International Journal of Coronaviruses

A Data Mining Methodology for Detecting Conspiracy Theories from Scientific Articles: The Covid-19 Case

Raúl Isea[1,*] Rafael Mayo-García[2]

[1]IDEA, Hoyo de la Puerta, Baruta, Venezuela, [2]CIEMAT, Av Complutense 22, Madrid, Spain

Abstract

The goal is to do a text mining analysis of all scientific publications and find out what journal and what aspects are studying about the conspiracy theories of Covid-19. For this purpose, all publications available in the National Center for Biotechnology Information (NCBI) database were consulted as they were peer-reviewed papers. Of all these papers, only the abstracts of each one were studied using artificial intelligence techniques to determine, for example, whether the subject is of importance depending on the journals where it has been published, and above all, what possible relationships could be extracted from the information published in them. In addition, the "Net Prevalence per Covid19" index was definedin those countries with a high value, greater campaigns should be sponsored to avoid the misinformation generated by Covid-19, although this comment should be verified in future publications. The main challenge was to unify the abstracts and for this purpose, a text summarizer was used under artificial intelligence schemes. The results obtained indicate the tendency of certain topics by the frequency of the words obtained where the focus on the conspiration are the Covid-19 vaccines, but further work is still needed to continue working on this methodology to unify the results.

Analysis Article

Open Access &

Peer-Reviewed Article

DOI: 10.14302/issn.2692-1537.ijcv-23-4586

Corresponding author:

Raul Isea, IDEA, Hoyo de la Puerta, Baruta, Venezuela

Keywords: Text Mining, Covid-19, Origin, Next Prevalence, NCBI

Received: May 05, 2023

Accepted: May 18, 2023

Published: June 22, 2023

Academic Editor:

John Akighir, Federal University Wukari, Taraba State, Nigeria.

Citation:

Raul Isea, Rafael Mayo-García (2023) A Data Mining Methodology for Detecting Conspiracy Theories from Scientific Articles: The Covid-19 Case. International Journal of Coronaviruses– 4(4):22-31. https://doi.org/10.14302/issn.2692-1537.ijcv-23-4586

Introduction

Thanks to advances in information technologies and how this information is distributed anywhere in the world in a wide range of database repositories focused on the management of publications and, in addition, in social networks [1], the main drawback is the large volume of information and, above all, that in many cases this information has the main problem that it has not been verified [2].

The confinement due to the Covid-19 pandemic has generated a large amount of information to know what was being done, what preventive measures should be taken, or what treatments and measures were carried out to contain the epidemic.

In fact, a range of conspiracy theories has been propagated claiming, for example, that vaccines are an attempt to commercialize health that Covid19 did not exist [3], that it was all a political strategy to control the masses [4], etc.

Other factors could be the failure to adopt measures to contain the spread of the disease, such as the use of face masks, as well as belittling vaccination campaigns motivated solely by conspiracy theories.

In fact, a study published in Sweden reported that, in 2021, one in three people thought that Covid-19 was a product of pharmaceutical campaigns [5], as well as there was a campaign speculating about the high mortality rate in that country compared to others [6].

In view of this, the present work conducts a search of peer-validated scientific literature in order to elucidate the extent of conspiracy theories concerning Covid-19 in the scientific community, i.e., to analytically determine whether it is possible to speak of conspiracy from the realm of science. Such knowledge can help to contain disinformation in future pandemics that may strike humanity or those campaigns orchestrated to seek the confusion of society.

Two epidemics, one virus

The Covid-19 pandemic has recorded since the first incident in December 2019 more than six hundred and ninety million cases worldwide until April 2023, of which six million eight hundred people died, according to data recorded at John Hopkins University (available at coronavirus.jhu.edu).

Since then, quarantine protocols (more or less strict in the different countries) as well as vaccination schedules have been established [7]. In all countries, most of the information came from the media and social networks [8], because the people need to keep informed. Moreover, new terms such as infodemic, which is a product of the union of an "epidemic" with "information", to highlight the excess of information that people could access without verifying the sources of it, being therefore many of them false or erroneous [9].

In fact, the director of the World Health Organization (WHO) pointed out on February 15, 2020, that there should be another front to combat the epidemic caused by Covid-19 focused on the fight against infodemic. This was due to the polarization of the debate and the questioning of the health measures of the various bodies responsible for counteracting the epidemic.

