CHAPTER 5
IMPLEMENTATION AND TESTING

5.1 Implementation

In the beginning, user is prompted to upload news article as the main cluster data (centroid) on index page. Then user will choose where and what date the online news articles will be take as the data. After user press the submit button, program will upload inputed news articles. The inputed news articles are read by program from folder and inserted into array. Next, program will take online news articles from selected online news web (kompas ekonomi, kompas otomotif, kompas tekno, kompas travel) and selected date with function getKompasArticles. The program then insert the online news articles title and content into array.

After that, program do the text preprocessing. First step in the text preprocessing is tokenizing with calling function tokenize from Tokenization class. Second step is stopword removal with calling function removal from Stopword class. The class load the list of stopwords that has been saved in a text file. It will check if there are any stopwords in the news article. If stopword exist, it will be removed. Third step is stemming with calling checkWord from Stemmer class. After the text preprocessing is done, existing words from user articles saved into bag-of-words, then program do the term weighting (TF-IDF). The result of TF-IDF inserted into two-dimensional array.

```
1.function KMeans($data,$c){
2. $this->cluster = $c;
3. $ed = $this->findEuclideanDistance($data,$c);
4. $nearest = $this->findNearestDistance($ed,$data);
5. $newc = $this->findNewCentroid($nearest,$data);
6. $this->euclidean = $ed;
7. $this->relate = $nearest;
```

```
8. if($newc!=$c){
9.  $this->KMeans($data,$newc);
10. }
11.}
```

The code above is the main function of K-means clustering. The result of TF-IDF from inputed news article and the result of TF-IDF from online news article inserted into this function. Inside the function, in row 2, store the main cluster data point into variable $this→cluster. Row 3 call the findEuclideanDistance function. Row 4 insert euclidean distance calculation result into findNearestDistance function. Row 5 call the findNewCluster function. Row 6 and 7 store euclidean distance and nearest data calculation result into variable. Row 8, if the new centroid data point not equals with the old centroid data point, program will do the recursive process until new centroid data point not changed. Like in row 9 where it will call the function KMeans again and insert new centroid data point.

```
1.function findEuclideanDistance($data,$c){
2. for($i=0;$i<count($c);$i++){
3.  if(isset($c[$i])){
4.   $keys = array_keys($c[$i]);
5.   for($k=0;$k<count($data);$k++){
6.    $pow = 0;
7.    for($j=0;$j<count($keys);$j++){
8.     $pow += ($data[$k][$keys[$j]]-$c[$i][$keys[$j]])²;
9.    }
10.   $cluster[$k] = sqrt($pow);
11. }
12. $ed[$i] = $cluster;
13. }
14. else{
15.  $c[$i] = null;
16. }
17.}
18.return $ed;
```

```
19.}
```

The code above is the function to calculate the euclidean distance. In row 2 is the code for looping process as much as number of centroid, in this case is user news articles. Row 3 will check if in a centroid there is data in the same cluster. If there is data in the same cluster, in row 4 it will store the data order in a cluster to variable $keys. Row 5 will loop as much as number of online news data. Row 6, set the variable $pow to 0. Row 7 will loop as much as data order in a cluster. Row 8 calculate the subtraction between data of term weighting result from online news article with data of term weighting result from user news article then the subtraction result will be powered. Row 10 save the root of the row 8 calculation result. Row 12 save a cluster euclidean distance calculation to 2-dimensional array. Then if there is no data in a cluster, a cluster euclidean distance will be set as null. Row 18 return the euclidean distance calculation result.

```
1.function findNearestDistance($ed,$data){
2. for($a=0;$a<count($data);$a++){
3.  $column = array_column($ed, $a);
4.  $min = array_keys($column, min($column));
5.  $neard[$min[0]][]= $a;
6. }
7. ksort($neard);
8. for($i=0;$i<count($neard);$i++){
9.  if(!isset($neard[$i])){
10.  $neard[$i] = null;
11. }
12. ksort($neard);
13.}
14.return $neard;
15.}
```

The code above is the function to find the nearest distance between a data with all the centroid. Row 2 loop process as much as number of online news articles. Row 3 return the data from the $a column. Row 4 find lowest value between a online news article data with all the centroid. Row 5 store the data row

number. Row 7 sort the centroid number. Row 8 loop process as much as number of centroid. Row 9, if there is no nearest data with the centroid, row 10 set the centroid nearest data as null. Row 14 return the nearest distance data.

```
1.function findNewCentroid($neard,$data){
2. $key = [];
3. $key = array_keys($data[0]);
4. for($i=0;$i<count($neard);$i++){
5.  if(isset($neard[$i])){
6.   $count = count($neard[$i]);
7.   for($k=0;$k<count($key);$k++){
8.    $sum = 0;
9.    for($j=0;$j<$count;$j++){
10.    $sum += $data[$neard[$i][$j]][$key[$k]];
11.   }
12.   $newc[$i][$key[$k]] = $sum/$count;
13.  }
14.}
15.else{
16. for($k=0;$k<count($key);$k++){
17.  $newc[$i][$key[$k]] = null;
18. }
19.}
20.}
21.return $newc;
22.}
```

