CHAPTER 6
CONCLUSION

Conclusions can be drawn from this project are :

1. To do the news article calculation with K-means, term weighting (TF-IDF) is used. The step of TF-IDF is count the term frequency in a document (TF), count the term occurrences from all the document (DF), count the inverse from DF (IDF), and multiply TF and IDF.
2. To make TF-IDF calculation more accurate, text preprocessing is used. Text preprocessing consists of tokenization, stopword removal, and stemming. Tokenization used to separate all the words. Stopword removal used to remove the less meaningful words. Stemming used to convert affixes word into root word.
3. To calculate the distance between online news data with user news data used euclidean distance. Euclidean distance is functionate to count the distance between data point.

Although there are still many shortcomings like the time it takes to scrapping/get the news from kompas.com website with cURL is quite long, when the internet connection is unstable, the program not work properly; the time taken for stemming process is much longer than when it skipped; sometimes there is a user news article that have more online news article data in the same cluster compared to other user news article; and there are some cases where the cluster will be empty, usually because there is no related news in the choosen date. In further research, program can be developed so it still can running well even with the unstable internet connection or improving stemming process speed.