Conspiracy theories

It is well known that conspiracy theories proliferate as a result of disinformation spread mainly through social networks [10-12]. In this regard, there was, for example, the belief that the Covid-19 pandemic was really a campaign to disfavor Donald Trump's administration as president of the United States of America [13].

In fact, there was a wide range of conspiracy theories resulting from the Covid-19 confinement, where there were many sources that pointed, for example, that the coronavirus came from a biological laboratory in China, or that Bill Gates wanted to sponsor a vaccine for humanity, without overlooking the defamation that it was a product of telecommunications resulting from 5G radiation [14], among others.

These conspiracy theories are usually the product of factors such as helplessness [15], anxiety and uncertainty [16], fighting the feeling of control that threatens people's safety [17], anti-vaccination campaigns [18], political factors [19] or the feeling of denial that the Covid-19 pandemic really existed [20].

Methodology

First, a search of all the scientific publications available in the National Center for Biotechnology Information (NCBI) was carried out through the web address https://www.ncbi.nlm.nih.gov/pmc, whose title of the work had the words "Covid-19" and "Conspiracy". It is assumed that at NCBI all papers have been peer-reviewed and the sources of information have been reviewed to avoid replicating misinformation on this topic. This repository has more than 8.3 million scientific articles available, despite the fact that in 2000 it started with only 2 scientific journals such as Proceedings of the National Academy of Science of

the United States of America and Molecular Biology of the Cell (available at https://pubmed.ncbi.nlm.nih.gov/).

Search terms were restricted to paper title and abstract only. The results were analyzed with a range of libraries developed in Python to perform text mining [21]. The Python library used was basically NLTK - Natural Language Toolkit [22], which is specialized for statistical natural language processing.

The next step was to calculate an indicator that we called "Covid-19 Net Prevalence", which is the ratio of the number of Covid-19 infected per country to the total population of that country. This indicator is derived from the concept of prevalence [23] which, as we know, is the proportion of a group or population in a given country that presents a given characteristic or event at a specific time or in a specific period. The statistics on the total number of infected persons consider indistinctly whether that person is the product of reinfection or not.

Data for both the number of infected persons and the total population were obtained from the Johns Hopkins University database. The frequency of words appearing in these papers was then determined after debugging with a text resumer using the NLTK libraries developed in Python.

To summarize the abstracts of the publications, all text was converted into lowercase and each paragraph into sentences. Subsequently, the frequency of words in each sentence of all publications obtained from NCBI was determined. With this data, each original sentence of the work is analyzed and those with the highest statistical value obtained from the frequency of occurrence are selected. Thus, the result will be those sentences with the highest frequency in the words that were obtained with the highest frequency from all the publications [more details at https://stackabuse.com/text-summarization-with-nltk-in-python/].

As a result of this methodology, the count of the most frequent words in the abstracts and also those sentences that appear more frequently than others is derived.

Results

The first step was to find out how many scientific publications are available in the NCBI whose title contains the words "Covid-19" and "Conspiracy". That search yielded 501 publications as of April 1, 2023, which is distributed as follows: 104 publications published in 2020, 189 in 2021, 216 in 2022, and 69 in 2023. From all this, it can be seen that there was an increase in the number of publications in 2021 of just over 55%, while in 2022 the increase was 87.5% with respect to the previous year.

Table 1 shows the countries that have published 10 or more scientific papers. The countries with the lowest number of papers are Bangladesh, France, India, Indonesia, Malaysia, and Singapore with 9 publications each, Belgium and Spain with 7, Iran, Oman, Peru, and South Africa with 7, Colombia, Mexico, and Turkey with 6, Argentina, Czech Republic, Finland, Niger and Saudi Arabia with 5, etc.

An examination of the types of publications shows that 470 are Scientific Articles, 20 are Reviews, 13 are Editorial, 10 are Comments, and 8 are Letters. In other words, there is a significant effort made by the scientific community to explain the role of Conspiracy theories in Covid-19.

Table 1. Number of scientific publications available in NCBI (see text for details).

Country	Number of publications
United States	155
United Kingdom	76
China	37
Canada	29
Italy	28
Australia	24
Germany	23
Korea	17
Poland	16
Pakistan	15
Sweden	15
Switzarland	15
Netherlands	14
Austria	11

Table 2. Journals that have 10 or more articles with the condition Origin Covid-19.

