Spleetter: a PACT-based Twitter Parser and Analyzer

Matteo Lissandrini
University of Trento
ml@disi.unitn.eu

Davide Mottin
University of Trento
mottin@disi.unitn.eu

ABSTRACT

Twitter is the biggest micro-blogging system and users therein produce every day a 175 milion of short messages[1], namely tweets. The ability to perform analyses fast on the tweets is not only useful, but also vital to retrieve news and real-time statistics. However, since the number of different statistics to compute on the same data is big (e.g., the hashtag trend over the time, the number of tweets per user), there is the need of performing several operations at the same time. On one hand we can consider a Map-Reduce system in order to split the operations into different machines running several map and reduce steps. On the other hand, map and reduce paradigm may not suffice. We propose to solve this problem with PACT flow, that is semantically reacher map-reduce like-system. Our solution is both simple and highly modular, being composed by simple operations that can be easily parallelized. We show how to compute a set of interesting statistics out of 22 milion tweets downloaded in a period of two weeks, showing the polarity of each tweet and filtering the data to produce interesting analyses.

Keywords

Twitter, Stratosphere, Data Analysis, Sentiment Analysis

1. INTRODUCTION

The ability to perform various statistical analyses with twitter has been attracted many research in the last few years [2, 3] and used to influence politics [4] and advertising [1]. This important trend must be taken into account, but the huge amount of information per day cannot be processed by a single machine. To this end, we propose a schema based on PACT paradigm implemented in Stratosphere, that is a enhanced map-reduce system able, to mix second order functions. Our goal is to conduct an analysis of a big dataset, automatically downloaded using the Twitter streaming APIs.

[1] http://www.mediabistro.com/alltwitter/twitter-stats_b32050

- **Dictionary Cleansing**
- **Polarity extraction**

2. DATA

We downloaded 33 milion of US tweets from the tweetstream using the streaming APIs[2] on a period of two weeks from December 18 2012 to January 18 2013. A summary of the main characteristic of the dataset is represented in table 1. The number of users is 16 milion, with an average of 2 tweets per user. The number of hashtags is one order of magnitude less than the number of tweets meaning that users often talk about same topics. From this dataset we generated three datasets 100K-TWEETS, 1M-TWEETS, 10M-TWEETS with 100k,1 milion and 10 milion tweets respectively, in order to test performance with various sizes. We also downloaded a dictionary of 213k English words to filter the meanngful tweets from those irrelevant or containing only symbols. To perform the sentiment analysis and (i.e., to assign a positive or negative polarity to the tweets) we used the senti-strength[3] library.

Period	2012-12-18,2013-01-18
Tweets	33774428
Users	16099129
Hashtags	1194691
Max tweets per user	2380

Table 1: Characteristics of the dataset used in the experiments

3. SOLUTION

A PACT flow is

3.1 Data cleaning

3.2 Compute statistics

4. EXPERIMENTAL EVALUATION

5. CONCLUSIONS

6. REFERENCES

[1] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 65–74. ACM, 2011.

[2] https://dev.twitter.com/docs/streaming-apis
[3] http://sentistrength.wlv.ac.uk/

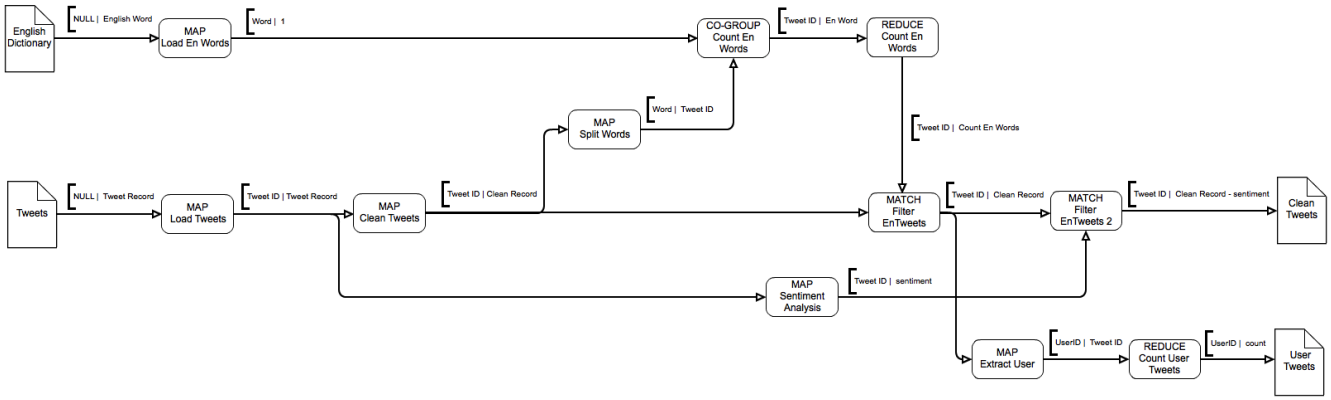

Figure 1: Data cleaning PACT

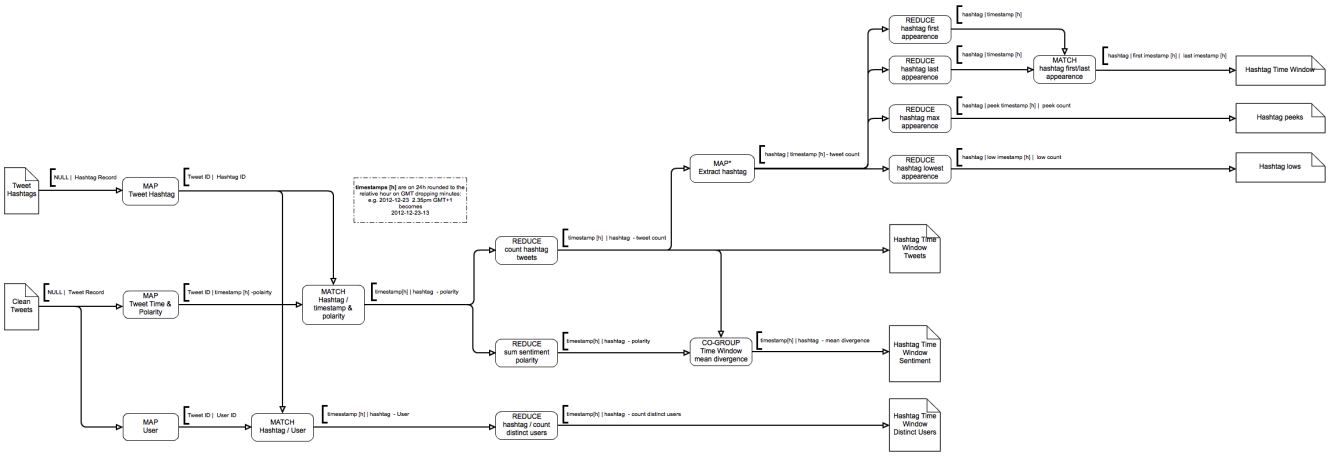

Figure 2: Compute statistics PACT

[2] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis. Discovering geographical topics in the twitter stream. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 769–778, New York, NY, USA, 2012. ACM.

[3] J. Lehmann, B. Gonçalves, J. J. Ramasco, and C. Cattuto. Dynamical classes of collective attention in twitter. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 251–260, New York, NY, USA, 2012. ACM.

[4] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the fourth international aaai conference on weblogs and social media*, pages 178–185, 2010.