Jurnal Mantik, 6 (3) (2022), ISSN 2685-4236 (Online)

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik

Natural Language Processing Analysis Of Frequently Used Words On Indonesia Website Names

Novianti Madhona Faizah[1], Luky Fabrianto[2*], Widyat Nurcahyo[3], Herlina Trisnawati[4]
[1,3]Computer Science Department, Universitas Tama Jagakarsa, Indonesia
[2]Digital Business Department, Universitas Nusa Mandiri, Indonesia
[4]Information System Department, Universitas Tama Jagakarsa, Indonesia

A R T I C L E I N F O

Article history:

Received Sep 2, 2022
Revised Sep 19, 2022
Accepted Oct 11, 2022

Keywords:

Domain Name System
Name
Python
Website
Wordninja

ABSTRACT

The increasing of internet use time after time is makes an impact addition of websites, the name of a website must be unique, eye catching and attractive, and in naming a website it should not use spaces, therefore it is often found that the website name consists of several words which are combined. This study aims to determine the most frequently used words on websites in Indonesia. The stages of this research briefly begin with the collection of 10,960 website names, word separation on each website name consisting of several words using Wordninja (one of packages available in Python programming language). The word separation process is carried out in several stages, starting from words containing at least 3 letters to 9 letters. Furthermore, from the word separation stage, ten words that appear most often are sorted. It was found that the word "Indonesia" most often appears at each stage of word separation, which is 139 times. Conclusion of this study is prove that Wordninja were very effective, as evidenced by an accuracy of 97.2%.

Corresponding Author:

Luky Fabrianto,
Digital Business Department,
Universitas Nusa Mandiri,
Jl. Raya Jatiwaringin No.2, RW.13, Cipinang Melayu, Kec. Makasar, Kota Jakarta Timur, Daerah Khusus Ibukota Jakarta 13620, Indonesia.
E-mail: luky.lfb@nusamandiri.ac.id

1. INTRODUCTION

The growth of the internet is increasingly widespread with higher quality, especially when the whole world has experienced a pandemic for more than two years, but we should be grateful that it is slowly starting to gradually become endemic. These conditions encourage internet usage to increase massively, but even if the pandemic does not occur, internet usage growth will continue to experience an increasing trend. The use of the internet cannot be separated from websites. A website is a collection of pages in which some topics are interrelated between one page and another. This is usually placed on a server that can be accessed via a local network (LAN) or a global network (internet) (Susilowati, 2019).

A website must have an easy-to-remember address which is usually called a domain. Technically the address of a website is in the form of a complex combination of numbers (ex. 192.168.11.12) called an Internet Protocol (IP). If these numbers are not converted into a name, certainly not be easy to reach the website address. So that visitors of a website

can easily access it, a domain is needed. A domain is a unique name that is a conversion from an IP address (Kim & Reeves, 2020). A domain also has an extension, often called a suffix that reflects the image of a domain, such as ".com" which describes the website being used in the business field, or ".sch.id" which represents a website owned by schools in Indonesia, and so on (Giannakoulopoulos et al., 2022)(SINAGA, 2019).

A study of English-language domain names found a correlation between domain name length and malicious intent, domain names consisting of only numbers were more likely to be considered bad websites, and domain names consisting of only one standard word were not popular websites (Smits, 2020). In the IEEE International Conference on Intelligence and Security Informatics (ISI), it is stated that attackers often use popular domains and top domains in designing websites with malicious intent (Mehedi et al., 2020). The global network could face serious threats with the emergence of sophisticated DGA Bots, DGA Bots have their dictionaries that can combine words to generate dynamic domain names that are not easy to distinguish from man-made domain names (Satoh et al., 2021). In 1999, the US Congress enacted the Anti-Cybersquatting Consumer Protection Act (ACPA) to deal with several cases of cybersquatting or the registration of domain names impersonating other people's trademarks for illicit profit, in which internet companies involved Internet Companies for Assigned Names and Numbers (ICANN) adopted the Uniform Domain Name Dispute Resolution Policy (UDRP) (*Legal Regulation of Internet Domain Names in North America by Jacqueline D. Lipton :: SSRN*, n.d.).

This study aims to determine the most frequently used words from about ten thousand domain names and their suffixes, that were managed, owned, or often accessed in the Indonesian territory. The domain name and its suffix are unique, there is no redundancy, the domain name also consists of one word or several words combined without spaces and ends with a '.' (dot) sign and its suffix. From the collection of domain names, the words that are often used or used as part of a domain name can be analyzed and separate using wordninja, wordninja is probabilistically split concatenated words using NLP based on English Wikipedia uni-gram frequencies. One of the result of this study is to find words that not trivial, usual, regular and ordinary, but eye catching and arouse curiosity to make new website name.

