Twitter User Recommender for Topics using Graphical Database

GROUP 7

Anton Lindqvist	Anna Lindelöf	Thiago Lobo	Helder Martins	Casper Renman
antoli@kth.se	anna@kth.se	thiago@kth.se	helderm@kth.se	casperr@kth.se

Abstract

1 Introduction

2 Background

3 Related work

4 Method

4.1 The Neo4j database

4.2 Crawling Twitter

4.3 Parsing tweets and extracting topics

The goal of the project is to recommend users given topics. In order to recommend a user, the user needs to be associated with the topics the user talks about. Therefore the users tweets are parsed and the topics of the tweets are extracted. The topics are extracted by parsing the freetext of the tweets and extracting the nouns and adjectives. The choice of extracting nouns and adjectives was an empiric decision made by the group.

Extracting topics from tweets is done using the Natural Language Toolkit (NLTK) [1] which provides interfaces in Python for things like classification, tokenization and stemming.

4.3.1 Cleaning tweets

A tweet can contain hyperlinks, hashtags, mentions and other symbols. These are removed in order to properly parse the text of the tweet. Specifically, words starting with *#, @, & or http* are ignored. A few other words that commonly occur in a tweet were also ignored as they would not contribute to the cause. These are *don't, i'll, retweet and rt.*

4.3.2 Extracting nouns

The nouns (topics) are extracted by performing the following actions, provided by NLTK:

1. Lowercase all letters and tokenize the text into separate tokens
2. Remove words that are shorter than three characters (This was also a decision made by the group)
3. For each word, remove ignored symbols and words starting with a ignored symbol
4. Part of Speech-tag [2] the words
5. Pick the words that are tagged as `NN` (noun) or `JJ` (adjective)
6. Stem the words and return the result which is a list of words

4.4 Ranking, PageRank and tf-idf

4.5 Graphical user interface

5 Experimental results

5.1 Ranking algorithms

5.1.1 Only PageRank

resulting list/table

5.1.2 Only tf-idf

resulting list/table

Small reflection (relevance feedback). What do we think? What should alpha be?

5.1.3 Combination

Try 3 different alphas and show list/table

5.2 Evaluation

Summary of what alpha should be and why.

6 Evaluation of the result

7 Summary and Conclusions

References

[1] Steven Bird. Nltk: the natural language toolkit. pages 69–72, 2006.

[2] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. 12:44–49, 1994.