INSTITUT TEKNOLOGI BANDUNG 1920
KKIK - FSRD

Jurnal Sosioteknologi

Website: https://journals.itb.ac.id/index.php/sostek/index

Word Frequencies in Linguistic Articles Published in SINTA Indexed Journals

Frekuensi Kata dalam Artikel Linguistik yang Diterbitkan di Jurnal Terindeks SINTA

Heri Heryono, Dadang Suganda, Susi Yuliawati, Nani Darmayanti

Doctoral Program of Linguistics, Universitas Padjadjaran, Bandung

heri20002@mail.unpad.ac.id

ARTICLE INFO

Keywords:
word frequency, data corpus, academic research

ABSTRACT

Multiword sequences are a language pattern that occurs when a bunch of words emerge in a similar register. In research papers conducted by lecturers and students, different topic areas and indexes has created various characteristics of lexical bundles. The method of this research is qualitative, combining corpus design to identify the sequence of words within the text. The corpus data were generated from five different indexing journals, yet the topic is linguistics. Initially, the whole papers were converted to text format to deal with readability in the program used. The program used was Orange Apps version 3.27 by applying the textable, data table, and text mining menus. The sources of the data are emphasized as being academic research indexed in SINTA 5, published in 2020. The main theory of used in this research is that of Biber's (2007) which discusses the main characteristics and number of criteria for defining word strings. This observation resulted in 207.896 characters and 33.636 words. There were 4,273 words based on the pre-processing analysis result, which included transformation, tokenization, and PoS-tagging. From a total of 4,273 words, virus, deixis, and slang were the most frequently occurring. Based on these results, it can be concluded that the majority of journal articles are about viruses and slang. They pertain to the prevalent topic of pandemics at the time the journals were published. When the process of writing a journal article is in progress, this information may aid the authors in identifying the journal's keywords and most frequent words.

INFO ARTIKEL

Kata kunci:
frekuensi kata, data korpus, penulisan akademis

ABSTRAK

Pola bahasa yang sering diamati adalah rangkaian kata ganda. Istilah ini mengacu pada frekuensi kata dari sekelompok kata yang muncul dalam register yang sama. Dalam jurnal penelitian yang diltulis oleh dosen dan mahasiswa, area topik dan indeks yang berbeda telah menciptakan berbagai karakteristik bundel leksikal. Metode penelitian ini menggunakan metode kualitatif dengan menggabungkan desain korpus untuk mengidentifikasi urutan kata dalam teks. Data korpus dihasilkan dari lima jurnal pengindeks yang berbeda, dengan topik linguistik. Awalnya, seluruh artikel diubah menjadi format teks agar terbaca dalam program yang digunakan. Program yang digunakan adalah Orange Apps versi 3.27 dengan menerapkan menu textable, data table*, dan* text mining*. Sumber data penelitian akademik diambil dari jurnal* *yang terindeks SINTA 5, yang diterbitkan pada tahun 2020. Artikel yang digunakan memiliki karakteristik yang beragam, baik dalam gaya penulisan maupun diksi. Teori utama penelitian ini menggunakan Biber (2007) yang membahas karakteristik utama dan sejumlah kriteria untuk mendefinisikan string kata. Pengamatan ini menghasilkan 207.896 karakter dan 33.636 kata. Terdapat 4.273 kata berdasarkan hasil analisis prapemrosesan yang meliputi transformasi, tokenisasi dan penandaan PoS. Dari total 4.273 kata tersebut, kata virus, deiksis, dan slang merupakan kata yang paling sering muncul. Berdasarkan hasil termuan tersebut dapat disimpulkan mayoritas artikel jurnal adalah tentang virus dan bahasa gaul. Kedua kata tersebut berkaitan dengan topik pandemi yang sedang marak pada saat jurnal-jurnal tersebut diterbitkan. Ketika proses penulisan artikel jurnal sedang berlangsung, informasi ini dapat membantu penulis dalam mengidentifikasi kata kunci jurnal dan kata-kata yang paling sering muncul.*

https://doi.org/10.5614/sostek.itbj.2023.22.1.8

Introduction

Language patterns have become essential to be researched by linguists in the last few decades. One of the language patterns that are quite common to observe is a series of words. This string of words refers to words that coincidentally and mutually follow each other more often by meeting certain criteria. There are several terms that are used to refer to a series of this word. For ideas and knowledge that should be effectively communicated to the reader, standard words, phrases, and expression language are required. In other words, the ability to write academic texts is not only based on the grammatical point of view but also should be based on the level of lexicon and syntax to arrange formulaic language, which is fundamentally basic in written academic research.

There is a set of expression sequences for the basic elements in that discourse. Along with the development of technology. It creates a greater requirement for fast and accurate information acquisition, which leads to the focus of this research. The advanced development should meet the information requirements as described. Language becomes the medium for obtaining, converting, and conveying information in the form of data. The computer stands as the mediator (machine) between data availability and human requirements; therefore, the machine did the approaching to comprehend natural language. It refers to a language understandable by someone in a certain location; for example, the natural language of Indonesian is Indonesian (Bahasa).

Theoretically, natural language processing is the development of various computing methods to analyze and display text in natural language on one or more levels of linguistic analysis to achieve goals in terms of language that include completing various tasks or applications (Liddy, 2001). The process of information retrieval leads to the process of withdrawing the required information within a certain desired period. This process depends on numerous factors for retrieving accurate data in an efficient time. Factors can be systems that are used to retrieve data, how it is stored, or how the schema is designed. In the era of the information explosion, improved and efficient techniques are required to collect smart data in addition to related data. Natural Language Processing (NLP) considers the manner of interaction between humans and machines, which processes to learn various rules and then apply the rules for the same task and for making intelligent data (Devale & Deshpande, 2011). The closest system related to linguistics in terms of classifying, parsing, identifying, and analyzing data texts is data mining, which later will be more specifically defined as “text mining”.

