ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

BAHASA INDONESIA TEXT CORPUS GENERATION USING WEB CORPORA APPROACHES

[1]AMALIA AMALIA, [2]OPIM SALIM SITOMPUL, [3]ERNA BUDHIARTI NABABAN, [4]MAYA SILVI LYDIA, [5]NADIA RAHMATUNNISA

[1,4,5]Departement of Computer Science, Faculty of Computer Science and Information Technology,

Universitas Sumatera Utara, Medan, Indonesia

[2,3]Departement of Information Technology, Faculty of Computer Science and Information Technology,

Universitas Sumatera Utara, Medan, Indonesia

E-mail: [1]amalia@usu.ac.id

ABSTRACT

A text corpus is a collection of texts stored electronically for various research and investigation for Natural Language Processing (NLP) needs. Bahasa Indonesia is the Indonesian language that is used officially by the Indonesian people, amounting to around 230 million people. However, the development of the NLP of Bahasa Indonesia is not as fast as English. One of the factors because Bahasa Indonesia still has limited linguistic resources like corpus. This study aims to generate Bahasa Indonesia text corpus using a web corpora approach for general purpose in NLP. We collected texts from the seven largest Indonesian online news sites with various categories covered in 52 URLs. The research stages contain resource observation, web structure analyzing, website crawling, scraping, and data cleaning. The last step, the clean data, which is a collection of sentences then arranged into a machine-readable format. In this study, the percentage of successful crawling content from the resources is 85.85% or 569.456 news articles, with 219.392 distinct tokens. It can be concluded that web corpora approaches can produce text corpus for Bahasa Indonesia.

Keywords: *Natural Language Processing (NLP), Corpus, Bahasa Indonesia, Web Corpora, Scrapy*

1. INTRODUCTION

Human language or natural language has evolved since thousands of years ago and used as a media to carry information in communication. Meanwhile, NLP is an important technology that is a sub of Artificial Intelligent field that aims to extract the meaning of natural language automatically. In processing natural language, we need some machine-readable linguistic resources or tools like corpus. The development of NLP Bahasa Indonesia is considered to be hampered due to a lack of linguistic resources. The problems are, the research of NLP Bahasa Indonesia still not having a well developed continuous long term plan[1]. Each researcher in this fieldwork independently and generate their linguistic resources for their research purpose only. Therefore, the availability of linguistic resources in Bahasa Indonesia still not well managed. A corpus is one of the linguistic tools that have many functions in NLP.

There are several types of a corpus, such as general corpus and specific corpus. The general corpus consists of a large number of sentences on various topics; meanwhile, a specific corpus is a corpus that contains specific topics like politics or economics. Some corpus is made for a specific function like for Name Entity Recognition (NER) [2] corpus or Part of Speech (POS) corpus. The sentences in this kind of corpus, not only contains natural language sentences but also contains inserted additional information like NER dan POS information. Another type of corpus is a parallel corpus that contains sentences in two or more different languages in one corpus. One example of a corpus generation for many languages has been done by the PAN project (http://www.panl10n.net). This project produces several corpora for various world languages, including for Bahasa Indonesia corpus. Most corpus of this project is a kind of parallel corpus. Universitas Indonesia has also created a Bahasa Indonesia corpus. This corpus is one of the specific function corpora to identify the POS tagger function in Bahasa Indonesia [3]. Another project in generating Bahasa Indonesia corpus already done by Sketch Engine www.sketchengine.eu this project called IndonesiaWaC [4], but this corpus is not a free corpus we have to pay if we want to use the corpus. The availability of the English corpus is relatively sufficient so that the researchers of NLP of English can focus on their research and not necessary to build an English corpus as the earliest step. Different from the researchers of Bahasa Indonesia, although several corpus Bahasa Indonesia is already available, many researchers still build their corpus for their research need. Various reasons like available corpus sometimes not meet the researchers' need, and the lack of information about availability corpus make the researchers have to build their corpus. The research question is how to generate an appropriate corpus that is used as NLP resources. Semantically, the corpus meaning is a collection of texts. Still, in modern linguistics, this understanding has expanded into a collection of texts that have four main headings sampling and representativeness, finite size, machine-readable form, and a standard reference (5). Building a corpus is one of the essential and earliest steps in NLP research. It is essential because the selection of corpus will impact to the NLP research result. One of the challenges in designing Bahasa Indonesia corpus is that there are still not many data sources to support a wide range of genres, such as spoken, fiction, magazines, newspaper, academics, etc. Therefore, a method is needed to build a corpus by the conditions of limited linguistic language resources like Bahasa Indonesia. There are several methods in generating a corpus for NLP purposes. In the beginning, before the internet era, a corpus is built manually by collecting many documents text. This traditional method is very time consuming and also labor-consuming.

One of the methods in generating a corpus is using web corpora approaches. The corpus is built from mining textual data from the web. This method is popularized by WaCky [5] and Leipzig Corpora Collection (LCC), a project by the University of Leipzig [6]. Although LCC is available in various language, LCC only give a sentence-wise shuffled translation from some specific words and not the entire corpus.

Some advantages to generate a corpus from a web corpus are, the corpus can be built without cost, the corpus size can reach a large scale, and the sentences can generate more natural. Other advantages, a web corpus approaches corpus can produce new words because of this method's coverage texts from natural language. One of the challenges of this method, we have to extract the information needed from HTML documents. It took an in-depth observation to filter and get only natural sentences from websites resources and discarded the noises like HTML tags, ASCII symbols, and redundant sentences. According to this background, the purpose of this study is to generate Bahasa Indonesia text corpus using web corpora approaches. The contribution of this study is a method that can be applied to generate a general-corpus where the source of the text is collected from websites. Besides Bahasa Indonesia, this method can also be applied to build a corpus for other languages. Another contribution of this study is the generated Bahasa Indonesia text corpus that can be used by other researchers for further NLP task research. The corpus is already in a machine-readable format and can be downloaded at www.risetahasa.id. The remainder of this paper is organized as follows: In section 2, we describe the theoretical background followed by the explanation of the related work in Section 3. Section 4 is the methodology followed by Section 4: the result and discussion section. The last section is Section 5 is described as a conclusion from this paper.

2. THEORETICAL BACKGROUND

2.1 Corpus

A term “corpus” is used to describe a collection of documents, whether written or spoken, that are stored and processed on a computer for investigation and research. A text corpus as one essential linguistic resource can be used for NLP applications like part-of-speech tagger, language parsers, speech recognizers, search engine optimization, literature studies, and other special needs in NLP.

Some several criteria and qualities must be fulfilled by a corpus [7], namely:

a. The text must be collected according to specific criteria, such as by content or list.
b. Text is available in machine-readable form.
c. Texts are collected to analyze certain linguistic phenomena.

A Web corpus approaches in generating an NLP corpus is a method to build a corpus from mining data on the websites. Websites as the most extensive available text resources in various languages can be utilized to generate an NLP corpus. Some advantages of these methods are:

1. The corpus can be built free of charge, excluding costs for storage space, CPU power, and bandwidth.
2. The resulting corpus size can reach a large scale.
3. The web corpora can reach the new word because it is available exclusively and includes texts that are much closer to natural language than the traditionally compiled corpus.
4. This method only requires standard tools and databases that are usually available locally.

