Who to follow on Twitter

Group 7: POSTER 1

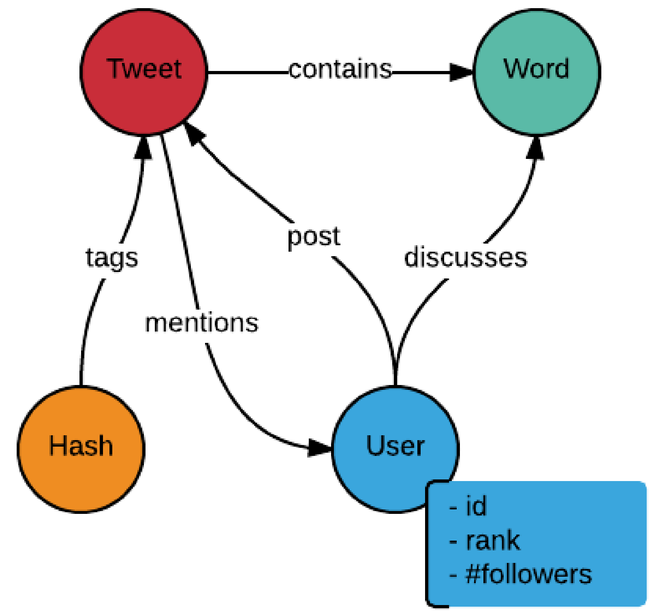
Problem description

The objective of this paper is to implement a web application of a Twitter user recommender system. The application should, given a topic, recommend Twitter users to follow. The resulting engine uses tf-idf in combination with PageRank to recommend users.

Conclusion

The conclusion is that it is feasible to create a Twitter user recommender engine using tf-idf and PageRank. Twitter data is suitable to be represented as a graph. Looking at the results, using appropriate parameters for weighing tf-idf and PageRank yields a result that is satisfying. Measuring the extent of the satisfaction is subjective, but for the group of this project it is deemed that the result is satisfactory enough for the given task.

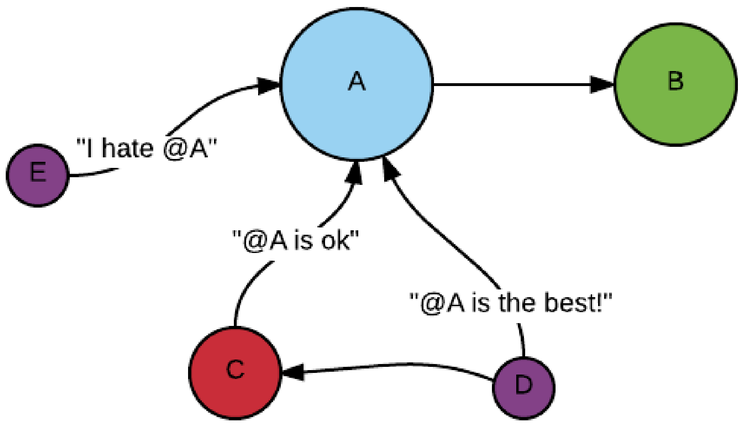
The database



- Number of users: 160.562
- Number of tweets: 1.021.876
- Number of words: 54.576
- Number of edges: 13.009.796

PageRank

Nodes are users. Edges are mentions. User mentions are the links between the users (documents).



tf-idf

tf: The term frequency is the number of times a term appears in a users tweets.
df: The document frequency is the number of users that have talked about the term.

tf-idf: The *cosine-similarity* is used to compute how close the query is to each of the documents.

A link between a *User* node and a *Word* node maps directly to a tf-idf score.

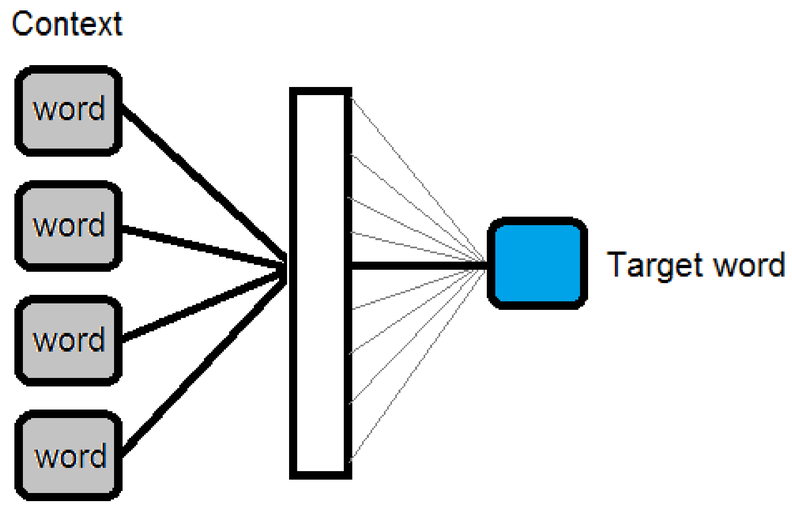
Final Score

After retrieving the sets of *PageRank* and *tf-idf* scores for all users that discuss any query term, they are normalized to zero-average and unitary variance so that they can be mixed together by means of a parameter, α. Finally both terms are combined, and the resulting score can be seen in Equation, where s_t is the tf-idf score for user u and query q and s_p the PageRank score of u.

$$s(u,q) = \alpha * s_t(u,q) + (1-\alpha) * s_p(u) \quad (1)$$

Word2vec - Generating Synonyms

Synonyms of the words in the query are generated using word embedding. A neural network model is trained where each word is represented by a real vector, mapping each word to a multidimensional vector space. Through training, word which have similar contexts, and therefore similar vector representations, are mapped close together in the vector space.



For example:
"president" ⇒ *presidency, presidents, chairmen, presidential, governor, chairman*
"liberal" ⇒ *conservative, liberals, liberalism, political, populist, moderate*