Collecting and Preprocessing Data Text in Social Media

Oktavian Abraham Lantang[a]

[a] University of Debrecen
oktavian_lantang@unsrat.ac.id

Abstract

Nowadays social media has become one of the largest data warehouses in the world, every day millions of texts are being entered into existing applications. Unfortunately, the data is not reused and becomes useless. In this paper we collect the data as text that contained in social media by using Web Scraping. Furthermore, we perform several preprocessing steps such as; Case Folding, Cleansing, Stopword Removal, Convert Emoticon, Convert Negation and Tokenisation to generate semi-structured data. Finally, we suggest some future work related the data that generated by this research.

Keywords: Preprocessing Data, Social Media, Semi-structured Data, Web Scraping

MSC: 68P01

References

[1] Prado H. A., Ferneda A., Emerging Technologies of Text Mining (Techniques and Application), *Information Science Reference 2008*

[2] Schrenk, M., Webbots, spiders, and screen scrapers: a guide to developing Internet agents with PHP/CURL, *No Starch Press 2007*

[3] Vargiu, E., Urru Mirko., Exploiting web scraping in a collaborative filtering-based approach to web advertising, *Artificial Intelligence Research SCIEDU Vol 2 No 1, 2013*