Data mining refers to a process of finding information where a user interacts with a set of documents that are full of texts. Data mining is also defined as a field of research that focuses on searching for or defining patterns in data. In terms of information necessity. Data mining stands as a science that scour large datasets for extracting implied information that is previously unknown and potentially useful (Gorunescu, 2011). Data mining is a term related to the use of algorithms and computers to find interesting patterns inside data (Dua & Xian Du, 2011). Data mining has two main achievements: predictions and descriptions. Prediction refers to the ability of studying the variables in the data set to predict the future unknown value of other variables of interest. In the description, it concerns finding arrangements that determine data so it can be interpreted by humans (Kantardzic, 2011). On the other hand, most documents are not classified or even well organized, so the related documents are complicated to find and analyze, as in academic journals, social media (Twitter), or even in an online newspaper. The goal of data mining depends on the amount of data and the knowledge and creativity possessed by the user or researcher when formulating it. In fact, data mining may be similar to solving a puzzle. The choice of method is determining in solving data mining problems since the data have their own characteristics (Athohillah et al., 2015).

Text mining is a method of mining useful information from data in the form of written texts, documents, or text in the form of classification and clustering (Han and Kamber, 2006). The data source does not emerge from set documents, and interesting patterns are not found in database records but in unstructured text data (Feldman, 2007). Finding the most relevant and exact information from various sources has become the most essential matter in text mining (Azam &Yao, 2012). This also leads to a research development in the field of sentiment analysis and opinion mining, which later become the intersection subject between language (linguistics) and informatics. By providing a system that can be automatically analyzed, users may review and extract information from the most relevant results (Brody, 2010).

Text mining is a part of data mining where the process of the data, text, as well as documents appears on an exceptionally large scale. To conduct large-scale data processing, it will consume a lot of resources that deal with data processing. At this point, pre-processing of text as data is required before the data are generated as a data set, which is a necessity in linguistics. Text mining has comparable properties to data mining, yet it focuses on text rather than a more structured form of data. Moreover, one of the initial steps in the text mining process is to organize and structure the data in such a way that it can be subjected to both qualitative and quantitative analysis. It specifically involves natural language processing (NLP) technology, which applies principles of computational linguistics to describe and interpret data sets. In the linguistics area, text mining relates to corpora, which have a large dataset. A set or group of texts may be categorized as a corpus when it is used as the object of language and literary research (Kilgarriff & Grefenstette, 2003). With the assistance of NLP technology and text mining, the process of analyzing texts can be much more efficient than manual processes. The use of NLP and text mining can help improve the efficiency of text document analysis by providing automation capabilities for the process.

Literature Review

Basically, the three core categories in lexical bundles are: 1) research-oriented bundles that help writers to structure their activities and experiences of the real world; 2) text-oriented bundles, which are concerned with the organization of the text and its meaning as a message or argument; and 3) participant-oriented bundles that focus on the writer or reader of the text. Taxonomy relates to the meaning and purpose of language. The function is to try to organize the discourse according to the situation and context (Hyland, 2008). There are a number of criteria for determining the sequence of words that can be referred to as lexical bundles. First, the series of words must occur together and repeatedly in the corpus of research data. Second, a series of three to four words that occur together at least ten times per million words in the research corpus cannot be referred to as lexical bundles because the three words series that occur frequently less than ten times cannot be referred to as lexical bundles.

The minimum limit of words in a string is five to six words, and they should at least appear five times per million words. At the very least, its use is limited. However, a string of two words can also be categorized as a lexical cluster if the word sequence is the abbreviation of the series of three words. The string of words should at least also be found in five different texts on the corpus-type level data. It aims to exclude lexical groups with a special language style used by the writer (Biber, 1999). Previous studies focused on the comparison of frequency and lexical functional category bundles in the translated text that was generated by student translators and by professional translators. Junior or student translators are those who are at the second level of the university and have completed a translation course, whereas a professional translator is a certified translator who is a member of the Translators Association of Indonesia (HPI). This research concludes that the production frequency and category of functional lexical bundles generated depend on the level of English as a second language (Novita & Kwary, 2018).

In addition, previous research discussed equivalent terms used to refer to lexical bundles in Indonesian as a lexical group. The term lexical bundles will be renamed "lexical group" in the next few paragraphs. The research focuses on finding the frequency and analyzing the structure of lexical groups in Indonesian when writing academics in law. The data corpus used consists of theses, dissertations, and journal articles containing 2,054.312 words. Meanwhile, the limit for the minimum set frequency is 40 times in at least five different texts. As a result, it was found that 475 lexical groups consisting of three words up to seven words were dominated by three words. Frequent use of vocabulary bundles can be used as a signal of competent use of the language in a particular register. More competent students use more natural, sometimes unintended, specific bundles to transmit specific discourse features in their arguments (Heng et al., 2014). On the other hand, according to other previous studies, students as writers rarely use vocabulary bundles in their academic papers, and language abuse arises from a lack of awareness of choosing correct, more natural expression patterns.

Additionally, some of them abused them inappropriately or did not match the usage of vocabulary bundles used by experienced authors, even when they used specific vocabulary bundles (Cortes, 2004). The area of linguistics was chosen because professional authors writing in the field of linguistics are highly aware of aspects or features of language, including fixed expressions expressed using vocabulary bundles. It is based on the assumption that the authors are experts in using language-related studies in a variety of fields. Linguistics is the use of knowledge about the properties of language obtained from linguistic research to improve the efficiency of practical tasks in which language is a central component (Corder, 1974). There are various aspects that should be noticed when composing good writing, especially for research papers, such as word choices, grammar, and punctuation. In this case, beside word choices and punctuation, word structure becomes an important aspect of writing to deliver ideas to be understood by readers (Agustina & Junining, 2015).