2. RESEARCH METHOD

This research was conducted in several stages, starting from data collection, data preprocessing, word separation from domain names, and the presentation of research results.

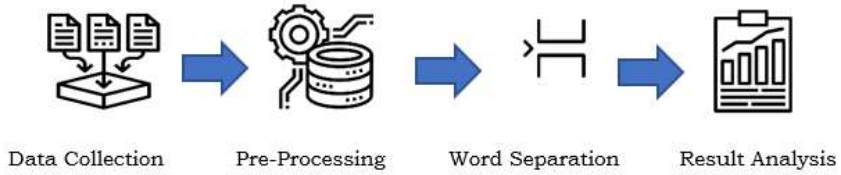

Figure 1. Stages research

The first stage of collecting domain name data was obtained from a DNS (Domain Name System) service provider company in charge of managing domain data information in a global network. Data preprocessing in this study is quite simple, from the raw data obtained, the attributes to be used are selected, namely the domain name and its extension/suffix.

The next stage is the separation of words from each domain name using one of the library packages that can be run in the Python programming language using Google Collaboratory (Kuroki, 2021)(Dacon & Tang, 2021). The most frequently used words in

domain names as a result of Wordninja, will be compared with a search using Microsoft Excel, to measure the accuracy. The last stage in this study is the presentation of the research results in the form of visualization of the most used words as domain names.

2.1 Wordninja

Wordninja is a library that can be run in the Python programming language. Wordninja works probabilistically (looks for possibilities) by splitting off concatenated words using Natural Language Processing (NLP) dependencies (Canesche et al., 2021). NLP is a branch of artificial intelligence that relate to interactions between humans and computers using natural language (Kang et al., 2020) . Wordninja uses the unigram frequency of the Indonesian Wikipedia, for example: "caramengobatipenebalandindingrahim.web.id" to "cara mengobati penebalan dinding rahim web id"(Dacon & Tang, 2021).

2.2 Tools

The tool used in this research was the Python programming language which is run at https://colab.research.google.com/. This software is technically similar to the free Jupyter Notebook but in cloud form and can be run using a browser, such as Mozilla, Firefox, Google Chrome, or others. This tool provides Python code without the need to do installation and setup processes, instead, all setting and adjustment needs are left to the cloud(Tock, 2020).

This application is the right tool for programmers who want to deepen or solidify their knowledge of Python code(Gunawan et al., 2020). Another plus, Google Collaboratory is also known for being able to drive the need for team collaboration. Projects created later can also be edited simultaneously by other team members, just like editing documents in Google Docs (Fan, 2017). The last thing that is most often needed from Google Collaboratory is that this tool has a collection of popular machine learning libraries and can be easily loaded on the project you are working on (Gunawan et al., 2020).

3. RESULTS AND DISCUSSIONS

The data set used in this study is 10,960 domain names and their suffixes. The components of a domain name are as shown in Table 1 below.

Table 1. The components of a domain name (Smits, 2020)

www	**.**	**contoh**	**.**	**com**
Subdomain		Second Level Domain (SLD)		Third Level Domain (TLD)/*Suffix*

As the most frequently used word analysis, this study also takes the SLD along with the TLD/suffix of a domain. The TLD distribution of the data set used can be seen in Table 2. Suffix Distribution and Figure 3 below is diagram of suffix distribution.

Table 2. Suffix Distribution

Suffix	Sum
ac.id	201
biz.id	65
co.id	4607
desa.id	194
go.id	112
id	3383
my.id	118
or.id	551
ponpes.id	9
sch.id	866
web.id	854