Construction of the web corpora makes it possible to take random samples from a population of web documents. Some parameters that must be considered in building an NLP corpus are the corpus representativeness, corpus balanced, corpus sampling, corpus size, and corpus format. A representative corpus is a corpus that representing a particular language or language variety where the corpus contains a large number of texts that cover all the varieties of a language; meanwhile, the balanced corpus contains these varieties in proportions that correspond to the reality of a (sub) language [8]. However, it is impossible to build a general corpus that represents an entire population of a language. It is also impossible to measure the real proportions of language varieties. Therefore, rather than measuring the proportion of the corpus that is built on a language, it is better to focus on building the corpus that primarily attempts to cover the variety of existing texts and their well-designed and documented classification [8]. To form this, the selection of text sampling is also important. In this case, it is necessary to explain the meaning of the text referred to here. 'Texts' are often assumed to be a series of coherent sentences and paragraphs [9]. Some of the consideration in generating a corpus is should a corpus include full texts or carried out only some part of texts. According to [9], the essential feature aspects for the corpus is stylistics and text linguistics. The corpus designer should neutralize as far as possible the effects of sampling bias and stylistic idiosyncrasies of one particular author. This problem can be reduced if texts that are included in the corpus are written by many authors. The next parameter to consider is the size of the corpus. Semantically, corpus-size is the information of the number of sentences, the number of articles or the number of words in the corpus. Ideally, a bigger corpus-size is better because it is increasingly representativeness of a language [10]. But there is no standard and strictly provision corpus-size should be made. However, according to the internal coherence of the component parts, the corpus-size is divided into three types, which are small-corpus, medium-corpus, and large-corpus [11]. Small-corpus is a corpus that was built only by a single author, medium-corpus is a corpus with several types of text and several authors but still in one genre. An example of medium-corpus is The LDC gigaword corpora because although composed almost entirely of news (journalistic process) by many authors, the text source is only from one source, which is from newspapers. In contrast, large-corpus is a corpus containing a variety of text types, genres, and styles that approximates to overall language usage. The brown corpus has 1 Million words and classified a larger range than the Reuters corpus at 100 Million words [11].

In this study, the text was collected from several URL sources using web crawlers and web scraper. Though sometimes crawler and scraper considered the same, there is the main difference between these two terms.

A web crawler more focused on managing which URLs to be download into local storage, and a web scraper is more focus on the content.

2.2 Web Crawler

Web crawler or also commonly known as a web spider or web robot is a program that works with a particular method and automatically collects and indexes information contained in a website [12]. The two main functions of web crawlers are:

a. Identify hyperlinks that are found in the content, and add the hyperlinks to the list of pages to be visited.
b. Conduct a recursive visit process according to application requirements.

Figure 1. describes the main idea of a web crawler work. A set of URLs was given to the crawlers as a starting point to a queue list. These URLs, also known as URL seeds, later on, will be visited one by one by the crawlers. Based on these URLs seed, the browser makes a URL request and download the websites through HTTP protocol into the local machine storage. The crawler also identifies the hyperlinks that appear on the web content and add the hyperlinks in the queue list of URLs seed. Each URL subsequently will be visited and be downloaded until the queue list is empty. It can be concluded; the crawlers make these repetitive works automatically work without too much human intervention.

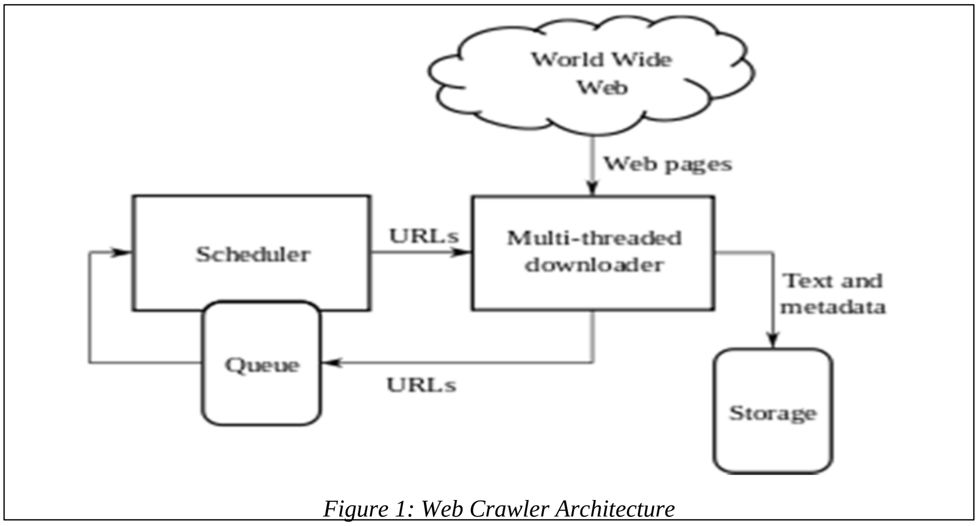


Figure 1: Web Crawler Architecture

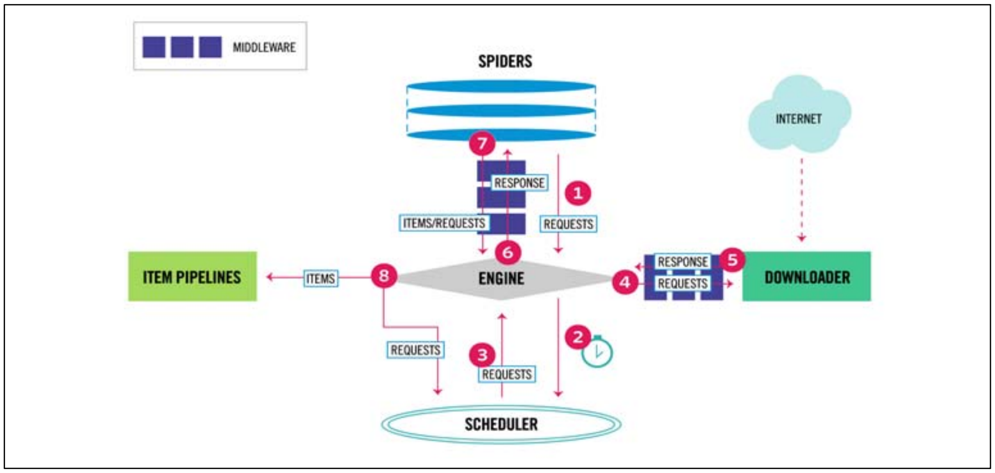


Figure 2: The Architecture of Scrapy [14]

2.3 Web Scraper

Web scraping [13] is a process to get documents from the internet. The documents commonly in the form of a markup language like HTML and XHTML. A web scraper also has tasks to analyze the document's content. A web scraper has some function, as follows:

1. *Create Scraping Template*: scraping template was designed based on the analysis of HTML pages. Each HTML page has various web structures. By analyzing the web structure make it easier to scrape the needed information.
2. *Explore Site Navigation*: A scraper also has to analyze the navigation structure method from websites. *Automate Navigation and Extraction*: Based on the information obtained in steps 1 and 2, a web scraper application was created to automate information retrieval from the specified website.
3. *Extracted Data and Package History*: the information obtained from step 3 is stored in a specific format.