The code above is the function to find new centroid value. Row 2 initiate array $key. Row 3 store the bag-of-words data order. Row 4 loop process as much as nearest distance data with a centroid. Row 5 check if there is there is nearest distance data with the centroid, row 6 will count amount of nearest data. Row 7 loop process as much as the bag-of-words data. Row 8 set the variable $sum to zero. Row 9 loop process as much as amount of a centroid nearest data. Row 10 sum all euclidean distance from the same cluster. Row 12 calculate the average by divide data summation with data amount to find the new centroid. Row 15 if there

is no nearest distance from a centroid, set the average as null in row 17. Row 21 return the new centroid.

Finally after all the clustering process done, the program will show the clustering result.

5.2 Testing

For the test, 3 news articles that taken from kompas.com and saved into text file will be used as main cluster data. The first article titled “Ini 3 Indikator yang Dipantau Pemerintah untuk Tetapkan Harga BBM” was taken from 13 June 2017. The second article titled “Pengusaha Tanggapi Target Pertumbuhan Ekonomi Pemerintah” was taken from 13 June 2017. The third article titled “Subsidi Akan Disalurkan ke 2.500 Desa yang Belum Teraliri Listrik” was taken from 13 June 2017. While the online news articles will be taken from Kompas Ekonomi (bisniskeuangan.kompas.com) on 8 July 2017.

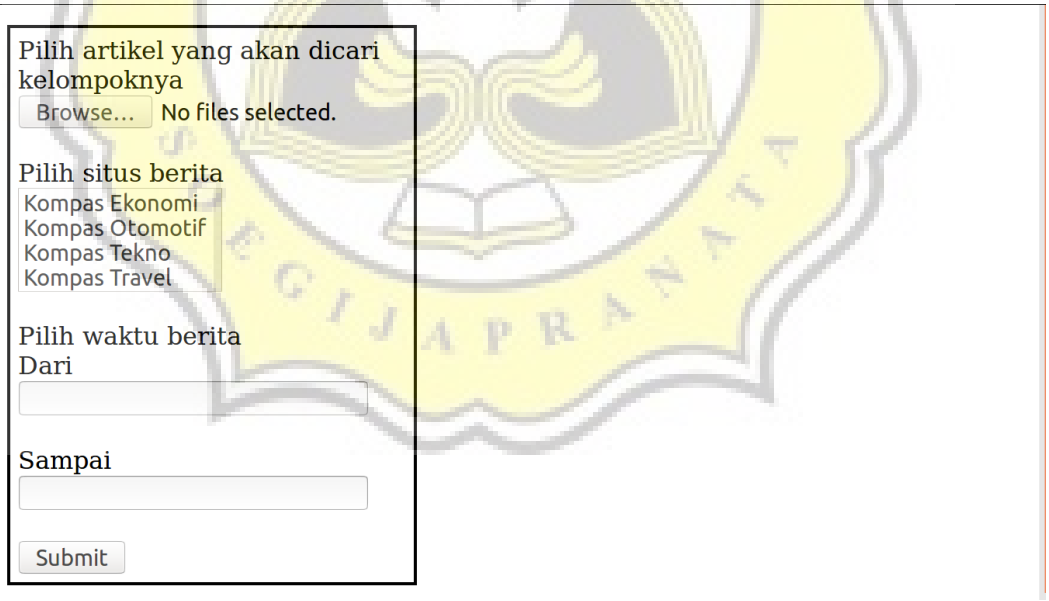

Illustration 5.1: Form Interface

After the form submitted, program will upload user articles and take online news articles based on the chosen source and date.

Illustration 5.2: Filled Form

Table 5.1: Euclidean Distance

	Cluster 1 - Ini 3 Indikator yang Dipantau Pemerintah untuk Tetapkan Harga BBM	**Cluster 2 - Pengusaha Tanggapi Target Pertumbuhan Ekonomi Pemerintah**	**Cluster 3 - Subsidi Akan Disalurkan ke 2.500 Desa yang Belum Teraliri Listrik**
Pastikan Tak Ada Kenaikan, PLN Justru Sebut Ada Penurunan Tarif Dasar Listrik	5.70608928447	6.0170898845406	3.022547215867
Tarif Kereta Ekonomi Bersubsidi Tak Jadi Naik	4.4076431691777	3.5334856593513	6.5743711799849
20 Tahun Setelah Krisis Finansial, Asia Lebih Tahan Banting	4.3798213997176	5.7249495646169	7.2782073515272
Pemerintah Janji Tak Ada Kenaikan Tarif Listrik hingga Akhir Tahun	4.2715794048639	4.7707900990283	3.7662120548307
PLN Klaim	7.6467495902449	7.8080135524837	4.5747783556214

