CS388 NLP Project Proposal

Text-based Mining for Hidden Relations and Trending

C. Vic Hu
vic@cvhu.org

Ali Unwala
aliunwala@gmail.com

I. Motivation

Knowledge spillovers have been a well-studied topic in economics, which makes assumptions that researchers in close regions and similar domains tend to influence each other. In this project, we will focus on relationships between US patents in terms of word usages and writing styles, and hopefully to utilize that knowledge to formalize hidden topics and trend prediction.

II. Objective

We want to distinguish features from US patents using the text of the patent. We are interested in being able to guess the date a patent was published, the author of the patent, and classification of the patent using its text. Ultimately,

III. Hypothesis

We believe that the patents text holds clues as to when the patent was published, author, and how it was classified.

Based on the authorship prediction model presented in [authorship paper], we assume that similar patterns could be formulated in the context of US patents. In addition, we assume hidden topics exist and can be discovered over some patterns, and that their causal relationships can be established in a chronological order.

IV. Method

Using " to be filled" we would like to extract information from the patent. We will use an subset of the US patent database as our treebank. We will use part of our treebank for training and part of it for testing. This way we are able to evaluate our accuracy concretely. (Since authorship/publication date/ classification are all known for all patents)

V. Future Work

If time permits we would like to expand on this work to try and make an HMM for inventions. Where Inventions are the "jars" and patents are the "marbles." And links between jars will tell us how patents cluster to one another. This will take domain knowledge and not be easily evaluable.

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.