Method

In this research, every part of the research was computerized. It started with data retrieval using web scraping techniques and text analysis using Orange. The coverage methodology for this research is the qualitative method by providing and serving datasets obtained from papers published in 2020. The research method is shown in Figure 1 below:

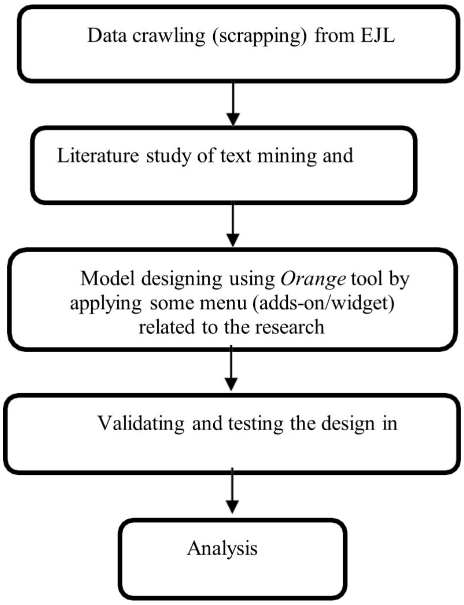

Data crawling (scrapping) from EJL
Literature study of text mining and
Model designing using *Orange* tool by applying some menu (adds-on/widget) related to the research
Validating and testing the design in
Analysis

Figure 1 Research method flow chart

Based on the flow chart above, the initial stage or step in this research was conducting a literature review and study of several papers or journals related to text mining and pre-processing, especially for linguistics purposes. The second step was starting to crawl or compile data sources from papers published in 2020; this resulted in thousands of characters, which were later segmented into thousands of words. The third step—words segmented from characters—was processed by the Orange data mining tool to obtain the required analysis. The fourth step was validating and testing the design that had been conducted in Orange, and the final stage was analyzing and evaluating the model.

The scope or data source for this research is taken from English Journal Literacy Utama, vol. 5, 2020. The issue has eight papers, which contain various linguistic scopes of discussion. The source is a peer-reviewed journal published in 2020, when the pandemic was progressing. English Journal Literacy regularly issues two publications a-year, in March and September. The data source is in the form of words that are generated and converted from character segmentations. Those datasets are processed in machine learning to obtain the results of word frequency and prediction through a word cloud. The technique is called text mining, yet in this research, the text mining process applies Orange data mining tools, which are simple and compact tools to be used. Text mining requires several primary stages, which should be prepared so the text can be changed to be more structured. Some of the main stages in text mining are pre-processing of text (text pre-processing), transforming text (text transformation), feature selection, and pattern discovery in text or data mining (Even, 2002).

This data crawler process is based on a program or script that crawls web pages; in this research, the data source was papers on the OJS (Open Journal System), targeted sequentially and automatically. This term is also known as spider-ing. The search process is based on the latest data available on the internet (OJS). On this occasion, the volume taken was the latest volume in 2020 to obtain and retrieve recent information. Almost all machines' searchers are now using the concept of crawlers to gather information from the internet as a major component of the search engine (Mironeanu, 2017). The web crawler becomes essential for collecting the latest data or information quickly. Data or information that has been collected by using a web crawler can be used for various needs, such as developing a business strategy to increase product sales.

Building massive-scale corpora for linguistics analysis in digital form can be considered the results of a collection of references collected and arranged in handwritten form over decades. In addition, the current term "corpus", which is most often used to refer to a linguistic data set collected for specific analytical purposes, assumes that it will be saved, managed, and analyzed in digital form (Sasongko, 2010). In this term, the papers in a certain volume were created to be corpora; the corpora were then available for some requirements of the linguistics research. In building a corpus from various languages, it can be scraped from a representative sample from various tests by providing the corpus with as accurate a description of the inclination as possible, including proportions between the corpus and information perspective. In conclusion, it is not solely based on the selected sample text, but also looking for samples from multiple sources taken from the original document, so that it will give you an idea that it is quite accurate of all the information to be obtained. Thus, it may be inferred that data crawling in this research refers to a stage intended to collect or download data from a database, which will be applied in the analysis as a model and main data source. Collecting data from this research is done using data downloaded from the OJS server in the form of texts in the papers and their attributes (George et al., 2014).

The text mining process is basically similar to data mining, except for some methods and the fact that managed data is like text data: unstructured, partially structured, and structured like email text, text HTML, as well as text comments from various sources (Vijayarani, 2015).

