CHAPTER 4
ANALYSIS AND DESIGN

4.1 Analysis

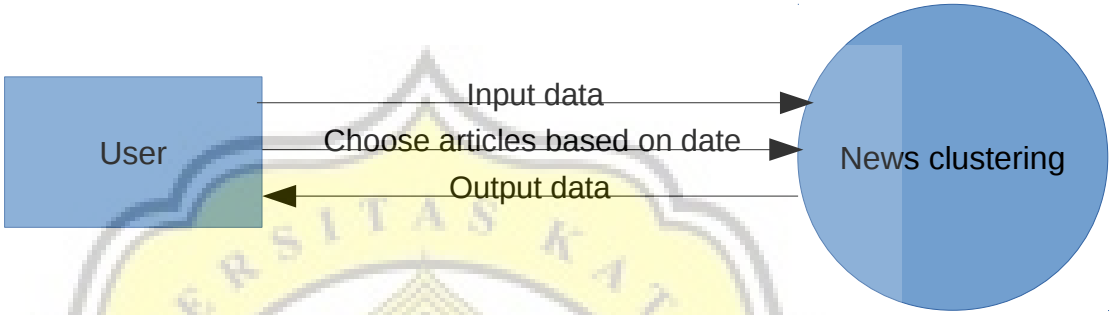

Illustration 4.1: Data Flow Diagram

As seen in the illustration, first user will upload news articles. For example user upload 3 news article titled “Ini 3 Indikator yang Dipantau Pemerintah untuk Tetapkan Harga BBM”, ”Pengusaha Tanggapi Target Pertumbuhan Ekonomi Pemerintah”, and “Subsidi Akan Disalurkan ke 2.500 Desa yang Belum Teraliri Listrik”. Then user choose online news source from bisniskeuangan.kompas.com from 8 July 2017. After user submit it, program do the text preprocessing that consists of tokenization, stopword removal and stemming. In stopword removal process, in the news titled “Subsidi Akan Disalurkan ke 2.500 Desa yang Belum Teraliri Listrik”, found the word “hingga” and “desa”. “Hingga” included in the stopword list, so it removed from the article. But the word “desa” not included in the stopword list, so it is not removed from the article. Then in stemming process, found the word “teraliri” and “alokasi”. The word “teraliri” stemmed to “alir”. While the word “alokasi” not stemmed.

After the text preprocessing done, save all the user news articles words to bag-of-words. Then do term weighting. Term weighting is done with TF-IDF.

TF(t) = Number of times term t appears in a document

IDF(t) = log_e(Total number of documents / Number of documents with term t in it)

TF-IDF(t) = TF x IDF

There are 207 words saved on bag-of-words, like “alih”, ”alir”, “desa”, “harga”, ”indikator”, etc. In the example news, for TF calculation, word “desa” appear 5 times. Then for DF calculation, that combining user news articles and online news articles, word “desa” appear in 2 news articles. For IDF calculation, word “desa” has value log_e(19/2) = 0.97772360528885. Then for TF-IDF calculation for the example news, word “desa” has value 5x0.97772360528885 = 4.888618026. Then make term weighting result as data for K-means algorithm calculation.

To determine data cluster, calculate the distance between data where the data is user news article and online news article. User news article in here will be the centroid. Calculation done with euclidean distance.

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

Where x is the online news article TF-IDF calculation result and y is the user news article TF-IDF calculation result. After the data clustered based on the nearest distance to the centroid, find new centroid with calculate the average data from each cluster.

$$C_K=(\frac{1}{n_K})\sum d_i$$

Where n_K is amount of data and d_i is the average data. After that, repeat the process from beginning with the new centroid . The process is repeated until new centroid value not changed. After the final centroid is found, can be known every data that located in the same cluster.

For example in the K-means calculation, online news titled “Pastikan Tak Ada Kenaikan, PLN Justru Sebut Ada Penurunan Tarif Dasar Listrik” selected.

The euclidean distance between selected news with the first user news article, “Ini 3 Indikator yang Dipantau Pemerintah untuk Tetapkan Harga BBM” is 14.645963757346, with the second user news article, ”Pengusaha Tanggapi Target Pertumbuhan Ekonomi Pemerintah” is 19.977734263988, and with the third user news article, “Subsidi Akan Disalurkan ke 2.500 Desa yang Belum Teraliri Listrik” is 11.706900342713. It means the selected news include in the same cluster with the third user news article. Then count the average data in the same cluster to find new centroid. Because the new centroid value is changed, the process is repeated. In here the process is repeated 5 times until the final centroid is found. After the final centroid is found the selected news still in the same cluster with the third cluster, which mean online news article “Pastikan Tak Ada Kenaikan, PLN Justru Sebut Ada Penurunan Tarif Dasar Listrik” related with user article “Subsidi Akan Disalurkan ke 2.500 Desa yang Belum Teraliri Listrik”.

