CHAPTER 4
ANALYSIS AND DESIGN

4.1 Analysis

1) Collecting data

In this study, the data needed can be retrieved through the web labs.semanticscholar.org/corpus. The data is a list of titles of scientific journals in the form of text. Therefore, pre-processing is needed to make the research data suitable for topic modeling. This Study uses 1047 document as the corpus. The example of the corpus can be seen in Table 4.1.

Table 4.1: Stages of Pre-Processing

No	ID	Abstract	Title
1	e1dffd3aba88348a140485822e36a8f17d07fc52	AIMnTo investigate whether a grading system of renal AMLs based on digital subtraction angiography DSA and computerized tomography CT, could help to select patients for embolization	Large renal angiomyolipomas: digital subtraction angiographic grading and presentation with bleeding
2	d90d81a6faeead07d9a1a653632f198ee5a41fc8	A method for reducing the approximation errors introduced in the Rytov inversion of the inhomogeneous wave equation is presented.	An Improvement to the Rytov Approximation in Diffraction Tomography
3	344d749968c98b2c3341174	A mathematical simulation of flow regulation in vascular networks is	Functional sympatholysis and

	0f943e36ec799cea3	used to investigate the interaction between arteriolar vasoconstriction due to sympathetic nerve activity SNA and vasodilation due to increased oxygen demand.	sympathetic escape in a theoretical model for blood flow regulation
4	38efa367a0bbb372bafd956d2a3ff385565486d6	-	Effect of occlusion of the aorta on the coronary and pulmonary circulation
5	7d5da51df21bbe4d1589dfdb6a080a3c1b1b9baa	OBJECTIVEnTo analyse the relationship between the presence of malnutrition MN, as measured by the NRS-2002 nutritional evaluation, and the rate of morbidity and mortality....	Association between nutritional risk based on the NRS-2002 test and hospital morbidity and mortality

2) Pre-Processing

There are several stages in preprocessing. The results of each step are stored in the bag of Words. In a journal written by Sriurai, Wongkot (2011) states that the Bag of Words is used to prepare the terms to be used in processing. The first step in pre-processing is Tokenizing. In this study, the tokenizing stage is done to separate each word in the title of a scientific journal. The next stage is cleaning and stopword. The cleaning stage is the stage of removing various symbols and numbers, while the stopword stage is done to delete words that are in the stopword list. The cleaning stage is also often used to change each letter in a word to lowercase (lowercase). Then, there is a stopword process that removes conjunctions and words that do not have special

or important meanings. The final step in pre-processing is stemming. Stemming is the process of converting existing words into standard words. Examples of pre-processing stages can be seen in table 4.2.

Table 4.2: Stages of Pre-Processing

Before pre-processing	After pre- processing	Stage
Effect of occlusion of the aorta on the coronary and pulmonary circulation	Effect of occlusion aorta on coronary pulmonary circulation.	Tokenizing
Effect of occlusion aorta on coronary pulmonary circulation.	effect of occlusion aorta on coronary pulmonary circulation	Cleanning + lowcase
effect of occlusion aorta on coronary pulmonary circulation	effect occlusion aorta coronary pulmonary circulation	Stopwords
effect occlusion aorta coronary pulmonary circulation	effect occlus aorta coronari pulmonari circul	Stemming

3) TF-IDF and Density

TF-IDF is a way of spreading words on each document and word on the entire corpus. The TF (Term-Frequency) - IDF (Inverse Document

Frequency) table is the media to look for the sparse value of the research material data examined.

$$\mathrm{tf}(t,d) = f_{t,d}.$$

Illustration 4.1: TF- formula

Source: (https://en.wikipedia.org/wiki/Tf%E2%80%93idf)

In Illustration 4.1 it is shown that each word in the designated document is the raw count of each word per document.

	larg	renal	angiomyolipoma	digit	subtract	angiograph	...
0	1	1	1	1	1	1	...
1	0	0	0	0	0	0	...
2	0	0	0	0	0	0	...
3	0	0	0	0	0	0	...
4	0	0	0	0	0	0	...
5	0	0	0	0	0	0	...

Illustration 4.2 : Term Frequency Table

lustration 4.2 is the result of TF calculation. On the top row, there are words that are on the corpus. Whereas the first column is the document index. Number 1 indicates the designated word in the document, while the number 0 indicates the absence of the designated word for the document.