Penyesuaian Tarif Listrik Tekan Laju Inflasi			
19 Produsen Teken Kontrak Penyaluran Biodiesel hingga Oktober 2017	6.1879787423667	4.4887837983777	7.8247299902809
Tahun Ajaran Baru, Pegadaian Bidik Transaksi Rp 11 Triliun	3.9597945764425	5.2164335611517	6.872172019009
PLN Bantah Cabut Subsidi Listrik Secara Menyeluruh	7.0489744674162	7.3311316696067	3.398092209739
Kelola Perhutanan Sosial, Petani Tetap Bisa Dapat KUR	3.8244663887051	4.3939809593255	6.403454709949
Cari Potensi Lokal, Bekraf Gulirkan Program IKKON 2017 di Lima Kota	4.3051426984061	3.5307354997127	6.6856590688795
PLN: Banyak Pihak Salah Persepsi Kebijakan Subsidi Tepat Sasaran	7.726906618212	8.0550759779834	4.2241417876051
Klaim Kecelakaan Mudik Lebaran 2017 Turun sekitar 50 Persen	2.440238316386	3.8188167340833	5.7789680915176
Pemerintah Tak Mau Terbuka soal Negosiasi dengan Freeport	3.5741649722648	2.951866699564	6.0501818945747
Pindahkan Ibu Kota, Pemerintah Siapkan Ratusan Ribu Hektar Lahan di 3 Lokasi	3.2313780119613	4.0639954359039	6.0805368287274

Kuartal I 2017, Laba Bank DBS Indonesia Tumbuh 72 Persen	3.2435184531188	4.6029885626	6.4804898249648
Ekspansi Bisnis, Sari Roti Akan Terbitkan Saham Baru	1.8477260611413	3.1405330571685	5.4126369012084

In the table above is the euclidean distance between user news article and online news article. The euclidean distance is compared between each user article. For example like the online news titled “Pastikan Tak Ada Kenaikan, PLN Justru Sebut Ada Penurunan Tarif Dasar Listrik”. Its euclidean distance between first, second and third user article compared and the nearest distance means it belong in the same cluster with the user article. In this case is cluster 3.

Illustration 5.3: Clustering Result Displayed on index.php

Table 5.2: Clustering Result

Cluster 1 - Ini 3 Indikator yang Dipantau Pemerintah untuk Tetapkan Harga BBM	**Cluster 2 - Pengusaha Tanggapi Target Pertumbuhan Ekonomi Pemerintah**	**Cluster 3 - Subsidi Akan Disalurkan ke 2.500 Desa yang Belum Teraliri Listrik**
20 Tahun Setelah Krisis Finansial, Asia Lebih Tahan Banting	Tarif Kereta Ekonomi Bersubsidi Tak Jadi Naik	Pastikan Tak Ada Kenaikan, PLN Justru Sebut Ada Penurunan Tarif Dasar Listrik
Tahun Ajaran Baru,	19 Produsen Teken Kontrak	Pemerintah Janji Tak Ada

Pegadaian Bidik Transaksi Rp 11 Triliun	Penyaluran Biodiesel hingga Oktober 2017	Kenaikan Tarif Listrik hingga Akhir Tahun
Kelola Perhutanan Sosial, Petani Tetap Bisa Dapat KUR	Cari Potensi Lokal, Bekraf Gulirkan Program IKKON 2017 di Lima Kota	PLN Klaim Penyesuaian Tarif Listrik Tekan Laju Inflasi
Klaim Kecelakaan Mudik Lebaran 2017 Turun sekitar 50 Persen	Pemerintah Tak Mau Terbuka soal Negosiasi dengan Freeport	PLN Bantah Cabut Subsidi Listrik Secara Menyeluruh
Pindahkan Ibu Kota, Pemerintah Siapkan Ratusan Ribu Hektar Lahan di 3 Lokasi		PLN: Banyak Pihak Salah Persepsi Kebijakan Subsidi Tepat Sasaran
Kuartal I 2017, Laba Bank DBS Indonesia Tumbuh 72 Persen		
Ekspansi Bisnis, Sari Roti Akan Terbitkan Saham Baru		

In the table above is the result of the news clustering. Online news article titled “Pastikan Tak Ada Kenaikan, PLN Justru Sebut Ada Penurunan Tarif Dasar Listrik” related/have the same topic with the user news article titled “Subsidi Akan Disalurkan ke 2.500 Desa yang Belum Teraliri Listrik.”