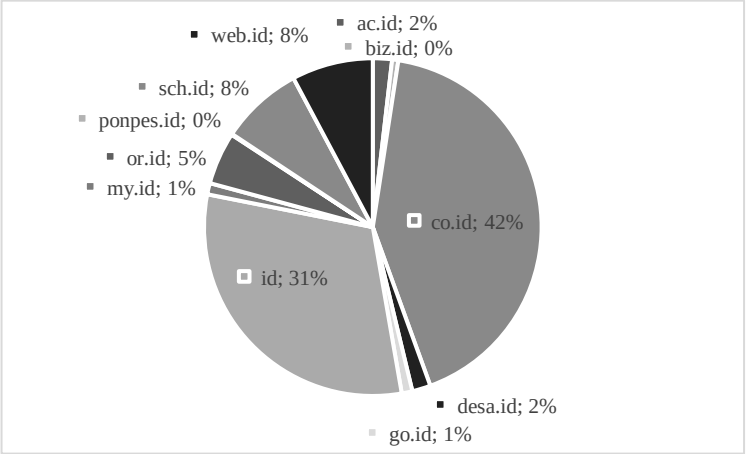

Figure 2. Suffix Distribution Diagram

3.1. Word Separation

The separation of domain names and their suffixes uses a package available in the Python programming language library, namely Wordninja. The separation of domain names and their suffixes is done several times based on the minimum number of letters desired. For example, a word that consists of at least 3 letters, 5 letters, and so on. The results of the separation of words based on the minimum number of letters desired are sorted and only the ten most frequently used words are taken. The following Tables 2. are sorted frequently used TLD and TLD if "id" word is the most frequently used that descript at Figures 3.

Table 3. Frequently Used TLD Parts

Word	Quantity
id	10976
co	4639
web	884
sch	867
or	613
desa	210
ac	206
man	183
go	161
my	159

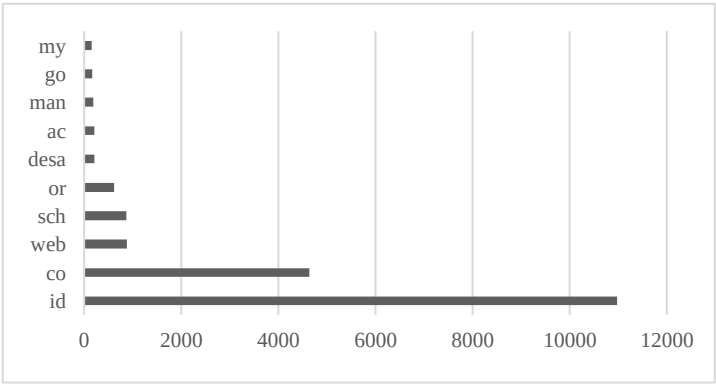

Figure 3. The Most Frequently Used TLD Parts

Table 3 and table 4 are the summary of the results obtained, website contain of word “web” is the most frequently in > 2 letters, word “desa” is the most frequently in > 3 letters, word “indonesia” is the most frequently in > 4 and < 8 letters, and word “collection” is the most frequently in > 9 letters.

Table 4. Summary of 10 Frequently Used Words (at least 2 – 5 letters)

> 2 letters	Sum	> 3 letters	Sum	> 4 letters	Sum	> 5 letters	Sum
web	884	desa	210	indonesia	139	indonesia	139
sch	867	indonesia	139	timur	73	online	48
desa	210	indo	121	group	59	travel	47
man	183	jaya	102	media	58	jakarta	38
indonesia	139	shop	81	store	54	digital	34
indo	121	prom	76	online	48	muslim	29
and	112	lamp	75	travel	47	bandung	27
jaya	102	timur	73	jakarta	38	global	25
ung	101	bali	67	digital	34	mandir	25
per	90	group	59	batik	33	studio	24

Table 5. Summary of 10 Frequently Used Words (at least 6 – 9 letters)

> 6 letters	Sum	> 7 letters	Sum	> 8 letters	Sum	> 9 letters	Sum
indonesia	139	indonesia	139	indonesia	139	collection	21
jakarta	38	collection	21	collection	21	production	11
digital	34	catering	16	nusantara	14	consulting	11
bandung	27	creative	15	consulting	11	advertising	7
fashion	22	makassar	14	production	11	technology	7
collection	21	nusantara	14	logistics	10	foundation	6
tanjung	18	property	13	adventure	9	management	6
catering	16	solution	11	furniture	9	distributor	6
creative	15	consulting	11	paramount	8	darussalam	5
laundry	15	production	11	technology	7	yogyakarta	4

Figure 4 and 5 are describe summary of results, website contain of word “indonesia” is dominate in almost all used word parameter.