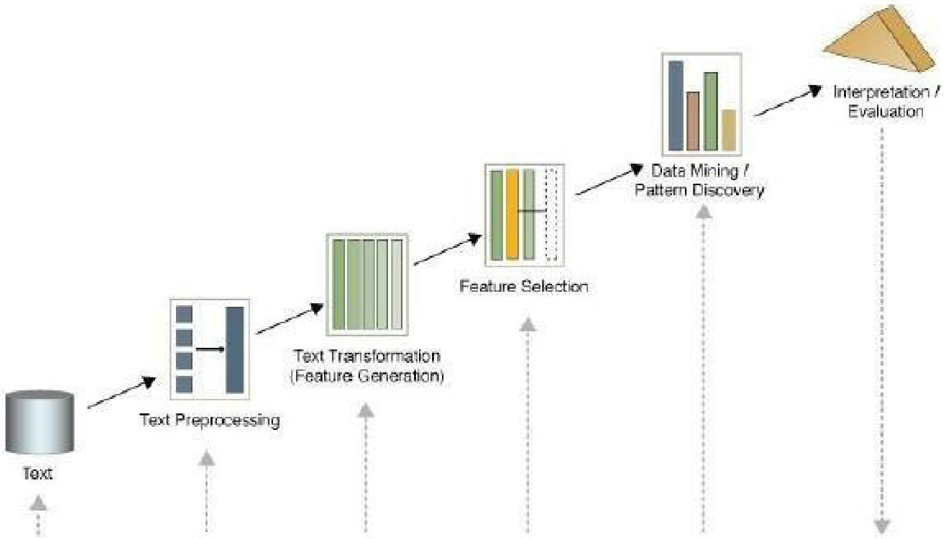


Figure 2 Text mining process based on Even.

Based on the irregularity of the text data structure, the process of information retrieval or text mining systems requires several essential stages, and one of the implementations of text mining is the text pre-processing stage. It also refers to a technique used to dig up hidden data, especially in texts. One of the methods in text mining is clustering. It is a grouping technique that is widely used in data mining. The main purpose of the clustering method is to group a number of data or objects into a cluster (group) so that each cluster will contain data as closely as possible. In addition, the overuse of features in the classification process may cause increased time-counting and decreased accuracy (Uysal & Gunal, 2012). Text mining is also referred to as a recent technology used for company data, which is always increasing, so that the unstructured text can be analyzed (Francis & Flynn, 2010). The pattern discovery stage is the most essential stage of the whole text mining process. It leads to the discovery of the patterns in or knowledge of the entire text.

Pre-processing text

The text pre-processing is a stage of selecting the data to be processed in each document. Commonly, in very basic steps or elements, this process will include case folding, tokenizing, filtering, and stemming. The pre-processing process is conducted to clean the data from noise; it has a smaller dimension, and furthermore, the data are more structured, so they can be processed further (Langgeni & Firdaus, 2010). The first stage of text pre-processing is case folding. It refers to a process in text to homogenize characters in the data. The case-folding process is also a process of changing all letters to lowercase. The data characters "A" through "Z" are converted into the characters "a" and "z" during this process. Characters other than the letters 'a' to 'z' (punctuation marks and numbers) will be removed from the data and considered delimiters (Raghavan and Schutze, 2009). Text mining searches for information from data sources through the identification and exploration of certain patterns; in this case, the data source is a collection of documents with patterns found in unstructured text. The pre-processing of text mining is focused on feature identification and extraction from representative natural language documents (Pratama et al., 2018). The flow chart below shows simple pre-processing stages that involve several stages to obtain raw data (a data set). The text mining process requires the arrangement of the input text based on grammar, which is followed by exploring patterns from structured data, evaluating, and interpreting the results. This process is used for classifying, grouping, meaning analysis, drawing conclusions from documents, and object relationship modeling in the form of words.

Text processing is the initial stage of the text process, which is to prepare the text as data that will be processed further. It is the process of breaking down the text process to prepare the text as data that will be processed further. This process breaks down text or even characters into sentences, words, or tokens.

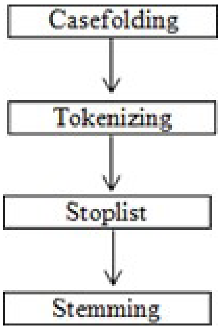


Figure 3 Basic pre-processing stages

The next step is tokenizing or parsing, which is the stage of cutting the input string based on each word that composes it, while the filtering stage is the stage of taking important words from the term results. This stage also eliminates certain characters, such as punctuation marks, and filters by text length (Krishna and Bhavani, 2010). For testing or evaluation methods tested on the models researched for information, a proposed model was used.

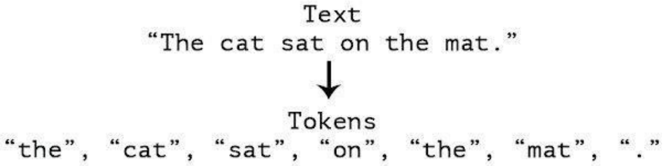

Text
"The cat sat on the mat."
Tokens
"the", "cat", "sat", "on", "the", "mat", "."

Figure 4 Tokenizing stages

In the context of coding, an algorithm that may be used is stop-list (removing less important words) or word-list (saving important words). Stop-words are non-descriptive words that can be discarded in the bag-of-words approach. Examples of stop-words are "that or which", "and", "at", "from", and so on. In a very dense document, there are many types of words such as conjunctions, prepositions, pronouns, and even interjections. The majority of those words have no potential for identifying the content of a document. Moreover, based on that definition of documents referring to word frequency, it becomes less effective even though the types of words above are not processed through filtering (Srividhya & Anitha, 2010).

Filtering can also be defined as a process to remove important words from the token process, which is also known as stop-words removal. It refers to a vocabulary that is not a feature (unique word) of a document (Dragut et al., 2009). After the words contained within documents pass the tokenizing process and stop list, the next stage that should be conducted is a stemming process. The stemming process aims to change or return the word to its original or basic form by omitting affixes to words in documents. The stemming stage is the stage of finding the root of a word from each filtered word. At this stage, the process is carried out by combining various word formations into the same representation. The stemming process is carried out by checking the word to see whether it contains affixes or not, especially suffixes. The stemming process in English has its own characteristics that cannot be separated from the influence of the grammatical point of view (Porter, 2000).

