Author profiling

Bulat Nasrulin

Innopolis University, Russia

Introduction

The goal of this assignment is to build a system for author's gender identification using covered NLP techniques.

As a natural extension of the assignment I named the problem as author's profile identification. Additionally, I identify the author's age.

Problem formulation:

- With a given train, test data build a model to identify author's text by his/her posts in social networks (Facebook, Livejournal or generally blogs).

Datasets

For our task I use two datasets:

1. The Blog Authorship Corpus with 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 [1].
   (Available here)
2. Blog corpus annotated with gender [2].
   (Available here)

Gender identification

First, we start with a author's gender identification on second dataset. This is relatively small dataset, having only 3231 documents, with total corpus of 66259 unique words.

First of all, the simplest approach of random gender guessing gives **46%**, since the problem is just a binary classification.

Models with CountVectorizer and TfIdfVectorizer from scikit-learn library:

1. SVM - score 0.653
2. Naive Bayes - score 0.647
3. RandomForest - score 0.623

Using n-grams has either no effect on score or small insignificant effect.

Models with Doc2Vec:

Tfidf represent the words as a features, but to identify the context of the word. The simplest way to do is is to use word embeddings: the words nearby.

The extension of Word2Vec approach is presented in gensim library, called Doc2Vec, that converts whole document into a vector.

1. SVM - score 0.71
2. RandomForest - score 0.61

One of the problems here is the increased number of parameters that can be tuned. Moreover, tuning and training increased compared to a simple tfidf approach.

Deep learning models:

Surprisingly, but the models with CNN and LSTM doesn't show great improvement on the score, moreover, the learning time increased from 10 min to 7 hours.

Score: 0.65-0.69

Age identification

To extend this assignment from simple author's gender identification to more general author's profiling I added author's age identification.

Naturally, age identification is viewed as an regression problem. To simplify this problem we divide all ages into three categories as described in [1]:

1. "10s" - 8240 blogs (ages 13-17)
2. "20s" - 8086 blogs (ages 23-27)
3. "30s" - 2994 blogs (ages 33-47).

Due to dissymmetry in data, a baseline solution is to predict the most common value - '10s' we get score around 0.42.

Models with CountVectorizer and TfIdfVectorizer from scikit-learn library:

1. SVM - score 0.498
2. Naive Bayes - score 0.48
3. RandomForest - score 0.49

Using n-grams has either no effect on score or small insignificant effect.

Models with Doc2Vec:

1. SVM - score 0.52
2. RandomForest - score 0.48

Deep learning models:

Similarly, deep learning models don't give significant improvement. Moreover training took 1 day.

Future work

Author profiling only from a text of posts is still a hard task. To get more accuracy we need to work more on feature selection. For example, in work [2] authors suggested 300 hand-engineered features like usage of specific slang words: oomph, ugh etc. In work [1] authors get similar score by using doc2vec, they increased the accuracy to 80 by adding POS tags as features.

Reference

1. J. Schler, M. Koppel, S. Argamon and J. Pennebaker (2006). Effects of Age and Gender on Blogging in Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs

2. Arjun Mukherjee and Bing Liu. "Improving Gender Classification of Blog Authors." Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-10). Oct. 9-11, 2010, MIT, Massachusetts, USA.