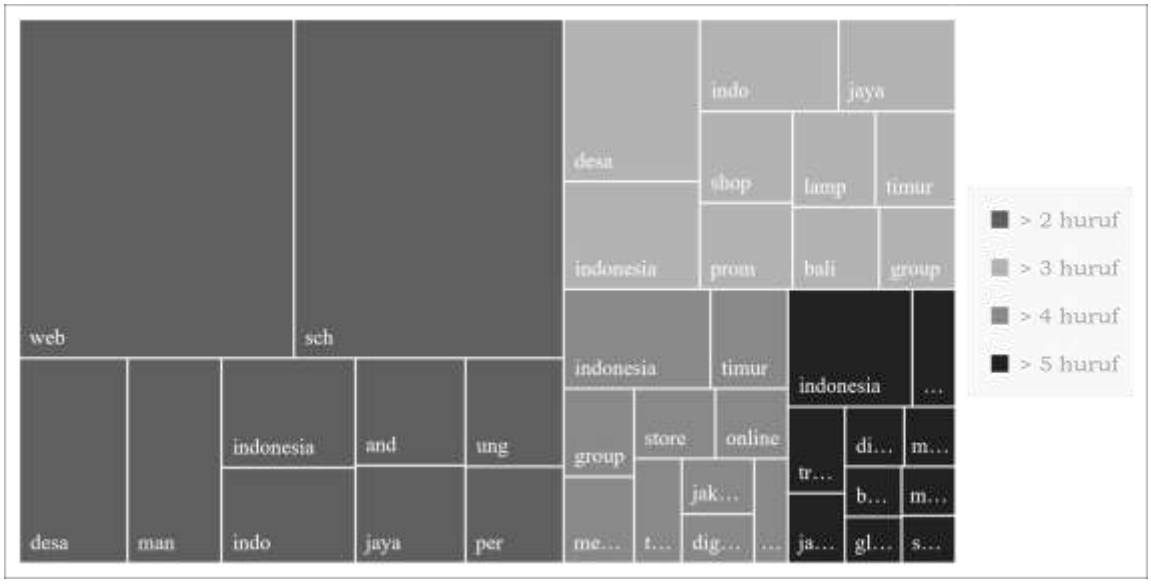

Figure 4. Graphics of frequently used words (minimum 2 – 5 letters)

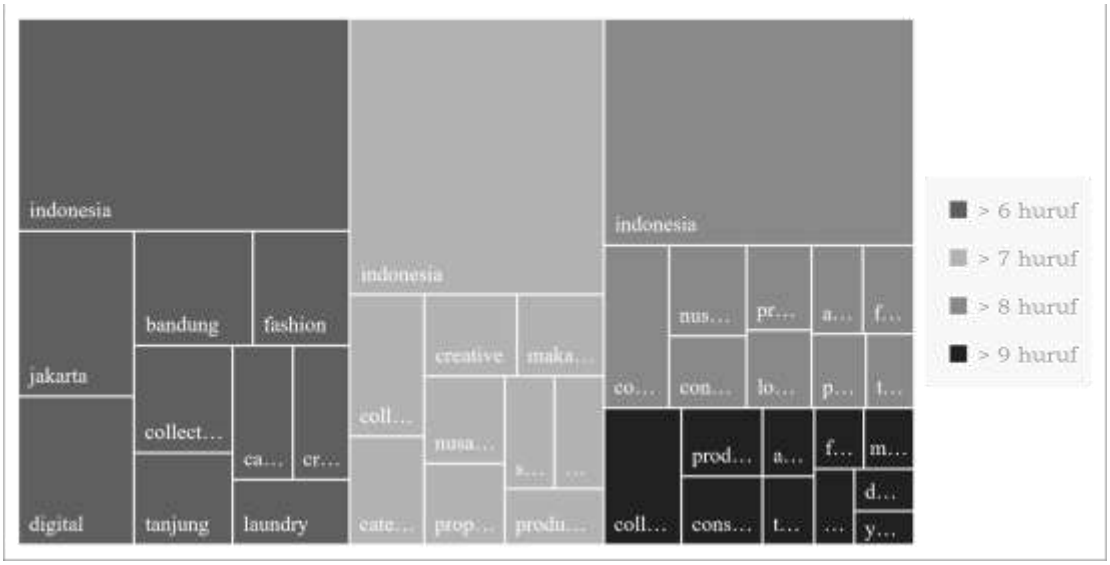

Figure 5. Graphics of frequently used words (minimum 6 – 9 letters)

From the summary above, it can be concluded that the word "indonesia" is most often used for domain naming although there are only 139 (1.27% of 10,960) domain names that use the word "indonesia". The words "web" and "sch" appear 884 (8.07%) and 867 (7.91%) times, respectively, these words can be confirmed as suffixes because they can be seen in Table 1. Another word that often appears is the word " jaya" and "indo" 102 (0.93%) and 121 (1.1%) times.