4.2 Design

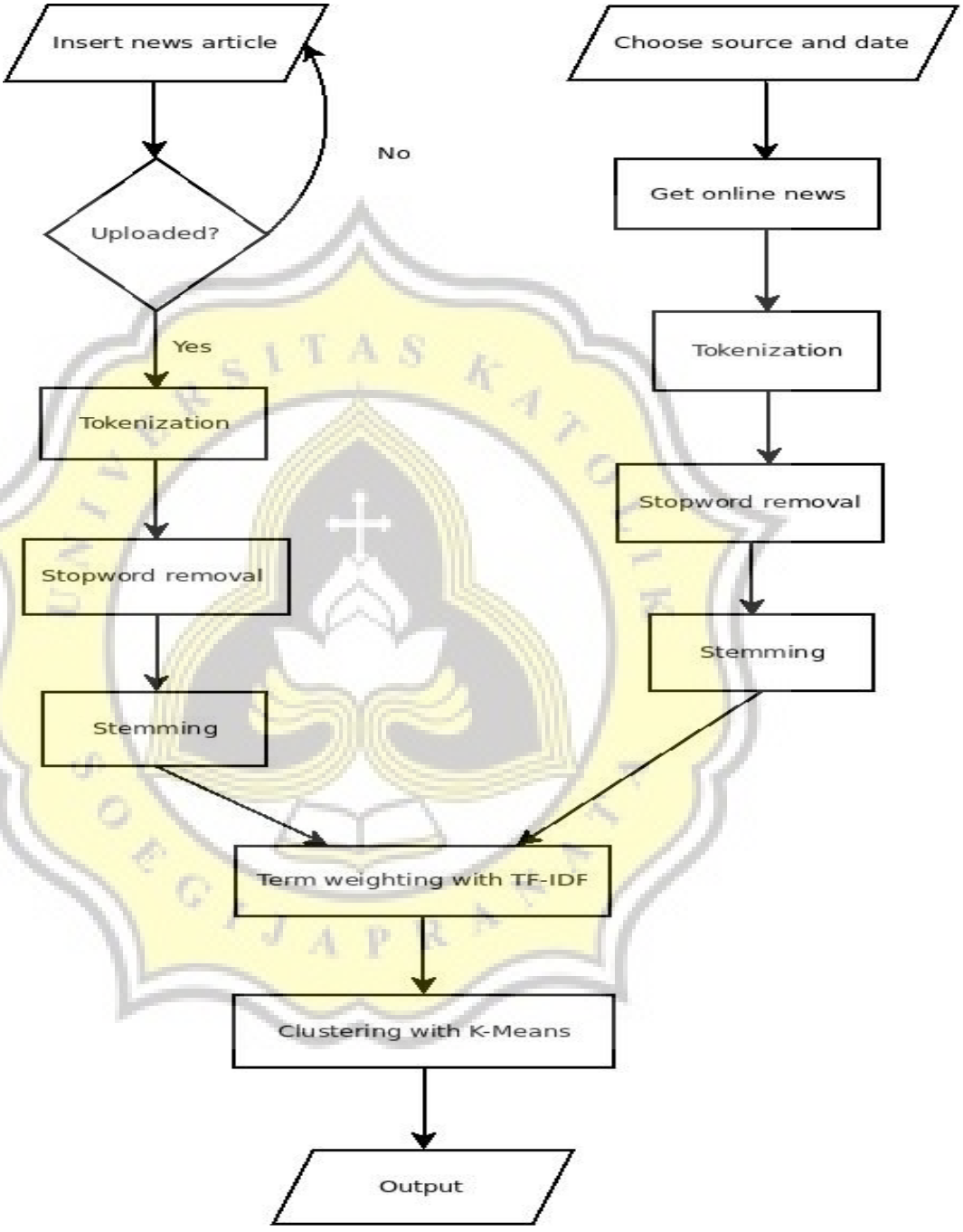

Illustration 4.2: Flowchart

In the beginning, program prompt user to submitting several news article file as the main cluster data (centroid). Then choose online news articles as the data from specific date that will be take from online news website. After

submitted, if the news articles are not uploaded, program will ask user to upload it again. If the news articles are uploaded, all the news stored into array. Then, program will take online news articles from selected online news web (kompas ekonomi, kompas otomotif, kompas tekno, kompas travel) and selected date. The online news articles also stored into array. Next program will do text preprocessing for the user news articles. First, tokenize the articles. The process will separate and lowercase all the words, remove all the html tag and character like dot, comma, strip, colon, semicolon,etc. Then remove the stop words like "ada", "dan", atau, etc. The stopwords checked from the list. If there is any stopword, it will be removed. Then stem all the words. It will change affixed word into root word. For example, "membantu" being "bantu". Nazief-Adriani algorithm is used for the stemming process. After the text preprocessing is done, make bag-of-words taken from existing words from user inputed articles. After that, program do term weighting process (TF-IDF) to make the data countable in K-means algorithm. TF-IDF process consists of TF, DF and IDF. TF (Term Frequency) is the process for count a word emergence. DF (Document Frequency) is the process for count a document that have the same word. IDF (Inverse Document Frequency) is the logarithm of DF. TF and IDF result multiplied and be the final result of TF-IDF.[2]

Next, do the preprocessing and term weighting to online news articles. The process is the same as that done to the user articles. After that, do the K-means clustering with online news articles term weighting results as data and user news articles term weighting results as centroid. Finally, program will display the clustering result. User can choose to read the news article directly from the page.

2 http://tfidf.com. Accessed June 30, 2017