yadav, V., & Bethard, S. (2019). A survey on recent advances in named entity recognition from deep learning models. *ArXiv Preprint ArXiv:1910.11470*. https://doi.org/10.48550/arXiv.1910.11470

Yu, W., Wu, L., Deng, Y., Mahindru, R., Zeng, Q., Guven, S., & Jiang, M. (2020). *A Technical Question Answering System with Transfer Learning*. 92–99. https://doi.org/10.18653/v1/2020.emnlp-demos.13

Zhang, N., Chen, X., Xie, X., Deng, S., Tan, C., Chen, M., Huang, F., Si, L., & Chen, H. (2021). *Document-level Relation Extraction as Semantic Segmentation*. 3999–4006. https://doi.org/10.24963/ijcai.2021/551