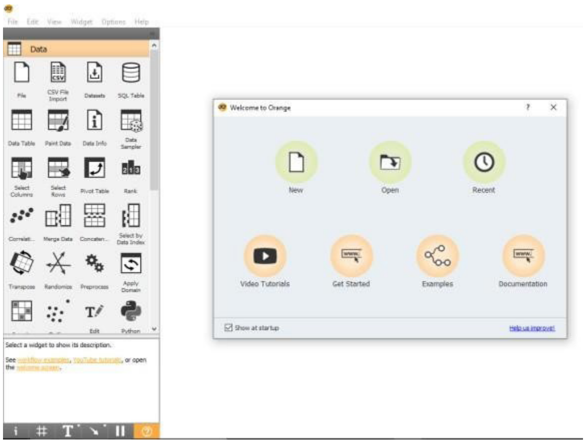
Filtering and classification are also used to recognize the patterns in which the data must be categorized by checking previous data that had been grouped by certain terms. For example, a credit card company or telecommunications company worries about losing customers. Classification helps to identify potential customers to stop that, so that it can provide an overview to help managers predict such customers. Then the manager can make an offer specifically to retain customers.

Orange is an application developed by the University of Ljubljana with the target of presenting an application for processing data visually without the need for experience in making previous programs. Dealing with data mining, the user will engage with a widget system. Every widget has its own function, and it can accept input or output. Orange is created by adapting a library of C++ core objects, which involves a wide variety of standard and non-standard machine learning techniques. It also has algorithms for data mining; in addition, it is also equipped with regular operations for data input and manipulation (Demsar and Zupan, 2004). Orange gives a graphical user interface to some particular information about data mining and how it relates to AI strategies. They include widgets for data entry and pre-processing, data visualization, classification, regression, association rules, and clustering. Furthermore, it gives a set of widgets for model evaluation and visualization of evaluation results, as well as widgets for exporting the models into PMML. Orange allows files to be readable in native tab-delimited format, and it can load data from any of the major standard spreadsheet file types, such as CSV and Excel. The native format starts with a header row with feature (column) names. When the users intend to analyze texts using Orange, the widgets that relate to the analysis requirement should be added through the menu "adds-on." Compared to other data mining software, Orange has a distinct advantage, especially when it comes to visualization or visual programming. Orange provides many of the widgets that are appropriate for users with a non-informatics background. It provides the canvas or drawing board and then connects it with other widgets that basically refer to the simplification of the coding process. With this "canvas", it will be easier for users to play with data and perform data analytics intuitively (Andri, 2015).

An additional stage within Orange was filtering; it removed or saved selected words, including stop-words (e.g., removing 'and', 'or', 'in' and etc.). By using the filtering process, language options could be used for many languages, and English was set as the default. Regexp removes words that match regular expressions \. |, |: |; |! | \? | \ (| \) | \
|| \ + | '| "|' | '|" | "|' | \ '| ... | \ - | - | - \ $ | & | \ * |> | <| \ / | and by default, it was set to remove punctuation marks.

Results and Discussions

This research started with downloading and installing Orange data mining to get the design of analysis texts from the data source (journal). When it was installed, the initial user interface was shown as in Figure 5 below:



Select a widget to show its description.

See workflow examples, YouTube tutorials, or open the welcome screen.

Figure 5 *Screen display of Orange*

As it could be seen, the welcome screen of *Orange* displayed some options, and it might have been started by the "New" file. The menu was on the left side, or it was called widgets. They interacted by means of tokens, which were conveyed from sender to receiver. The widget helps users communicate with the data, gives some operations, and commands to the data, and conducts coding by only connecting the required components to get the result of models. Besides, *Orange* also provides several add-ons or modules for problems in certain domains, such as text mining (text analytics), bioinformatics, network data, social networks, model maps, prototype processes, and others. The most useful model for linguists is sentiment analysis; users might conduct the research for SA from several sources, for example, *Twitter* or the *New York Times.*

1. Inputting Data Source Stage

The use of *Orange* data mining displays the *Textable Widget Design* that is presented in the process flow of Text Clustering, which contains the corpus, corpus viewer, data table, pre-process text, and word cloud. As it is shown below:

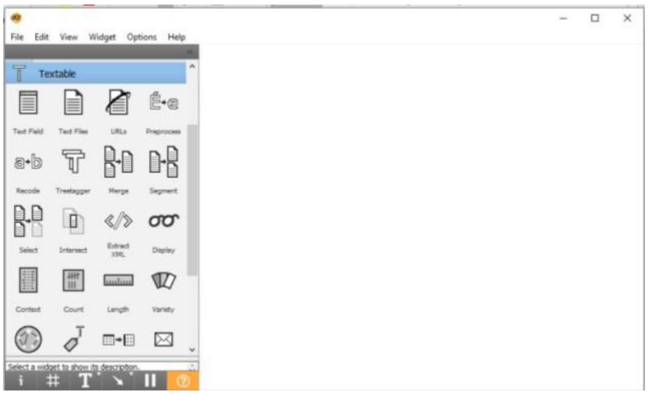
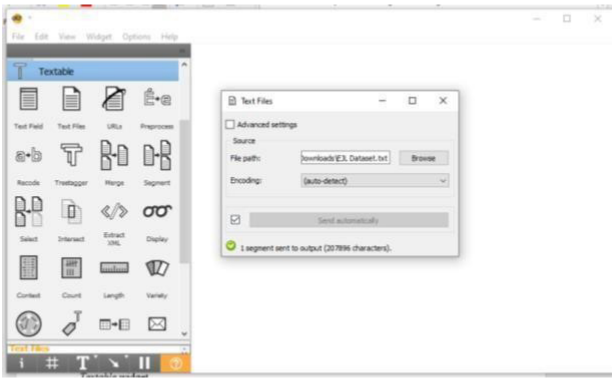


Figure 6 Textable widget

The first step before choosing the sub-widget on the menu above, the source of the data should be provided. In this research, the data source was obtained from the OJS website, which provided archives for whole issues and volumes. This research determined the volume of the journal to limit the analysis. The data sources consisted of eight papers with different authors and topics, commonly about linguistics. The whole papers from volume were converted by conducting copy-paste activities to Notepad, and later, those papers were converted to the .txt format.