Search data density values are often referred to as density. The sparse value is a numeric value that has a range of 0 to 1. The results of sparsity are used to inform the appearance of each word in each

document. Density requires the number of occurrences of words in the entire corpus to be divided by multiplying the large TF matrix. Calculation of density can be written as follows:

```
a = index.shape #get matrix size
density = 0 #declare variable
total = a[0] * a[1] #calculate matrix area
for i in range(rownumber): #looping until end ofdocument
 for n in range(len(data2)): #looping until end of bag of words
  if index[data2[n]][i] > 0: #search words with not null value
            density += 1 #counting words total
density = (density*100) / (total*1.0) #calculate the density
```

Through these calculations, the density value of this research corpus is known as 0.197346369779%

4) Topic Modeling

Through the search results of density, you can know that the data used has a very broad word spread, in other words, the data used is very sparse. Based on these results, it can be known that this research requires special algorithms to do topic modeling. This is because the data used cannot do topic modeling by means of statistics (conventional). One topic modeling algorithm that can be used is Latent Dirichlet Allocation (LDA). In the Handbook of the Latent Semantic Analysis Journal titled “Probabilistic topic models” stated that the LDA can simplify the problem of drawing conclusions by the statistical method by (Steyvers & Griffiths, 2017). Through the sample body in table 4.1, the LDA do grouping so that groups of documents and keywords are formed on each topic.

The LDA grouping process requires several parameters such as Betta, Alpha, number of topics, and the number of iterations so that it can produce groups of data. Group data keywords on each topic can be seen

in table 4.3 while the document group on each topic can be seen in table 4.4.

Table 4.3: Sample List of Keywords

No	Topic 0	Topic 1	Topic 2
1	digital	angiomyolipoma	bleed
2	presentation	angiograph	tomography
3	presentation	sympathetic	blood
4	base	sympatholy	aorta
5		nutrition	coronari
6			pulmonari

Table 4.4: Sample List of Document's Title

No	Topic 0	Topic 1	Topic 2
1	-	-	Large renal angiomyolipomas: digital subtraction angiographic grading and presentation with bleeding
2	An Improvement to the Rytov Approximation in Diffraction Tomography	-	-

3	-	Functional sympatholysis and sympathetic escape in a theoretical model for blood flow regulation	-
4	-	-	Effect of occlusion of the aorta on the coronary and pulmonary circulation
5	-	-	Association between nutritional risk based on the NRS-2002 test and hospital morbidity and mortality