In this study, we utilized a Scrapy framework to generate a crawler and scraper. The work architecture of Scrapy can be seen in Figure 2.

Figure 2. describes the data flow in Scrapy controlled by the execution engine, and runs as follows [14]:

2. *The Engine* gets an initial *request* from *Spiders* to *crawling* the websites.

3. The Engine schedules the request in the Scheduler and asks the next request for crawling
4. *Scheduler* returns the next *request* to the *Engine*.
5. The Engine sends the Requests to the Downloader, passing through the Downloader Middlewares.
6. Once the page finishes downloading, the *Downloader* generates a *Response* and sends it to the *Engine*.
7. *The Engine* receives a *response* from the *Downloader* and sends it to the *Spiders* or the next step.
8. *The Spiders* process the *Response* and return an *item that already scraped* and new *requests* to the *Engine*.
9. *The Engine* sends the processed *item* to *Item Pipelines*, then send process *Request* to the *Scheduler* and ask for possible next request for the crawl.
10. *These processes repeated* until no more requests from the *Scheduler*.

3. RELATED WORKS

As stated earlier, the corpus is one of the most important and also the initial stages of the NLP overall process. The corpus of English already started to build seriously since 1961. The Brown Corpus [15] is an English corpus contains written English text published in the United States in 1961. The published text contains a list of the collection of books, newspapers, magazines, and periodicals in the Brown University. The size of the Brown corpus approximately a million words. Besides Brown Corpus, there is also a corpus that was built from written English published in the United Kingdom around 1961. This Corpus is known as Lancaster-Oslo / Bergen Corpus, which is often abbreviated as LOB Corpus. Corpus LOB is also known as part of the Brown Corpus Family. The Brown Corpus family is the first corpus that was built computerized. The others examples of English corpus that have been widely used by NLP researchers such as the British National Corpus (BNC), the Bank of English (BoE), Corpus of Contemporary American English (COCA), the ICE Corpora, The London Lund Corpus, Helsinki Corpus of English Texts and many more. Unlike English, which has been equipped with many corpora, Bahasa Indonesia is still very limited for the number of the corpus. The research and studies in NLP Bahasa Indonesia mostly oriented to implement NLP methods and not focus on producing Bahasa Indonesia Corpus. However, some studies need corpus as their linguistic resources; therefore, they generate a corpus for their study's purpose. Some of these corpora are shared with the community, and some of the corpora are not published. According to

the study to analyze the development of Computational Linguistic Research for Bahasa Indonesia by [16], one of the first Bahasa Indonesia corpus built by R.R Hardjadibarata (1969) from Monash University. The text resources of this corpus are manually collected from Indonesia newspapers. The study by [16] also generated Bahasa Indonesia text corpus from Kompas newspapers, Kompas is one of the largest newspapers in Indonesia. This study collected 52 editions of Kompas, which are published in the year 1994. This study generated a corpus with 2.200.818 words that were formed by 74.559 unique words. Another study to generate Bahasa Indonesia corpus was done by [17]. This study generates a bilingual corpus that produces Indonesian sentences paired with English sentences. This study generates 45.000 sentences. The text resources of this study are from the PAN localization project [18]. Some further research, like [1] and[19], utilized this corpus as their linguistic resources. Another research also generates Bahasa Indonesia text corpus for specific purposes, like study by [20] [21] to build Bahasa Indonesia text corpus from twitter for sentiment analysis purposes. Based on the background, the purpose of this study is to generate a text corpus Bahasa Indonesia using web corpora approaches.

4. METHODOLOGY

The methodology of this study is described in Figure 3. There are several steps, which are Resources Observation step, Web Structure Analyzing step, Crawling dan Scraping step, and Data Cleaning step

4.1 Resources Observation

The aim of this study is to generate a general corpus for NLP Bahasa Indonesia, where the source of the text comes from the websites. Ideally, the good corpus should meet several parameters like corpus representativeness, corpus balance, corpus sampling, corpus size, and corpus format. The representative corpus contains a large number of texts that should cover many varieties of language. In this study, we observed several resources variation in Bahasa Indonesia, like fiction, magazines, newspapers, and academics papers. However, we face some obstacles for the resources of magazines, fiction, and academics papers. Not many online magazines were available in Bahasa

Indonesia. Most of these magazines are printed-version magazines, and they only display the excerpt of the magazine on their websites. Therefore, many articles in these online versions not in a full version article. As well as fictions in Bahasa Indonesia, most of them are in printed-book versions, and on the internet only show the excerpt of the book. Another obstacle in collecting academics papers in Bahasa Indonesia. Most of the academics paper is in PDF format. We face the copyright and permission issue if we want to collect these academics papers. It can be a time-consuming and increasingly difficult operation. For those reasons, we exclude magazines, fictions, and academics from the resources of our corpus. One of the most available text resources in the formal language of Bahasa Indonesia is from online newspapers. To produce representativeness and balance corpus from newspapers in Bahasa Indonesia, we collected newspaper articles from various newspapers, various authors, and various news categories. As text resources, we chose seven the most significant online newspaper in Bahasa Indonesia with various news categories. We collected the articles from the available archives from the resources in the time-range between the year 2006 until 2018. However, these newspapers have various available archives, therefore the time-interval of collecting the articles is different for each newspaper. The URLs list of the newspaper and time-interval of collecting the articles can be seen in Table 1.

4.2 Web Structure Analyzing

The next step is *the structure of web analyzing*. In this step, the structure of HTML of these URLs is analyzed. The aim of this stage is to optimize the work of Scrapy as a *crawler* and *scraper engine*. For example, in this study, we only collected the title of the news article, the name of the article author, the tags of the article, and the main of the articles. Therefore, we have to analyze the HTML tags and elements that bring the needed information. Manually, we can see the HTML sources from a website by open the "view page source" in the Chrome browser and then choose to inspect. Fortunately, Scrapy already has a tool, namely fetch, to analyze the web structure. Furthermore, the Scrapy also has response.css().extract()commands to extract the data from CSS elements. The extraction process can be done using Xpath. The result of this extraction step can be seen in Figure 4.

Analyzing web structure is a crucial stage because each website has various unique structures to extract the information or elements from the websites. Therefore, it is necessary to test the extraction from the URLs list initially. In this study, the extracted data contains the metadata (title, date, author, and tags) and the main article. This step generates a pattern of web structure analyzing for each URL list from Table 1.