4. CONCLUSION

Word separation using Wordninja in Python is good enough to be used in Indonesian. By using Wordninja, the word "indonesia" was the word that appears most often in this study, 139 times. However, when the search was done using the find feature in Microsoft Excel, the word "indonesia" was found 143 times. In other words, the accuracy of Wordninja can be considered quite large (97.2%). One of result of this study is to find words that not trivial, not usual, not regular and not ordinary, but eye catching and arouse curiosity to make new website name.

In the future, research can be carried out with a much larger number of domain names, it can also use the hashtag name dataset. Furthermore, from the resulting word separation, it is also possible to identify bad words or sentences that are considered hoaxes. This study has actionable insight on larger scale like in Big Data, the Wordninja package for Python programming is very reliable.

REFERENCES

Canesche, M., Bragança, L., Neto, O. P. V., Nacif, J. A., & Ferreira, R. (2021). Google Colab CAD4U: Hands-on cloud laboratories for digital design. *Proceedings - IEEE International Symposium on Circuits and Systems*, *2021-May*. https://doi.org/10.1109/ISCAS51556.2021.9401151

Dacon, J., & Tang, J. (2021). *What Truly Matters? Using Linguistic Cues for Analyzing the #BlackLivesMatter Movement and its Counter Protests: 2013 to 2020*. https://doi.org/10.5281/zenodo.4056563

Fan, M. (2017). Google Docs as a tool for collaborative writing in the middle school classroom. *Journal of Information Technology Education: Research*, *16*, 391–410. http://www.informingscience.org/Publications/3870

Giannakoulopoulos, A., Pergantis, M., Limniati, L., & Kouretsis, A. (2022). Investigating the Country of Origin and the Role of the .eu TLD in External Trade of European Union Member States. *Future Internet 2022, Vol. 14, Page 174*, *14*(6), 174. https://doi.org/10.3390/FI14060174

Gunawan, T. S., Ashraf, A., Riza, B. S., Haryanto, E. V., Rosnelly, R., Kartiwi, M., & Janin, Z. (2020). Development of video-based emotion recognition using deep learning with Google Colab. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *18*(5), 2463–2471. https://doi.org/10.12928/TELKOMNIKA.V18I5.16717

Kang, Y., Cai, Z., Tan, C. W., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. *Https://Doi.Org/10.1080/23270012.2020.1756939*, *7*(2), 139–172. https://doi.org/10.1080/23270012.2020.1756939

Kim, T. H., & Reeves, D. (2020). A survey of domain name system vulnerabilities and attacks. *Journal of Surveillance, Security and Safety*, *1*(1), 34–60. https://doi.org/10.20517/JSSS.2020.14

Kuroki, M. (2021). Using Python and Google Colab to teach undergraduate microeconomic theory. *International Review of Economics Education*, *38*, 100225. https://doi.org/10.1016/J.IREE.2021.100225

Legal Regulation of Internet Domain Names in North America by Jacqueline D. Lipton :: SSRN. (n.d.). Retrieved June 24, 2022, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3290646

Mehedi, M., Pritom, A., Schweitzer, K. M., Bateman, R. M., Xu, M., & Xu, S. (2020). *Data-Driven Characterization and Detection of COVID-19 Themed Malicious Websites; Data-Driven Characterization and Detection of COVID-19 Themed Malicious Websites*. https://doi.org/10.1109/ISI49825.2020.9280522

Satoh, A., Nakamura, Y., Fukuda, Y., Nobayashi, D., & Ikenaga, T. (2021). An approach for identifying malicious domain names generated by dictionary-based DGA bots. *IEICE Transactions on Information and Systems*, *E104.D*(5), 669–672. https://doi.org/10.1587/TRANSINF.2020NTL0001

SINAGA, D. (2019). Comparative Study of Cohering Suffix in English and Indonesia. *Jurnal Ilmiah Simantek*, *3*(1). https://www.simantek.sciencemakarioz.org/index.php/JIK/article/view/28

Smits, J. (2020). *What does a Domain Name say*.

Susilowati, Y. (2019). *MODUL E-COMMERCE untuk Siswa Kelas XI Teaching Factory*. https://books.google.co.id/books?id=I6LGDwAAQBAJ&pg=PA38&dq=Website+statis+dan+website+dinamis&hl=id&sa=X&ved=0ahUKEwi38LmnnJHpAhVs73MBHVopAXMQ6AEIRDAE#v=onepage&q=Website statis dan website dinamis&f=false

Tock, K. (2020). *Google CoLaboratory as a Platform for Python Coding with Students*. *2*(1), 1–13. https://doi.org/10.32374/rtsre.2019.013.