Text Gender Classification

Weiming Li

Abstract

Males and females usually have different style of talking and writing. It is reasonable to think that their words can reflect their gender, such as “my husband” or “my wife”. I would use data from twitter as examples to show how the gender influence text. My goal is to determine the gender of user from a random one tweet. I choose Naïve-Bayes classifier in NLTK to train the data provided by Kaggle. I get an accuracy result with 55%. I found that males and females tend to use different words and lengths of sentence.

Introduction

The linguistic styles between males and females has been study for decades. Trudgill found that people in different gender may have different speech speed (Trudgill, 1972). Informal writing (Anthony Mulac, Lisa B. Studley, Sheridan Blau, 1990) and books (Lai, 2009) are also study by previous work.

Twitter is known as one of most famous micro-blogging sites. One of the things that is missing from Twitter is an indication of a person’s gender. The tweets are kind of different with the text we use in daily life, characters like “@” or “#” are use frequently. It would be interest to determine the user gender given a tweet.

User gender analysis may have useful applications, such as machine translation. Some languages employ different grammatical structures for depending on the gender of the user, whereas in English no such thing exists. The result can also work for other online forum, not only twitter.

The Natural Language Toolkit (NLTK) is a suite of libraries and programs for symbolic and programs for symbolic and statistical natural language processing (NLP) for English written in the Python programming language. The Naïve-Bayes classifier implanted in NLTK is good for NLP

Data

Kaggle is an online community of data scientists and machine learners, owned by Google, Inc. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges. I would use data from Kaggle to train my model.

The data set was originally used to train a CrowdFlower AI gender predictor (Using machine learning to predict gender, 2015). The data was collected in a small time around Oct 26th ,2015 with 20000 tweets. I divide 10% as test set and 90% train set.

The labels of gender include three type: male, female, or brand. Since not all users are individual, some twitters are managed by multiple people, like twitter used by companies, which I called brand twitter.

The original data contains multiple features like user name, color tweet count or retweet count. I would only consider tweets as feature and abandon the other features here.

Approach

I choose Naïve-Bayes classifier to train the data. At first, I try to use SVM to train but that was time consuming and we have three kinds of genders here. Neural network may be good at NLP, but it is slow to converge with so much features. Finally, Naïve-Bayes classifier is good at handling numerous distinct features and distinct results.

The feature model utilized in this project is the appearance of certain word. The frequency and the length of test may also influence the accuracy. Bigram works with similar result but increase the running time, and I already have enough features. There many be too much for the collection of all the words, so I only select frequent word as necessary.

Experiments

Initially, I tried to check all the words in the corpus as feature of tweets. A feature may look like "contains {boyfriend}: true". The appearance did get an accuracy over 70% on training set. However, checking every word may easily be overfit, most words in the tweets only appear one time, which cause the test set result quiet poor. The words with long length appear really and slow down the training time.

Then I try to restrict the length of words and use only frequent words to train the data. The length of sentence also added as a feature since some points that males prefer shorter sentence (Trudgill, 1972). The frequency to reference other (counting of "#" also differ in gender. I add all those features and the result gets better performance at test set.

Result

The result of using Naïve-Bayes gets an accuracy around 55%. Some technique like bigram can increase the accuracy by 1%, with the cost of much longer time to converge.

The results also give interesting information about words used by males and females. It is confirmed that males prefer shorter tweets and females prefer longer. The

preference in pronouns are also differ. Female tends to use of first-person pronoun, such as ”I”, “my”, “our” and “we”. On the contrary, males prefer second-person and third person. The most frequently word “the” and “and” cannot tell much different in gender.

Conclusion

The text gender analysis is still hard to decide by a single text, since for most sentence males and females may write the same way. Also, brand user is kind of difficult to confirm, since the true user of twitter are still humans. Even though, most brand tweets have more “official” and most the content is to claim the truth.

References

.

Anthony Mulac, Lisa B. Studley, Sheridan Blau. (1990). The gender-linked language effect in primary and secondary students' impromptu essays. *Sex Roles, 23*, 439. doi:https://doi.org/10.1007/BF00289762

Argamon, S., Koppel, M., Fine, J. and Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. *Text, 23*(3), 321-346. doi:https://doi.org/10.1515/text.2003.014.

Lai, C.-Y. (2009). *Author Gender Analysis.* University of California.

Trudgill. (1972). Sex, covert prestige and linguistic change in the urban British English of Norwich. Language in Society. *Language in Society, 1*, 179-195. doi:https://doi.org/10.1017/S0047404500000488

Using machine learning to predict gender. (2015, Nov 6). Retrieved from figure eight: https://www.figure-eight.com/using-machine-learning-to-predict-gender/