Text Files

□ Advanced settings

Source

File path: ownloads\E3. Dataset.txt

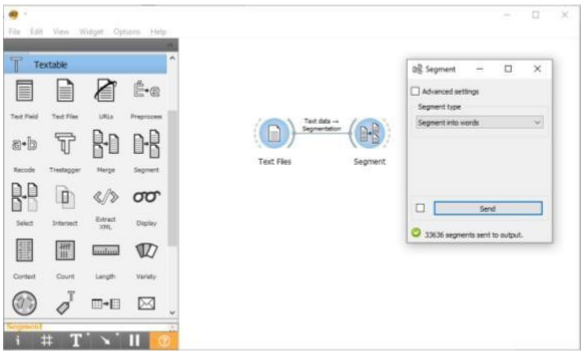
Encoding: (auto-detect)

☑

1 segment sent to output (207896 characters).

Figure 7 Calling the data source

The image above shows that the converted data source was called using the text file option. It automatically reads the total number of characters in the data source that was converted to .txt format. Based on the process, it was reported that “1 segment was sent to output 207896 characters”. It referred to the total number of characters in a very raw mode; it contained letter, number, punctuation, and other elements. The next stage was applying segments and choosing segment type into words. In this section, the characters from the first stage would be converted to word segments. When it was done, the result was 33.636 words, as can be seen in the image below:



Segment

☐ Advanced settings

Segment type: Segment into words

33636 segments sent to output.

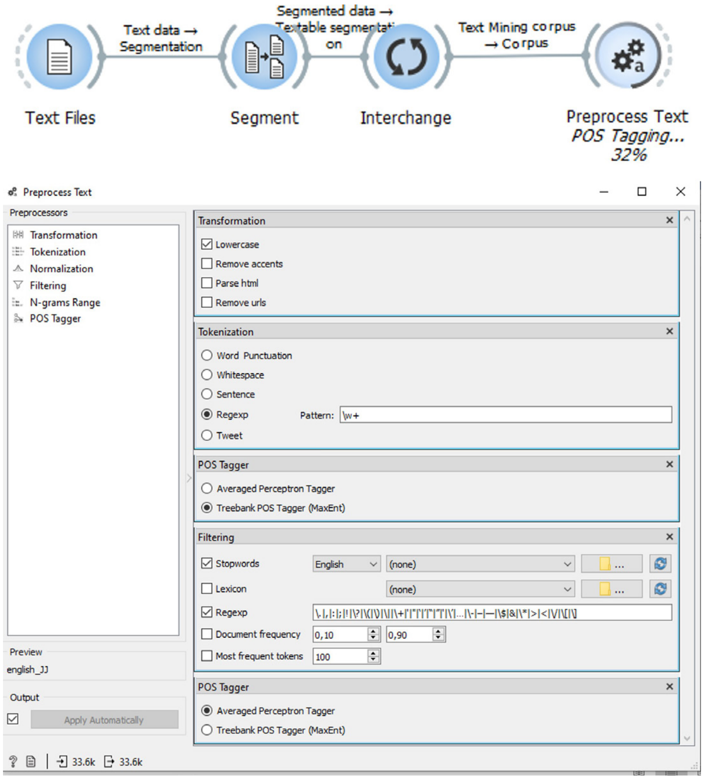
Figure 8 Word segmentation

Those were the stages in the first section of obtaining and inputting a data source to Orange, especially in the particular widget, which in this case was textable. Later on, the pre-processing stage was conducted by applying the other menu (widget) to run the program and obtain the result.

1. Pre-processing Stage

In this stage, the converted data source might start to be processed for several purposes and requirements. The pre-processing stage was the most essential step to obtain various outputs: a word cloud, a bag of words, word frequency, sentiment, and many other outputs. In this case, since the data source was paper, it started to be processed for several purposes and requirements. The pre-processing stage was the most essential step to obtain various outputs: a word cloud, a bag of words, frequency, sentiment, and many other outputs. In this case, since the data source was papers taken from a journal, the output was limited to a word cloud and word frequency to find information about the journal in general.

The process of obtaining a word cloud should be initiated by attaching a mediator, Interchange, that becomes the connector between segment and pre-process. By using interchange, the stage of pre-processing would run perfectly.



Preview: english_JJ

Output: ☑ Apply Automatically

Transformation
☑ Lowercase
☐ Remove accents
☐ Parse html
☐ Remove urls

Tokenization
○ Word Punctuation
○ Whitespace
○ Sentence
● Regexp Pattern: \w+
○ Tweet

POS Tagger
○ Averaged Perceptron Tagger
● Treebank POS Tagger (MaxEnt)

Filtering
☑ Stopwords English (none)
☐ Lexicon (none)
☑ Regexp \.|,|:|;|!|\?|\(|\)|\||\+|'|"|'|"|'|"|'|\'|...|\-|–|—|\$|&|*|>|<|\/|\[|\]
☐ Document frequency 0,10 0,90
☐ Most frequent tokens 100

POS Tagger
● Averaged Perceptron Tagger
○ Treebank POS Tagger (MaxEnt)

Figure 9 Pre-processing

At this point, the data source started to be processed by several elements as described before: transforming, case folding, tokenizing, filtering, and PoS-tagging. All those steps were conducted automatically. It resulted in 4.273 words cleaned from whole words before filtering out unnecessary elements (stop words, punctuation, etc.), and the main words appeared in the word cloud. Filtering for display is the process of combining words that have the same frequency. This form also gives a view of the total number of words that have been processed and the number of words that have been sorted.