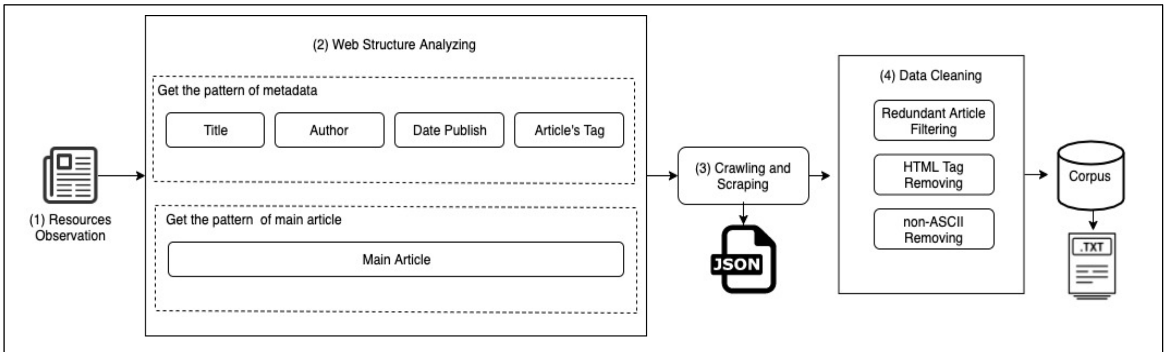

(1) Resources Observation
(2) Web Structure Analyzing
Get the pattern of metadata
Title
Author
Date Publish
Article's Tag
Get the pattern of main article
Main Article
(3) Crawling and Scraping
JSON
(4) Data Cleaning
Redundant Article Filtering
HTML Tag Removing
non-ASCII Removing
Corpus
.TXT

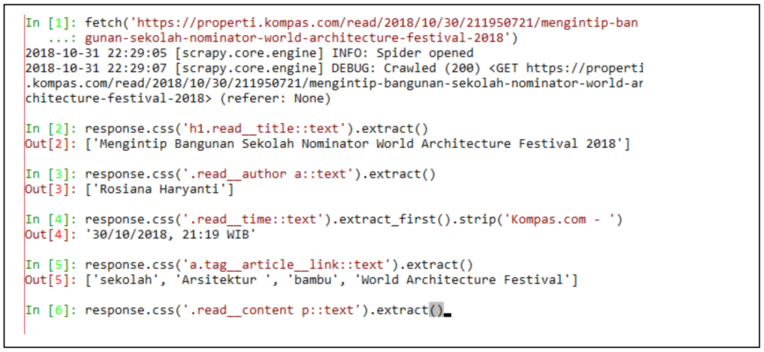
Figure 3: Methodology

Table 1: Newspapers URLs

Newspapers URL	Time-Interval Available Articles
https://edukasi.kompas.com/	19/09/2018 - 24/09/2018
https://lifestyle.kompas.com/	31/03/2017 - 19/10/2018
https://tekno.kompas.com/	07/07/2016 - 19/10/2018

https://sains.kompas.com/	21/10/2016 - 19/10/2018
https://entertainment.kompas.com/	17/01/2018 - 19/10/2018
https://otomotif.kompas.com/	31/05/2017 - 19/10/2018
https://ekonomi.kompas.com/	08/01/2018 - 19/10/2018
https://travel.kompas.com/	16/09/2016 - 19/10/2018
https://properti.kompas.com/	13/07/2016 - 19/10/2018
https://news.kompas.com/	06/09/2018 - 19/10/2018
https://www.merdeka.com/peristiwa/	22/10/2018 - 26/10/2018
https://www.merdeka.com/travel/	04/08/2018 - 25/10/2018
https://www.merdeka.com/politik/	17/10/2018 - 26/10/2018
https://www.merdeka.com/uang/	18/10/2018 - 26/10/2018
https://www.merdeka.com/dunia/	01/10/2018 - 26/10/2018
https://www.merdeka.com/properti/	19/10/2018 - 24/10/2018
https://www.merdeka.com/gaya/	26/08/2018 - 26/10/2018
https://www.merdeka.com/sehat/	11/07/2018 - 25/10/2018
https://www.merdeka.com/teknologi/	18/09/2018 - 26/10/2018
https://www.merdeka.com/sepakbola/	25/09/2018 - 25/10/2018
https://www.antaranews.com/warta-bumi	02/07/2006 - 18/10/2018
https://www.antaranews.com/tekno	01/07/2006 - 18/10/2018
https://www.antaranews.com/hiburan	03/01/2006 - 19/10/2018
https://www.antaranews.com/ekonomi	10/05/2018 - 19/10/2018
https://www.antaranews.com/dunia	01/01/2006 - 21/10/2018
https://www.antaranews.com/nasional	01/01/2006 - 21/10/2018
https://sumutpos.co/rubrik/lifestyle/	21/03/2014 - 27/10/2018
https://sumutpos.co/rubrik/teknologi/	14/01/2015 - 27/11/2018
https://sumutpos.co/rubrik/health/	31/03/2012 - 27/10/2018
https://sumutpos.co/rubrik/internasional/	10/08/2015 - 27/10/2018
https://sumutpos.co/rubrik/ekonomi/	19/03/2016 - 27/10/2018
https://sumutpos.co/rubrik/politik/	27/07/2016 - 27/10/2018
https://sumutpos.co/rubrik/hukum-kriminal/	18/08/2015 - 27/10/2018
http://waspada.co.id/warta/mancanegara/	27/03/2015 - 30/07/2018
http://waspada.co.id/warta/politik/	09/07/2015 - 13/05/2015
http://waspada.co.id/warta/ekonomi-bisnis/	09/07/2015 - 09/07/2018
http://waspada.co.id/warta/teknologi/	02/06/2015 - 26/05/2018
http://waspada.co.id/ragam/kesehatan/	25/06/2015 - 11/06/2018
http://waspada.co.id/ragam/hiburan/	31/10/2015 - 11/06/2018
http://waspada.co.id/ragam/gaya-hidup/	25/07/2015 - 22/06/2018
http://mediaindonesia.com/ekonomi	13/06/2015 - 23/06/2018
http://mediaindonesia.com/internasional	01/06/2015 - 07/06/2018
http://mediaindonesia.com/megapolitan	27/02/2015 - 17/10/2018
http://mediaindonesia.com/nusantara	22/06/2015 - 29/06/2018
http://mediaindonesia.com/olahraga	06/06/2015 - 20/06/2018
http://mediaindonesia.com/politik-dan-hukum	11/06/2016 - 30/06/2018
https://www.suara.com/tekno	27/02/2015 - 18/06/2018

https://www.suara.com/health	04/06/2015 - 03/06/2018
https://www.suara.com/bisnis	17/11/2016 - 22/06/2018
https://www.suara.com/news	30/10/2018 - 30/10/2018
https://www.suara.com/sport	14/08/2015 - 07/06/2018
https://www.suara.com/lifestyle	06/01/2016 - 21/06/2018



```
In [1]: fetch('https://properti.kompas.com/read/2018/10/30/211950721/mengintip-ban
   ...: gunan-sekolah-nominator-world-architecture-festival-2018')
2018-10-31 22:29:05 [scrapy.core.engine] INFO: Spider opened
2018-10-31 22:29:07 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://properti
.kompas.com/read/2018/10/30/211950721/mengintip-bangunan-sekolah-nominator-world-ar
chitecture-festival-2018> (referer: None)

In [2]: response.css('h1.read__title::text').extract()
Out[2]: ['Mengintip Bangunan Sekolah Nominator World Architecture Festival 2018']

In [3]: response.css('.read__author a::text').extract()
Out[3]: ['Rosiana Haryanti']

In [4]: response.css('.read__time::text').extract_first().strip('Kompas.com - ')
Out[4]: '30/10/2018, 21:19 WIB'

In [5]: response.css('a.tag__article__link::text').extract()
Out[5]: ['sekolah', 'Arsitektur ', 'bambu', 'World Architecture Festival']

In [6]: response.css('.read__content p::text').extract()
```