Journal	Number of Publications	Scholar Google h5	Journal nationality
Int J Environ Res Public Health	32	152	Nigeria
J Med Internet Res	22	-	Germany
Front Psychol	21	-	Switzerland
PLoSOne	16	198	USA
PersIndividDif	16	-	England
Vaccines	15	-	Netherlands
HealthCommun	14	-	USA
JMIR Infodemiology	10	-	Canada
BMJ	9	190	UK

NOTE: The hyphen means that the journal has no h5 value from Google Scholar.

In the Table 2 shows the h5 value used by Google Scholar indicates the impact of the articles published in that journal in the last 5 years without interruption (in our study, this corresponds to the period between 2017-2021). It highlights the fact that of the nine publications indicated in Table 3, only 3 of them have an h5 index in Google Scholar which indicates that this topic is not usually published in high impact journals.

The next step was to determine the "Net Prevalence by Covid19" (defined in this manuscript) parameter from the data obtained from Johns Hopkins University as shown in Table 3.

Table 3. Net Prevalence by Covid19 according to the first countries with the highest number of confirmed infections according to the data obtained from Johns Hopkins University and integration of the data in Table 1.

Country	Covid19 cases	Population	Covid-19 Net Prevalence	Number of publications
France	39.850.030	65.584.518	60,76%	-
Greece	5.965.643	10.316.637	57,83%	-
Switzarland	4.399.088	8.773.637	50,14%	15
Netherlands	8.610.372	17.211.447	50,13%	14
Germany	38.368.891	83.883.596	45,74%	23
Australia	11.352.930	26.068.792	43,55%	24
Italy	27.715.384	60.262.770	42,67%	28
UnitedKingdom	24.448.729	68.497.907	35,69%	76
UnitedStates	106.363.949	334.805.269	31,77%	155
Spain	13.798.747	46.719.142	29,54%	-
Turkey	17.232.066	85.561.976	20,14%	-
Poland	6.504.194	37.739.785	17,23%	16
Pakistan	1.580.153		17	15
Canada	4.634.277	38.388.419	12,07%	29
Venezuela	552.483	29.266.991	1,89%	
China	503.302	-	<1%	37

NOTE: “-” means data not indicated

As it can be seen in Table 3, the first countries that have a high Net Prevalence value for Covid-19 have a non-negligible number of cases and published works that disprove the Covid-19 conspiracy theories; therefore, it is interesting to carry out further studies to infer that those countries with a high prevalence rate are those countries where more communication campaigns should be sponsored to avoid the propagation of conspiracy theories.

The next step was to determine the number of words in each Abstract. To do this, the longest abstract had 765 words, and to find out if this is a trend in all the papers, we proceeded to determine the average number of words per article, which were 196 words (exactly 195.87 average words per article). The distribution of words per article is shown in Figure 1.

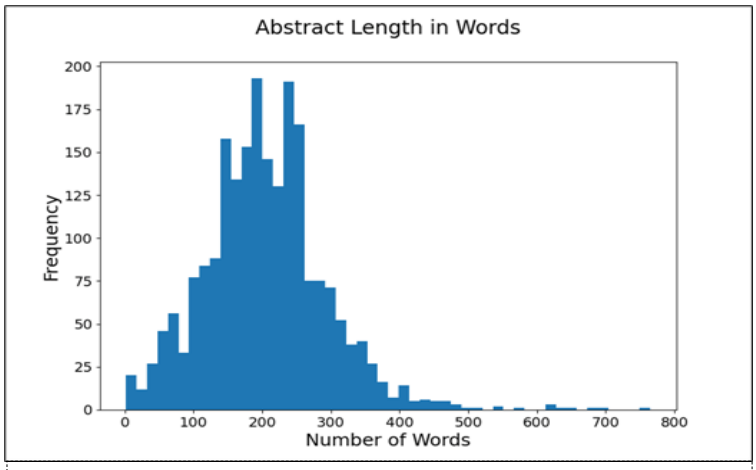

Fig.1. The distribution of words per article

This distribution is not uniform, and to achieve this, we proceeded to calculate the unique number of words in each Abstract. The average word count was 117 words per article (117.29) and the distribution is shown in Figure 2. It should be noticed that 116 publications presented an abstract with only one word.