2. Running the program

The last stage in this research was running the pre-processing stage to produce the required output, which in this research was a word cloud and word frequency..

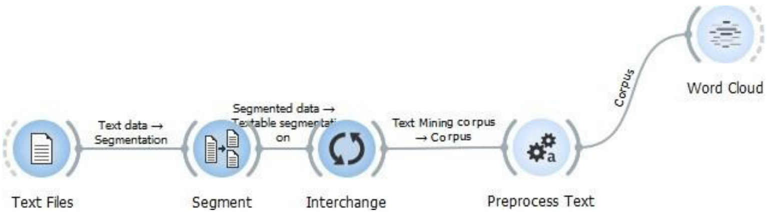


Figure 10 Word Cloud stage

From the image above, the program is running well without any warning signs (!). After setting the pre-processing stage, add the widget word cloud to obtain the output is required.

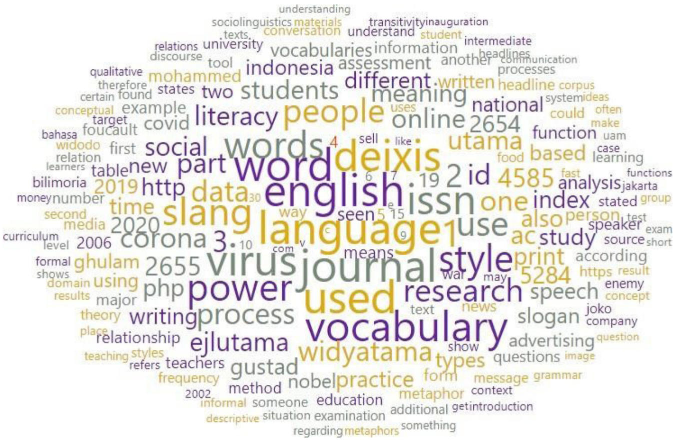


Figure 11 Word Cloud result

From the image above, it can be seen that several dominant words emerged because of the frequency of their appearance. The more frequently a word appears, the larger the letters in the word cloud, especially in the journal published in 2020. From that image, the reader might conclude that the journal commonly discusses English, language, deixis, viruses, slang, etc.

In the display result, it was a word cloud with a frequency of the number of words. It was a variation for displaying the results of the stages pre-processed text. The color of the word made its appearance more attractive and easier to understand. On the other hand, readers might also determine the topics from the journal in a particular issue. It could be references for the readers to obtain detailed information through the whole issue or volume of the journal. Furthermore, the authors could also predict that the next issue would discuss certain topics based on the topic's keywords provided by the Orange output, especially in topic modeling.

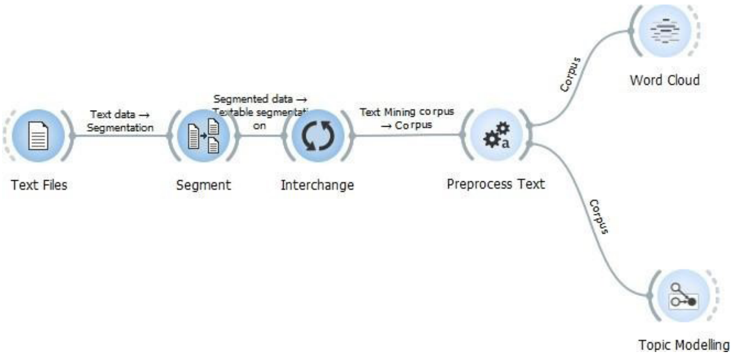


Figure 12 Topic modeling

The topics were basically obtained from the connection line between pre-process texts, which are considered corpora. Basically, there was a corpus viewer widget intended to view text (example corpus). Yet, in this research, the corpus was a data set obtained from converting papers (content) to .txt form. It would return an instance of the corpus, and after the analysis process was completed, the corpus viewer was applied to display the corpus (data review) by providing information in the form of how many documents were there and revealing the explanation. Based on the result, it could be seen that the top five words that frequently appeared in the journal, especially in its 2020 publication, were: language (245 times), English (152 times), journal (151 times), and virus (135 times).

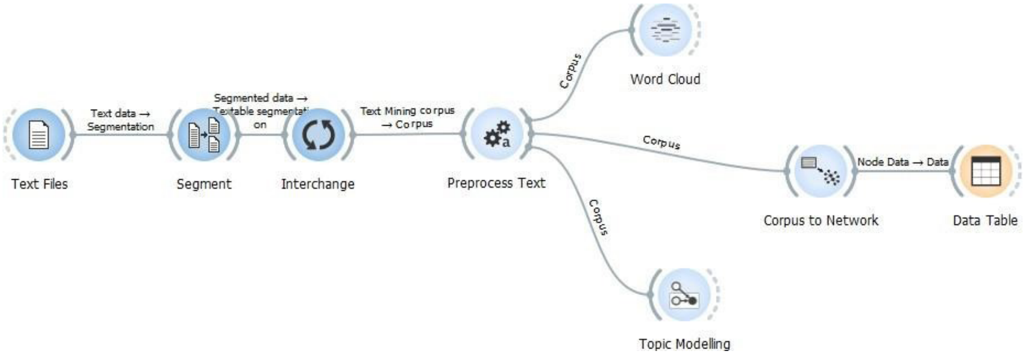


Figure 13 Final stage of process

The strings that appeared in the image above were the final model strings that would result in certain outputs: a word cloud, topic modeling, and word frequency. By generating the pre-process text widget, it opened all the access to various outputs. An automatic search process of core sentences was developed to produce a product text that contained an important part of the original article, although the grammar was not good.