Figure 4: Result of Data Extraction

4.3 Crawling and Scraping

The next step is a process to crawl and scrape based on the pattern of web structure analyze from the previous step. A Scrapy provides spiders that can be filled with orders according to user needs. The crawler engine is responsible for recursively browsing URLs. As long as the crawler engine visits these links, the scraper engine will do its job to extract the page. Within each Scrapy project, there is a file named items.py. This file is useful for storing scraped data and storing it in a specific format according to user needs; in this study, we use the JSON format.

4.4 Data Cleaning

Data Cleaning is a process to clean the scraped data from unwanted data or noisy. The unwanted data, for example, are HTML tags, non-ASCII characters and symbols, and also duplication of spaces. In this step, we also filter the redundant articles by comparing the metadata of each article.

The cleaning process in this study uses the Python programming language, which has many libraries to support the process. One way to do the cleaning process is to use regular expressions or regex. Regex is a language construction to match text based on specific patterns. Regex will help eliminate unneeded characters like HTML tags, and will only leave the data needed. The last steps are to convert clean data into a machine-readable format. In this study, the corpus is stored in text format.

5. IMPLEMENTATION AND EVALUATION

5.1 Scrapy Implementation

The first job that Scrapy does is crawling web pages based on the URLs list. In this study, the Scrapy program is set to visit one web page at a time and will stop at a given time interval. The length of time the crawling process depends on how many web pages are visited by the engine. The more web pages visited, the more time is needed. Apart from these factors, the internet speed factor will also affect the crawling process. In testing 52 news URLs, the crawler engine successfully visited 717,639 web pages. When the crawler engine in Scrapy visits a web page, it also extracts the contents of the web page with its scraper engine. After that, the scraper engine saves the extraction results into an item. One visited and extracted page will represent one item. Each item represents a news article consisting of the title, date published, author or editor, tags, and news content. To better understand, Figure 5. shows the average number of crawled pages, and Figure 6. shows the average number of scraped items.

Not all visited pages can be scraped by the Scrapy. Here is the percentage of data that was successfully scraped using Scrapy on seven news sites.

1. Kompas, (89205 / 98070) x 100% = 90.96%
2. Merdeka, (2751 / 3037) x 100% = 90.58%
3. Antara, (301364 / 328326) x 100% = 91.79%
4. Sumut Pos, (12829 / 13478) x 100% = 95.18%

5. Waspada, (23529 / 26012) x 100% = 90.45%
6. Media Indonesia, (98285 / 207575) x 100% = 47.35%
7. Suara, (38495 / 40674) x 100% = 94.64%

From seven URLs of newspapers, six URLs reach a successful percentage above 90% in scraping web pages. However, there is one news site that is unable to reach 50 percent, namely Media Indonesia. The reason is that the page visited by a crawler engine does not meet the criteria for a scraper machine to do scraping it because the page does not contain text; meanwhile contains photos, videos, and other content that cannot be scraped by Scrapy. Therefore, the average percentage of the scraping process from seven newspaper URLs is 85.85%.