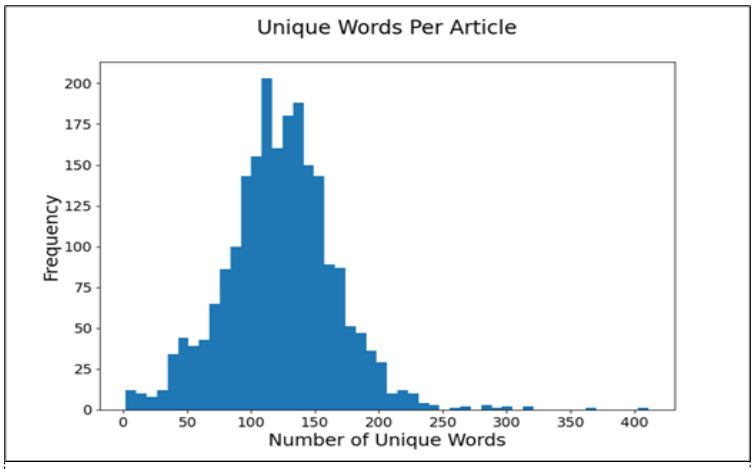

Fig.2. The distribution of only words per article (see text for details)

An important point is that there is a series of words that are frequent in the language and do not contribute anything to the analysis being carried out in the work, known as stopwords, such as "is", "that", "by", "so much", "then", among many others. These have been eliminated according to the methodology described in [24]. By eliminating all of them, and counting the most repeated words we obtain that Covid19 is repeated 6106 and SARS-CoV-2 was used 2888 times.

This is a list of the frequency of the 30 most repeated words in the abstracts, it means, 'covid19' (6106) ir., covid19 appears 6106 in all papers, ‘sarscov2’ (2888), 'pandemic' (2026), 'health' (1671), 'origin' (1623), 'study' (1535), 'vaccine' (1513), 'patients' (1456), 'coronavirus' (1454), 'disease' (1366), 'virus' (1093), 'results' (1061), 'conspiracy' (1013), 'infection' (965), 'data' (938), 'public' (888), 'vaccination' (840), 'severe' (815), 'information' (809), 'respiratory' (806), 'analysis' (791), 'may' (759), 'cases' (728), 'social' (714), 'associated' (698), 'among' (684), and 'risk' (657).

The next step is to corroborate that all the Abstracts deal with Conspiracy Theories and COVID-19, and for this purpose, we find the keywords which said the general topics covered in the papers, buy with help of resumer obtained from abstract. The most frequent keywords obtained after applying the resumer are (from most frequent to least frequent) conspiracy (304 times), Covid (192), Belief and Beliefs (192), Theory(ies) (128), Study (110), Vaccine (93), Health (89), Pandemic (86), Vaccination(s) (92), Social (60), Public (45), Trust (38) and Perceived (34). According to this list, it can be seen that most of the papers deal with truth in conspiracy theories where Vaccine and Vaccination(s) have a high occurrence. Figure 3 shows a word cloud with the words that appear most frequently in the cloud where a virus has been used as a template for the figure.

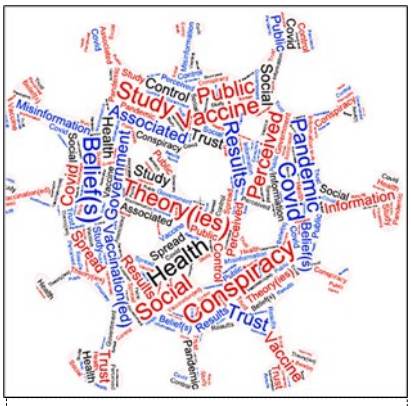

Fig.3 The words that appear most frequently in the abstract

Finally, as well as the words, we proceeded to identify phrases that occur more frequently in the texts and that can serve as an orientation of the topics being studied in the works, where phrases such as (the frequency of these phrases was indicated in parentheses):

“this study aimed to evaluate covid 19 vaccine acceptance among” (0,20)

“this study explored the associations between covid 19 conspiracy beliefs” (0,20)

“conspiracy theories during the covid 19 pandemic” (0,29)

“feelings of anxiety and lack of control” (0,29)

“belief in conspiracy theories and the” (0,33)

“belief in covid 19 conspiracy theories” (0,33)

“belief in covid related conspiracy theories” (0,33)

“beliefs conspiracy theories about covid 19” (0,33)

