Project Proposal of Blog Authorship Detection

Context

With the development of the internet, more and more people love to put their ideas and thoughts online as blogs. Often, people love to read blogs that they might be interested in. One quick way for someone to find blogs is to check the author's features. If the machine learning technique of NLP is able to automatically identify the author features by analyzing texts of a given blog with existing other blogs and the author information, it will be a practical way to tackle this issue.

Problem Identification

An NLP analysis model will be built to identify which industry an author might belong to if provided with the content of a blog.

Data sources

The Blog Authorship Corpus consists of the collected posts of 19,320 bloggers gathered from blogger.com in August 2004. The corpus incorporates a total of 681,288 posts and over 140 million words - or approximately 35 posts and 7250 words per person.

Each blog is presented as a separate file, the name of which indicates a blogger id# and the blogger's self-provided gender, age, industry and astrological sign. (All are labeled for gender and age but for many, industry and/or sign is marked as unknown.)

All bloggers included in the corpus fall into one of three age groups:

- 8240 "10s" blogs (ages 13-17),
- 8086 "20s" blogs(ages 23-27)
- 2994 "30s" blogs (ages 33-47).

For each age group there are an equal number of male and female bloggers.

Each blog in the corpus includes at least 200 occurrences of common English words. All formatting has been stripped with two exceptions. Individual posts within a single blogger are separated by the date of the following post and links within a post are denoted by the label urllink.

Constraints

There are 251015 blogs or 36.84% of the total blogs don't have identification for the industry the author belongs to. The numbers of blogs for the rest of the different industries are unbalanced. For example, there are 153903 blogs written by students , and only 280 written by people from maritime.

Approaches

We will use the Natural Language Toolkit (nltk) to analyze the content of blogs, and use the TF-IDF method for the frequencies of words. We will then use models such as Logistic Regression, Decision Tree, SVM, KNNs to classify the blogs.

Criteria for success

We will use accuracy_score, f1_score, average_precision_score and recall_score from sklearn.metrics to evaluate the model behavior.