4.2 Desain

A) Flowchart Workflow Program

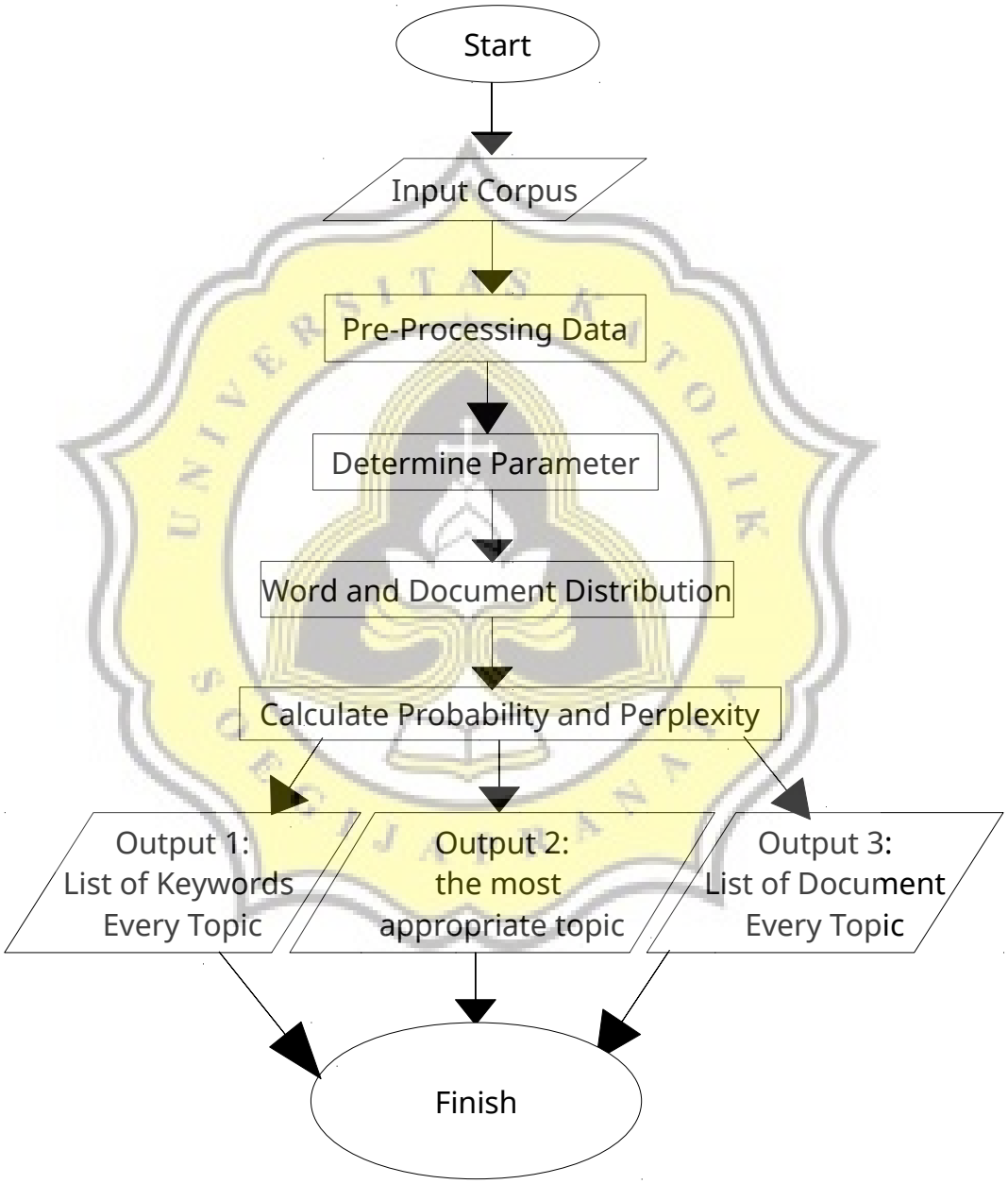

Illustration 4.3 : Main Flowchart

Illustration 4.3 is the stage in working on topic modeling using LDA. The stages of research start from inputting data, pre-processing, filling in parameter values, distributing the matrix, calculating the probability (Gibbs Sampling) and perplexity. The topic modeling process produces 3 outputs. The first result is a list of keywords for each topic, the most appropriate number of topics, and a list of documents for each topic.

B) Flowchart Pre-Processing

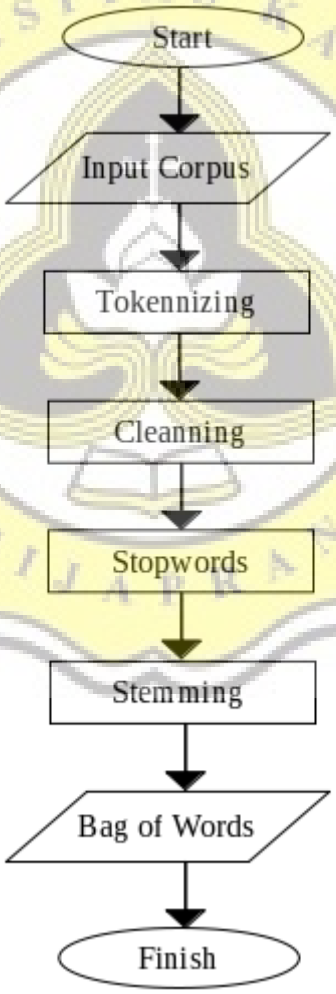

Illustration 4.4: Pre-Processing Flowchart

In the pre-processing stage, it is divided into several stages including tokenizing, cleaning, lowercase, stopword, and stemming. The end of the pre-processing stage is stemming by producing bags of words to be processed using the Latent Dirichlet Allocation Algorithm Collapse Gibbs Sampling, but before entering the topic modeling phase, steps are taken to find the spread of words in the document using Term-Frequency and density. The steps in preprocessing can be seen in Illustration 4.4.