“conspiracy beliefs about covid 19 and” (0,33)

“conspiracy theories related to covid 19” (0,33)

“and covid 19 conspiracy beliefs” (0,40)

“as well as conspiracy mentality” (0,40)

“association between conspiracy mentality and” (0,40)

“belief in the conspiracy theory” (0,40)

“beliefs in conspiracy theories and” (0,40)

"conspiracy beliefs are associated with" (0,40)

"vaccine acceptance results showed that" (0,40)

"and vaccine conspiracy beliefs" (0,50)

"believing in conspiracy theories" (0,50

"compliance with public health" (0,67)

"conspiracy theories was positively" (0,50)

"conspiratorial thinking and the" (0,50)

"covid related conspiracy beliefs" (0,50)

"study examined the association between conspiracy beliefs" (0,29)

"infodemics conspiracy beliefs and religious fatalism" (0,33)

"less support for public health policies" (0,40)

"web interest in conspiracy hypotheses and" (0,40)

"pollution and climate change" (0,50)

"an infodemic of" (0,67)

"anxiety and depression" (0,90)

"and conspiracy mentality" (0,67)

From all the above sentences, it is easy to observe that the studies are focused on the association of the occurrence of Covid-19 cases by a belief in a conspiracy associated with vaccines, where the relationship of the conspiracy with religion could be inferred, but it is a very punctual aspect.

The last (and future work) is to calculate the similarity using the cosine likelihood test, which is based on the cosine mathematical function that seeks to measure how similar texts are once they are reduced to a vector (details in [25]), as a result of which a word is represented as a vector so that the text documents are represented in an n-dimensional vector space.

Conclusions

The paper performs a search of all publications that have focused on conspiracy theories centered on COVID-19 available in the NCBI database. It is seen that there is a significant correlation between the Covid-19 Net Prevalence rate with the number of cases and publications in some countries, so it should be analyzed country by country to determine the reasons for this. In fact, it was surprising that China and Russia do not publish about conspiracy theories, while the United States and the United Kingdom are the most prominent in this regard. On the other hand, the keywords as well as the phrases indicate that conspiracy theories are focused on vaccines. Also, when examining what kind of articles have been written on the subject, the vast majority are scientific articles, not reviews or letters, so the subject cannot be taken lightly and should serve as a warning to face future pandemics that may strike mankind.

Finally, we are beginning to work on a cluster concept that involves the similarity between texts to group them. However, it is not a simple calculation to analyze due to the number of variables to be used and we are still developing a new cluster concept that will take into account the degree of similarity between papers.

References

1. Qin Ye and Chen Rongrong. Social Network Analysis of COVID-19 Research and the Changing International Collaboration Structure. J. Shanghai Jiao Tong Univ. (2022). https://doi.org/10.1007/s12204-022-2558-7.
2. Cheng Cheng and Rita Espanha. Critical Review: A Review of the Studies About the Usage of Social Media During the Covid-19 Pandemic. Comunicação e Sociedade, (2021), 40: 149-167. https://doi.org/10.17231/comsoc.40(2021).3174
3. Clare Dyer. Surgeon who said covid-19 was a hoax loses appeal to have suspension overturned (2023) BMJ 2023: 381.
4. HT Logemann and S. Tomczyk. How Media Reports on COVID-19 Conspiracy Theories Impact Consensus Beliefs and Protective Action: A Randomized Controlled Online Trial. Sci Commun. 2022 Dec 14:10755470221143087. doi: 10.1177/10755470221143087
5. Elena Raffetti, Elena Mondino & Giuliano Di Baldassarre. Epidemic risk perceptions in Italy and Sweden driven by authority responses to COVID-19. Scientific Reports volume 12, Article number: 9291 (2022)
6. Mariam Claeson, Stefan Hanson. COVID-19 and the Swedish enigma. Lancet. Vol 397 (10271),pp: P259-261, JANUARY 23, 2021
7. John E Oeltmann, Divya Vohra, Holly H Matulewicz, Nickolas DeLuca, Jonathan P Smith, Chandra Couzens, R Ryan Lash, Barrington Harvey, Melissa Boyette, Alicia Edwards, Philip M Talboy, Odessa Dubose, Paul Regan, Penny Loosier, Elise Caruso, Dolores J Katz, Melanie M Taylor, Patrick K Moonan, Isolation and Quarantine for Coronavirus Disease 2019 (COVID-19) in the United States, 2020–2022, Clinical Infectious Diseases, 2023
8. Panagiotis D. Michailidis. Visualizing Social Media Research in the Age of COVID-19. Information 2022, 13(8), 372
9. Carrion-Alvarez D, Tijerina-Salina PX. Fake news in COVID-19: A perspective. Health Promot Perspect. 2020 Nov 7;10(4):290-291. doi: 10.34172/hpp.2020.44. PMID: 33312921; PMCID: PMC7722992.
10. Uscinski, JE; Enders, AM; Klofstad, C.; Seelig, M.; Funchon, J.; et al Why do people believe COVID-19 conspiracy theories? 2020. Available at: https://misinforeview.hks.harvard.edu/article/why-do-people-believe-covid-19-conspiracy-theories/ (April 1, 2023).
11. Allington, D.; Duffy, B.; Wessely, S.; Dhavan, N.; Rubin, J. Health-protective behaviour, social media usage and conspiracy belief during the COVID-19 public health emergency.. psicol. Medicina. 2021 , 51 , 1763–1769.
12. Oliver, E. y Wood, T. (2014). Medical Conspiracy Theories and Health Behaviors in the United States. JAMA, 174 (5), 817-818. doi:10.1001/jamainternmed.2014.190
13. By Peter Baker and Annie Karni. The New York Times, 2020. Disponible en línea https://www.nytimes.com/2020/02/28/us/politics/trump-accuses-media-democrats-coronavirus.html (Consultado el 1 de abril de 2023)