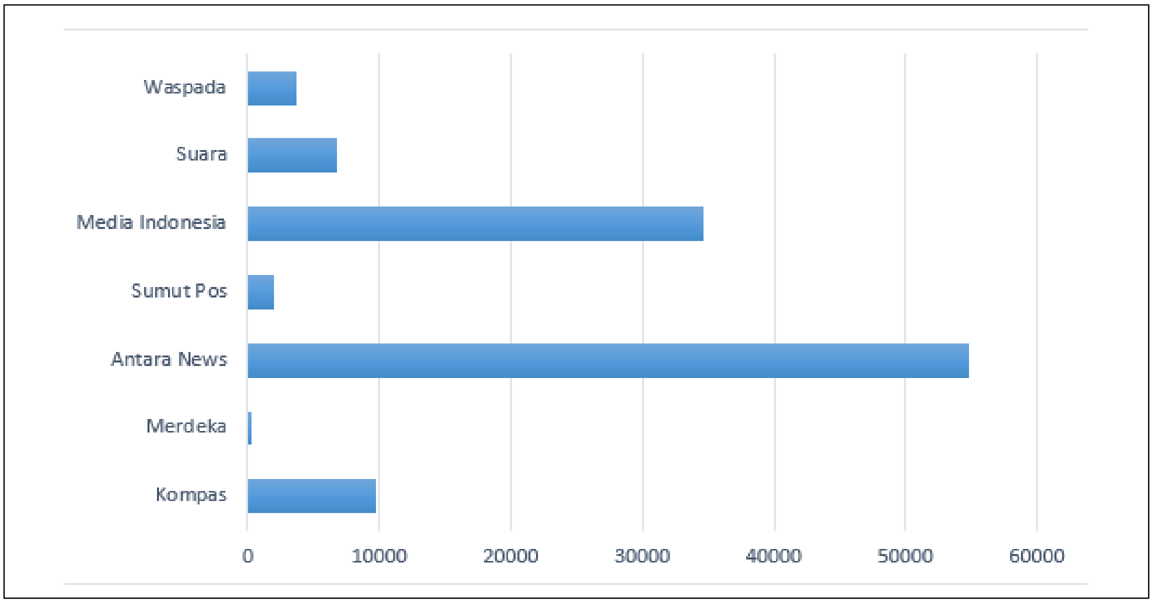


Figure 5: Number of Crawled Pages

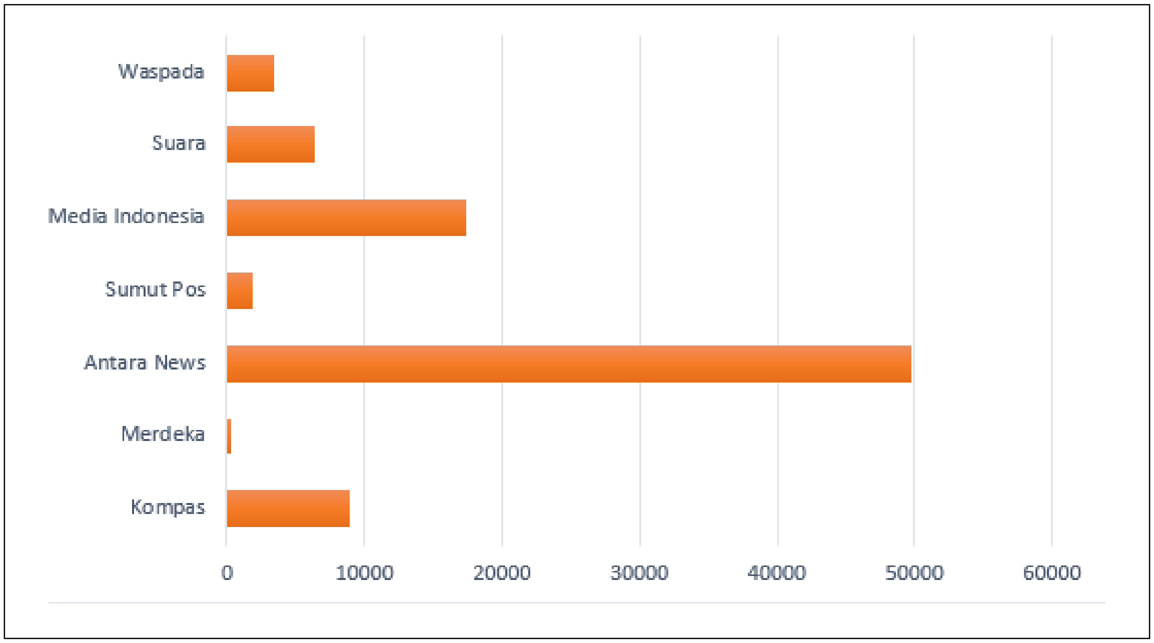


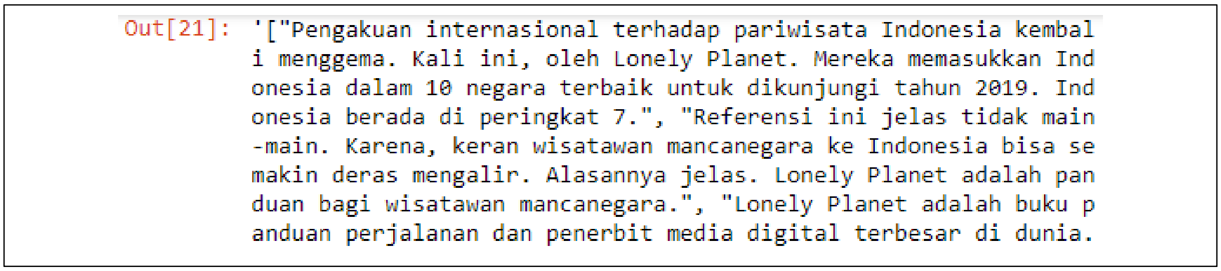
Figure 6: Number of Scraped Pages

5.2 Data Cleaning Implementation

In this study, the final result is a text corpus that contains many sentences in Bahasa Indonesia. The sentences were obtained from crawling and scraping news content from 7 URLs of the largest newspaper sites in Bahasa Indonesia. The cleaning process di a process to remove unwanted data like HTML tags, non- ASCII characters, symbols, and else. The implementation of data cleaning can be described as follows:

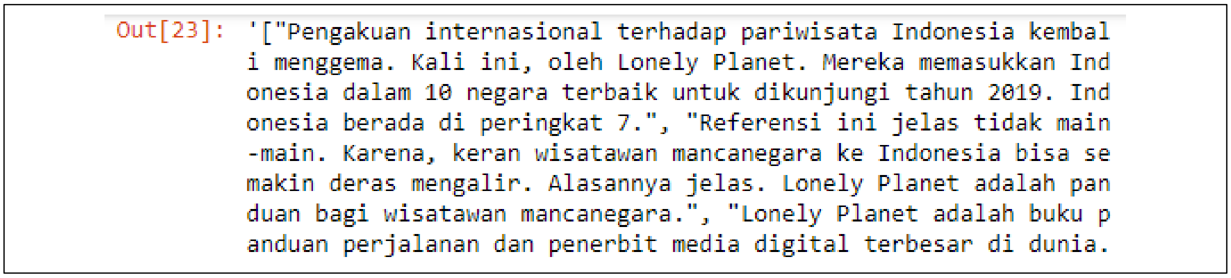
a. Remove non-ASCII characters, with function:
re.sub(‘[^\x00-\x7F]+’, ‘’, text)
b. Remove unwanted symbols, with function:
re.sub(‘\W+’, ‘ ’, text)
c. Remove numeric character, with function:
re.sub(‘\d+’, ‘ ’, text)
d. Remove duplicate space, with function:
re.sub(‘ +’, ‘ ’, text)

The result of the function can be seen in Figure 7, 8, 9, 10, and 11.



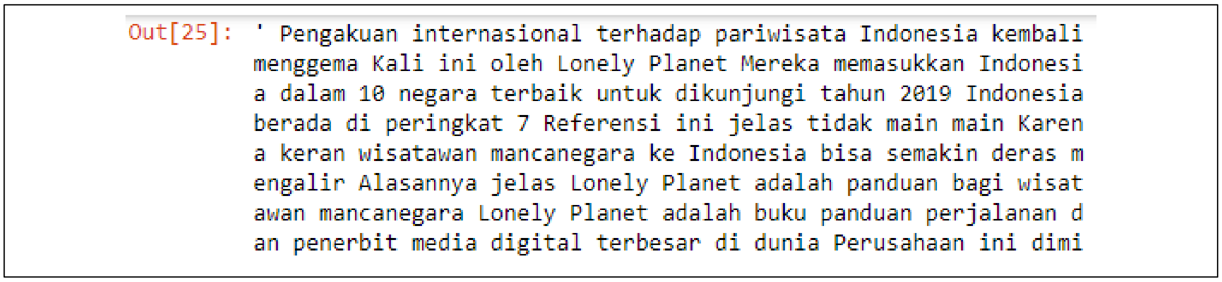
```
Out[21]: '["Pengakuan internasional terhadap pariwisata Indonesia kembal
         i menggema. Kali ini, oleh Lonely Planet. Mereka memasukkan Ind
         onesia dalam 10 negara terbaik untuk dikunjungi tahun 2019. Ind
         onesia berada di peringkat 7.", "Referensi ini jelas tidak main
         -main. Karena, keran wisatawan mancanegara ke Indonesia bisa se
         makin deras mengalir. Alasannya jelas. Lonely Planet adalah pan
         duan bagi wisatawan mancanegara.", "Lonely Planet adalah buku p
         anduan perjalanan dan penerbit media digital terbesar di dunia.
```