Conclusion

Based on the model analysis conducted in Orange, it could be concluded that the journal had 207.896 characters and 33.636 words. Along with the pre-processing stage, which consisted of transformation, tokenization, and PoS-tagging, 4.273 cleaned words were produced, which are commonly used as the data set for several purposes. Through the process, the data source was converted to a word cloud output that also described the prediction of the whole topic of the journal upon its publication in 2020. Based on the model running, the most frequent words that emerged were: language (245 times), English (152 times), used and journal (151 times), and virus (135 times). In addition, the words slang, metaphors, and corona also emerged in minor totals. It might be concluded that papers published in 2020 mostly discussed linguistics and language under the sub-topic "Corona virus," since the journal volume 5 was issued during the pandemic situation and the discussion about deixis, metaphor, and slang language, which were related to the recent situation, was being hyped.

References

Alamsyah, Andri. (2015). More than words: Social networks "Data Analytics Menggunakan Orange." https://andrya.staf.telkomuniversity.ac.id/data- analytics- menggunakan-orange

Athoillah, M., Irawan, M., & Imah, M. (2015). Study Comparison OF SVM-, K-NN- and Backpropagation-Based Classifier for Image Retrieval. J. Comput. Sci. Inf., vol. 8, no. 1, pp. 11– 19.

Azam, N., & Yao, J. (2012). Comparison of term frequency and document frequency-based feature selection metrics in text categorization. Expert Systems with Applications, 39(5), 4760–4768. https://doi.org/10.1016/j.eswa.2011.09.160

Brody, S. (2010). An Unsupervised Aspect Sentiment Model for Online Reviews, (June), 804–812.

Demsar J, Zupan B (2004) Orange: From Experimental Machine Learning to Interactive Data Mining, White Paper (www.ailab.si/orange), Faculty of Computer and Information Science, University of Ljubljana.

Devale, P., & Deshpande, A. (2011). Probabilistic Context Free Grammar: An Approach to Generic Interactive Natural Language Interface to Database. Journal of Information, Knowledge and Research in Computer Engineering, vol. 01, no. 02.

Dragut, E., Fang, F., Sistla, P., Yu, S. & Meng, W. (2009). Stop Word and Related Problems in Web Interface Integration. http://www.vldb.org/pvldb/2/vldb09- 384.pdf.

Dua, S. & Xian Du. (2011). Data Mining and Machine Learning in Cybersecurity. USA: CRG Press.

Feldman, R. & Sanger, J. (2007). The Text Mining Handbook. New York: Cambridge University Press

Francis, L., & Flynn, M. (2010). Text Mining Handbook. Casualty Actuarial Society.

George, S Antonia, & G. Dimitros. (2014). A Faceted Crawler for the Twitter Service. WISE 2014, Oc 12-14, 2014, Thessaloniki, Greece.

Gorunescu. (2011). Data Mining, vol. 12. Berlin, Heidelberg: Springer Berlin Heidelberg. EForum. Spring.

Han & M. Kamber. (2006). Data mining: concepts and techniques. San Francisco: Morgan Kaufman Publisher.

Kantardzic, M. (2011). Data Mining: Concepts, Models, Methods, & Algorithms. Wiley Online Library: IEEE Press.

Kilgarriff, A. & Grefenstette, G. (2003). Introduction to the Special Issue on the Web as Corpus. Computational Linguistics Volume 29, Number 3.

Krishna & S. Bhavani. (2010). An efficient approach for text clustering based on frequent item-sets. Eur. J. Sci. …, vol. 42, no. 3, pp. 385–396.

Langgeni, Baizal & Firdaus. (2010). Clustering Artikel Berita Berbahasa Indonesia Menggunakan Unsupervised Feature Selection. Seminar Nasional Informatika, Yogyakarta.

Liddy, E.D. (2001). Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Edition, Marcel Decker Inc, NY, USA.

Mironeanu Catalin, A.C. (2017). An efficient method in pre-processing phase of mining suspicious web crawlers. International Conference on Intelligent Computing and Control Systems (ICICCS), 2017.

Porter, A. L., Schoeneck, D. (2000). Mining Electronic R&D Information in Support of Resource Management. 8 th International Symposium on Society and Resource Management Bellingham, WA.

Pratama, J. A., Sunengsih, N., & Suherman, M. (2018). Analisis Klaster Pada Dokumen Teks Opini Pengguna Twitter Terhadap Kasus Miras Oplosan Menggunakan Metode K-Means. Jurnal Statistika Universitas Muhammadiyah Semarang, 6(1).

Raghavan & Schutze. (2009). Introduction to Information Retrieval. New York: Cambridge University Press.

Sasongko, Jati. (2010). Aplikasi untuk Membangun Corpus dari Data Hasil Crawling dengan Berbagai Format Data Secara Otomatis. Jurnal Teknologi Informasi DINAMIK Volume XV, No.1, Januari 2010 : 16-26

Srividhya & R. Anitha. (2010). Evaluating pre-processing techniques in text categorization. Int. J. Computer Sci. Appl., no. 2010, pp. 49–51.

Uysal, A. K., & Gunal, S. (2012). A novel probabilistic feature selection method for text classification. Knowledge-Based Systems, 36, 226–235. https://doi.org/10.1016/j.knosys.2012.06.0 05

Vijayarani, M. J. Ilamathi, & M. Nithya. (2015). Pre-processing Techniques for Text Mining - An Overview, vol. 5, no. 1, pp. 7–16.