14. Flaherty E, Sturm T, Farries E. The conspiracy of Covid-19 and 5G: Spatial analysis fallacies in the age of data democratization. Soc Sci Med. 2022 Jan;293:114546. doi: 10.1016/j.socscimed.2021.114546. Epub 2021 Nov 6. PMID: 34954674
15. Abalakina-Paap, M.; Stephan, WG; Craig, T.; Gregory, WL Beliefs in conspiracies. Political Psychol.. 1999 , 20 , 637–647
16. Van Prooijen, J.-W. An Existential Threat Model of Conspiracy Theories. . EUR. phicol. 2020 , 25 , 16–25
17. Douglas, KM T COVID-19 conspiracy theories.. Intergroup Relat.l 2021 , 24 , 270–275.
18. Jolley, D.; Douglas, KM The Effects of Anti-Vaccine Conspiracy Theories on Vaccination Intentions. PLoS ONE 2014 , 9 , e89177
19. Motta, M.; Stecula, D.; Farhart, C. How Right-Leaning Media Coverage of COVID-19 Facilitated the Spread of Misinformation in the Early Stages of the Pandemic in the US Can. J. Political Sci. 2020 , 53 , 335–342.
20. Uscinski, JE; Enders, AM; Klofstad, C.; Seelig, M.; Funchon, J.; Everett, C.; Wuchty, S.; Premaratne, K.; Murthi, M. Why do people believe COVID-19 conspiracy theories? 2020. Disponible en línea: https://misinforeview.hks.harvard.edu/article/why-do-people-believe-covid-19-conspiracy-theories/ (consultado el 1 de abril de 2023).
21. S. Bird, E. Klein, and E. Loper, Natural Language Processing with Python. Sebastopol, CA: O'Reilly Media, 2009. ISBN-10: 0596516495, ISBN-13: 978-0596516499
22. Benjamin Bengfort, Rebecca Bilbro, Tony Ojeda. Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning. 1ra edición. O'Reilly Media; 1er edición (17 Julio 2018). ISBN-13 : 978-1491963043
23. I Spronk I, JC Korevaar, R. Poos, R. Davids, H. Hilderink, FG Schellevis, RA Verheij and MMJ Nielen. Calculating incidence rates and prevalence proportions: not as simple as it seems. BMC Public Health. 2019 May 6;19(1):512. doi: 10.1186/s12889-019-6820-3
24. Los detalles para eliminar las palabras communes (stopwords) está disponible en https://www.geeksforgeeks.org/removing-stop-words-nltk-python/
25. D Gunawan, C A Sembiring, M A Budiman. The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents. IOP Conf. Series: Journal of Physics: Conf. Series 978 (2018) 012120 doi :10.1088/1742-6596/978/1/012120