Figure 7: The Result of Scraped Data



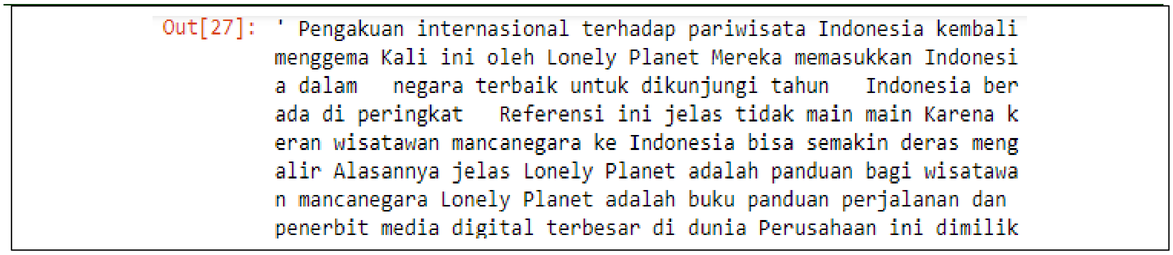
```
Out[23]: '["Pengakuan internasional terhadap pariwisata Indonesia kembal
         i menggema. Kali ini, oleh Lonely Planet. Mereka memasukkan Ind
         onesia dalam 10 negara terbaik untuk dikunjungi tahun 2019. Ind
         onesia berada di peringkat 7.", "Referensi ini jelas tidak main
         -main. Karena, keran wisatawan mancanegara ke Indonesia bisa se
         makin deras mengalir. Alasannya jelas. Lonely Planet adalah pan
         duan bagi wisatawan mancanegara.", "Lonely Planet adalah buku p
         anduan perjalanan dan penerbit media digital terbesar di dunia.
```

Figure 8: Non-ASCII Character Removal



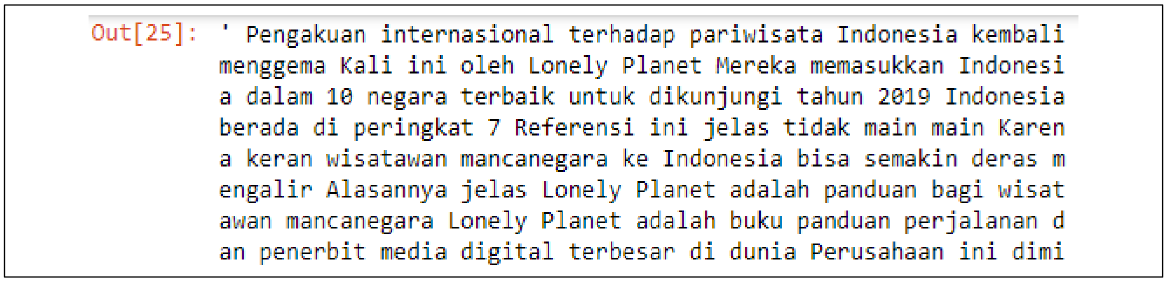
```
Out[25]: ' Pengakuan internasional terhadap pariwisata Indonesia kembali
         menggema Kali ini oleh Lonely Planet Mereka memasukkan Indonesi
         a dalam 10 negara terbaik untuk dikunjungi tahun 2019 Indonesia
         berada di peringkat 7 Referensi ini jelas tidak main main Karen
         a keran wisatawan mancanegara ke Indonesia bisa semakin deras m
         engalir Alasannya jelas Lonely Planet adalah panduan bagi wisat
         awan mancanegara Lonely Planet adalah buku panduan perjalanan d
         an penerbit media digital terbesar di dunia Perusahaan ini dimi
```

Figure 9: Symbol Removal



```
Out[27]: ' Pengakuan internasional terhadap pariwisata Indonesia kembali
         menggema Kali ini oleh Lonely Planet Mereka memasukkan Indonesi
         a dalam   negara terbaik untuk dikunjungi tahun   Indonesia ber
         ada di peringkat   Referensi ini jelas tidak main main Karena k
         eran wisatawan mancanegara ke Indonesia bisa semakin deras meng
         alir Alasannya jelas Lonely Planet adalah panduan bagi wisatawa
         n mancanegara Lonely Planet adalah buku panduan perjalanan dan
         penerbit media digital terbesar di dunia Perusahaan ini dimilik
```

Figure 10: Numeric Removal



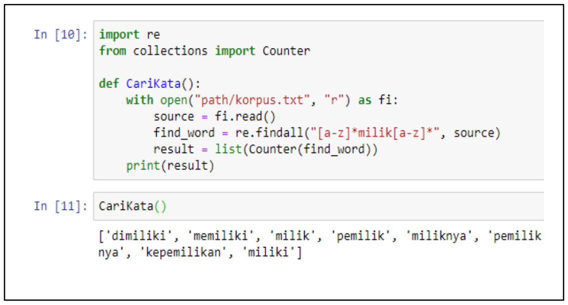
```
Out[25]: ' Pengakuan internasional terhadap pariwisata Indonesia kembali
         menggema Kali ini oleh Lonely Planet Mereka memasukkan Indonesi
         a dalam 10 negara terbaik untuk dikunjungi tahun 2019 Indonesia
         berada di peringkat 7 Referensi ini jelas tidak main main Karen
         a keran wisatawan mancanegara ke Indonesia bisa semakin deras m
         engalir Alasannya jelas Lonely Planet adalah panduan bagi wisat
         awan mancanegara Lonely Planet adalah buku panduan perjalanan d
         an penerbit media digital terbesar di dunia Perusahaan ini dimi
```

Figure 11: Redundant Space Removal

5.3 Corpus Evaluation

Corpus evaluation is a process to evaluate the corpus whether the corpus is good enough to runs in accordance with user needs.

The corpus that is generated in this study using web corpora approaches from 52-page categories from 7 URLs Indonesian newspaper 569,458 news articles with 219,392 distinct words. To evaluate the corpus, we implemented the expansion of the stem word. For example, word "didik" (in English: educate) has word expansion like "didikan" (in English: education), "pendidik" (in English: educator), "Pendidikan" (in English: education) and "pendidikannya" (in English: his education). Figure 12 is the excerpt of syntax to searching word expansion using the corpus that is generated in this study.



```
In [10]: import re
         from collections import Counter

         def CariKata():
             with open("path/korpus.txt", "r") as fi:
                 source = fi.read()
                 find_word = re.findall("[a-z]*milik[a-z]*", source)
                 result = list(Counter(find_word))
             print(result)

In [11]: CariKata()

         ['dimiliki', 'memiliki', 'milik', 'pemilik', 'miliknya', 'pemilik
         nya', 'kepemilikan', 'miliki']
```

Figure 12: The excerpt of Syntax to word expansion

Some word expansion examples from this evaluation step can be seen in Table 2.

