Abstract

Developing corpora from social media content involves convoluted cleaning. In this paper we propose and implement the automation of corpora building for facilitating social media mining and analytics. This automation process incorporates: a) metadata extraction and structuring b) semantic cleaning with tagging and c) learning domain terms/entities. The implementation performs comprehensive cleaning including abbreviation and slang correction, phonetic matching using metaphorone algorithm, splitting joined words and identifying/learning entities. It identifies the entities, tags them and creates/updates a knowledgebase (KB) comprising of domain terms. The corpus thus constructed, facilitates multidimensional analysis and summarization. This proposed technique was tested with an experiment in which real world streaming tweets pertaining to Indian politics were collected, structured, cleaned and tagged. The results of the automation experiment can be stated as follows: a) the tweets although primarily in English, contained at times words from the regional languages. The algorithm does not recognize these words and they are construed as domain terms. An accuracy of 85.55% was achieved in identifying the correct domain terms and entities. b) The automation required
human feedback and intervention which progressively reduced and reached a figure of 18% with the update and enhancement of the KB. This paper assumes relevance because the implementation automates the entire process of collecting and cleaning the tweets and yields a corpus suitable for multi-faceted analysis.

References

12. Times of India 19th July 2015 “Governments have understood the potential of social media”, Arun.Dev@timesgroup.com
Index Terms

Computer Science

Information Sciences

Keywords

Corpora, Tweets, Social Media, Mining, Analytics, Knowledgebase