Final Report Proposal

Name: Guangyu Lin & Huihuang Zheng
Proposed Topic: Trust Filters - Identity & Seperation
Instructor: Ray Mooney
Course Name: CS 388 Natural Language Processing

Introduction: Your social network information can reveal a lot. In this project, we will implement part of a fused trust filter model to analyze the social network, such as Twitter, WordPress, Wiki, Instagram and etc. Our work is mainly inspired by ”You are what you tweet” [1] and from Center for Identity (Dtra Project), which reflects that we can use social media data to measure users' characteristics, including public health. The model is hierarchical, with the top-most level representing author-level attributes and the lower levels characterizing authors documents and words. As is customary in the topic modeling literature, we shall refer to a users posts on social media platforms as documents. The model assigns topic and category distributions to authors as well as to the documents written by the authors. Topics are identified by the algorithm, while categories are labels for words which are known in advance. We will handle the separation and identity parts of this fused trust filter model.

Approach: The process of the project is collecting raw text, tokenizing words, stemming lemmatization and removing punctuation normalize case, designing classifier like SVM and LDA Algorithm to retrieve related information and using NLP to do subject words predict. For NLP part, there are three core problems to solve,

1. Retrieve related information from corpus (influenza in our project but our model can handle other topics)
2. Identify the subject(the person) of a sentence
3. Identify the relationship between the subject and the author

For the first problem, we want to use LDA or SVM algorithms to do the document summarization and information filter. For the second problem, we want to classify if a user is talking about themselves or not. We would like to use Name Entity Recognizer to identify the subject of the sentence. For the third problem, which is more complicate, we want to seperate different Named Entity Recognizer and absorb the relationship between them. And their range could be from 0 to 1. We will use LDA from NLTK and SVM from Libsvm.

Evaluation: we have already collected raw text data from Twitter and WordPress. These data are the project from CiD. The data will be labeled so we can evaluate our model and compare with other algorithms.

References

[1] Paul, Michael J., and Mark Dredze. ”You are what you Tweet: Analyzing Twitter for public health.” ICWSM 20 (2011): 265-272.