Table 2: Word Expansion Examples

Root Word	Word Expansion
milik	dimiliki, kepemilikan, kepemilikannya, miliki, milknya, memiliki, pemilik, pemiliknya
salah	bersalah, bermasalah, kesalahan, kesalahannya, masalah, msalahnya, mempermasalahkan, permasalahan
hasil	alhasil, berhasil, dihasilkan, hasilnya, keberhasilan, menghasilkan, penghasilan
harap	berharap, diharap, diharapkan, harapan, harapannya, harapkan, mengharapkan
beda	bedagai, bedakan, berbeda, dibedakan, membeda, membedakan, perbedaan

dukung	didukung, dukungan, mendukung, pendukung
tahu	diketahui, ketahuan, ketahui, memberitahu, memberitahukan, mengetahui, pemberitahuan, dengetahuan, pengetahuannya, sepengetahuan, sepengetahuannya, setahun, tahukah, tahun, tahunnya
guna	digunakan, gunakan, menggunakan, mendayagunakan, multiguna, pendayagunaan, pengguna, penggunaan, penggunanya, penyalahgunaan, serbaguna
jaya	baktijaya, berjaya, brawijaya, dharmawijaya, jayakarta, jayakerta, jayamakmur keberjayaan, menjayakan, sanjaya, sukmajaya, sukmawijaya, sujaya, wijaya, wijayanto, wirawijaya
jaya	baktijaya, berjaya, brawijaya, dharmawijaya, jayakarta, jayakerta, jayamakmur keberjayaan, menjayakan, sanjaya, sukmajaya, sukmawijaya, sujaya, wijaya, wijayanto, wirawijaya

According to the experiments, we can find many word expansions from the corpus-based on a root word. This experiment proves that the corpus can be utilized for the NLP application in Bahasa Indonesia. Moreover, the corpus can reach some words that are rarely used in Bahasa Indonesia like “bedagai”, “sujaya” and “dharmawijaya”.

5. CONCLUSION

The aim of this study is to generate a general corpus for NLP Bahasa Indonesia, where the source of the text comes from the websites. The research stages contain resource observation, web structure analyzing, website crawling, scraping, and data cleaning. Ideally, the good corpus should meet several parameters like corpus representativeness, corpus balance, corpus sampling, corpus size, and corpus format. However, in the resource observation stage, we face some obstacles for resources of magazines, fiction, and academics papers. Therefore in this study, we choose the newspaper as the main source. To produce a representativeness and balance corpus from sampling newspapers in Bahasa Indonesia, we collected newspaper articles from various newspapers, various authors, and various news categories. We chose seven online newspapers in Bahasa Indonesia, which are Kompas, Merdeka, Antara News, Sumut Pos, Media Indonesia, Suara, and Waspada. We collected the articles of the available archive between the year 2006 until the year 2018. In the web structure analyzing stage, we found each newspaper has various range of time-interval in providing the archive articles. In this study, the percentage of successful crawling content from the resources is 85.85% or 569.456 news articles, with 219.392 distinct tokens. All the collection of the sentence then arranged into a corpus with the machine-readable format. The corpus evaluation by implementing the expansion of the stem word. For example, word "didik” (in English: educate) has word expansion like “didikan" (in English: education), "pendidik" (in English: educator), "Pendidikan" (in English: education) and "pendidikannya" (in English: his education). According to the experiments, we can find many word expansions from the corpus-based on a root word. This experiment proves that the corpus can be utilized for the NLP application in Bahasa Indonesia. It can be concluded that web corpora approaches can produce text corpus for Bahasa Indonesia.

ACKNOWLEDGMENT

The authors gratefully acknowledge that the present research is supported by Lembaga Penelitian Universitas Sumatera Utara. The support is under the research grant TALENTA USU of Year 2019 Contract Number 196/UN5.2.3.1/PPM/KP-TALENTA USU/2019

REFERENCES

[1] S. D. Larasati, V. Kuboň, and D. Zeman, “Indonesian morphology tool (MorphInd): Towards an Indonesian corpus,” in *Communications in Computer and Information Science*, 2011, vol. 100 CCIS, pp. 119–129.

[2] A. S. Wibawa and A. Purwarianti, “Indonesian Named-entity Recognition for 15 Classes Using Ensemble Supervised Learning,” *Procedia Comput. Sci.*, vol. 81, no. May, pp. 221–228, 2016.

[3] “POS Tag - Corpus Site.” [Online]. Available: http://bahasa.cs.ui.ac.id/postag/corpus.

[Accessed: 12-Nov-2019].

[4] "Indonesian web corpus search | Sketch Engine." [Online]. Available: https://www.sketchengine.eu/indonesianwac-corpus/?highlight=indonesian. [Accessed: 12-Nov-2019].

[5] M. A. Baroni Silvia Bernardini AE Adriano Ferraresi AE Eros Zanchetta, "The WaCky wide web: a collection of very large linguistically processed web-crawled corpora," 2009.

[6] C. Biemann, G. Heyer, U. Quasthoff, and M. Richter, "The Leipzig Corpora Collection: Monolingual Corpora of Standard Size."

[7] E. Tognini-Bonelli, "Corpus linguistics at work," 2001.

[8] M. Křen *et al.*, "SYN2015: Representative corpus of contemporary written Czech," in *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, 2016, pp. 2522–2528.

[9] S. Atkins, J. Clear, and N. Ostler, "Corpus design criteria," *Lit. Linguist. Comput.*, vol. 7, no. 1, pp. 1–16, 1992.

[10] A. Kilgarriff and G. Grefenstette, "Introduction to the Special Issue on the Web as Corpus," *Computational Linguistics*, vol. 29, no. 3. pp. 333–347, Sep-2003.

[11] A. Kornai, P. Halácsy, V. Nagy, C. Oravecz, V. Trón, and D. Varga, "Web-based frequency dictionaries for medium density languages," in *aclweb.org*, 2006, pp. 1–8.

[12] A. Amalia, D. Gunawan, A. Najwan, and F. Meirina, "Focused crawler for the acquisition of health articles," in *Proceedings of 2016 International Conference on Data and Software Engineering, ICoDSE 2016*, 2017.

[13] R. Mitchell, *Web Scraping with Python Collecting Data From The Modern Web*. 2015.

[14] "Architecture overview — Scrapy 1.2.2 documentation." [Online]. Available: https://scrapy-ja.readthedocs.io/ja/stable/topics/architecture.html. [Accessed: 09-Nov-2019].

[15] W. N. Francis, "A Standard Corpus of Edited Present-Day American English," *Coll. English*, vol. 26, no. 4, p. 267, Jan. 1965.

[16] B. Nazief, "Development of Computational Linguistics Research: a Challenge for Indonesia," *Proc. 38th Annu. Meet. Assoc. Comput. Linguist. - ACL '00*, pp. 1–2, 2000.

[17] S. D. Larasati, "IDENTIC corpus: Morphologically enriched Indonesian-english parallel corpus," in *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, 2012, pp. 902–906.

[18] "PAN Localization." [Online]. Available: http://www.panl10n.net/english/OutputsIndonesia2.htm. [Accessed: 02-Nov-2019].

[19] A. Dinakaramani, F. Rashel, A. Luthfi, and R. Manurung, "Designing an Indonesian Part of speech Tagset and Manually Tagged Indonesian Corpus."

[20] A. F. Wicaksono, C. Vania, B. T. Distiawan, and M. Adriani, "Automatically building a corpus for sentiment analysis on Indonesian tweets," in *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation, PACLIC 2014*, 2014, pp. 185–194.

[21] A. Amalia, W. Oktinas, I. Aulia, and R. F. Rahmat, "Determination of quality television programmes based on sentiment analysis on Twitter," in *2nd International Conference on Computing and Applied Informatics 2017*, 2018.