Author Profiling in Twitter Variety and Gender classification.

Màster Big Data Analytics
Universitat Politècnica de València
2016-2017

Òscar Garibo i Orts

Author Profiling.

Author profiling consists of identifying some text's authors based on characteristics found in the text.

In this task Tweets written in Spanish language will be used as the object of study.

Base Line

Model based on BoW with the 1000 most frequent words.

Results, accuracy:

Gender: 0,6643

Variety: 0,7721

Python as a choice.

For our best convenience we decided to start coding from raw in Python.

We have processed the train and test datasets to store them in csv files so we can easily import them to pandas dataframes.

Further preprocessing has been performed. We have built separate train and test dataset with accents being removed.

Variety classification.

There are 7 different classes, which corrspond to 7 different countries where Spanish is the mother language. These classes are: Spain, Colombia, Perú, Argentina, Chile, México and Venezuela.

First approach:

- Corpus with accents.
- Spanish stop-words removed.
- Tf-Idf considering unigrams, bigrams and trigrams.
- SVM used for modeling.

Best result:

Tf-Idf with unigrams. Accuracy: 0.9121

Variety classification.

Second approach:

- Corpus with accents.
- Spanish stop-words removed.
- Tf-Idf considering unigrams, bigrams and trigrams.
- Corpus specific stop-words removal by setting Tf-Idf max_df parameter to 0.90.
- SVM used for modeling.

Best result:

- Tf-Idf with unigrams. Accuracy: 0.9128

Variety classification.

Third approach:

- Corpus without accents.
- Spanish stop-words removed.
- Tf-Idf considering unigrams.
- Corpus specific stop-words removal by setting Tf-Idf max_df parameter to 0.90.
- Multi-Layer Perceptron used for modeling.

Best result:

- Tf-Idf with unigrams. Accuracy: 0.9179

Variety classification.

Other (discarded) approaches:

- Corpus without accents.
- Spanish stop-words removed.
- Tf-Idf considering unigrams.
- Corpus specific stop-words removal by setting Tf-Idf max_df parameter to 0.90.
- Random Forest Classifier used for modeling.

Best result:

Tf-Idf with unigrams. Accuracy: 0.9043

Variety classification.

Other (discarded) approaches:

- Corpus without accents.
- Spanish stop-words removed.
- Tf-Idf considering unigrams.
- Corpus specific stop-words removal by setting Tf-Idf max_df parameter to 0.90.
- Multi-Layer Perceptron used for modeling.
- Mixes of additional features: POS tags, word count, emoji count, hashtag count, etc.

Best result:

Tf-Idf with unigrams. Accuracy: 0.9100

Variety classification.

Other (discarded) approaches:

- Corpus without accents.
- Spanish stop-words removed.
- SnowBall stemmer.
- Tf-Idf considering unigrams.
- Corpus specific stop-words removal by setting Tf-Idf max_df parameter to 0.90.
- Multi-Layer Perceptron used for modeling.

Best result:

Tf-Idf with unigrams. Accuracy: 0.9164

Gender classification.

There are 2 different classes, which corrspond to the gender of the tweet's author. Obviously these classes are: Male and Female.

First approach:

- Corpus with accents.
- Punctuation marks removed.
- Spanish stop-words removed.
- Tf-Idf considering unigrams, bigrams and trigrams.
- SVM used for modeling.

Best result:

- Tf-Idf with unigrams. Accuracy: 0.7468

Gender classification.

Second approach:

- Corpus with accents.
- Punctuation marks removed.
- Spanish stop-words removed.
- Tf-Idf considering unigrams, bigrams and trigrams.
- Corpus specific stop-words removal by setting Tf-Idf max_df parameter to 0.90.
- SVM used for modeling.

Best result:

Tf-Idf with unigrams. Accuracy: 0.7507

Gender classification.

Third approach:

Corpus with accents.
Punctuation marks removed.
Spanish stop-words removed.
Tf-Idf considering unigrams.
Corpus specific stop-words removal by setting Tf-Idf max_df parameter to 0.90.
Multi-Layer Perceptron used for modeling.

Best result:

Tf-Idf with unigrams. Accuracy: 0.7664

Gender classification.

Fourth approach:

Corpus with accents.
Punctuation marks kept.
Spanish stop-words removed.
Tf-Idf considering unigrams.
Corpus specific stop-words removal by setting Tf-Idf max_df parameter to 0.90.
Multi-Layer Perceptron used for modeling.
Part Of Speech included.

Best result:

Tf-Idf with unigrams. Accuracy: 0.7707

Gender classification.

Fifth approach:

Corpus with accents.
Punctuation marks kept.
Spanish stop-words removed.
Tf-Idf considering unigrams.
Corpus specific stop-words removal by setting Tf-Idf max_df parameter to 0.90.
Multi-Layer Perceptron used for modeling.
Part Of Speech included.
Word Count included.

Best result:

Tf-Idf with unigrams. Accuracy: 0.7743

Gender classification.

Other (discarded) approaches:

Corpus with accents.
Punctuation marks kept.
Spanish stop-words removed.
Tf-Idf considering unigrams.
Corpus specific stop-words removal by setting Tf-Idf max_df parameter to 0.90.
Multi-Layer Perceptron used for modeling.
Part Of Speech included.
Word Count included.
Twitter specific items count (hashtag, retweets, mentions, etc)

Best result:

Tf-Idf with unigrams, POS tags, emoji count, word count, hashtag count.
Accuracy: 0.7714

Conclussions (I)

Our model for Variety classification relies mostly in statistical features. The accents in the tweets just mean noise for our models, adding no usable information. Differences in vocabulary are enough for the model to classify the tweets.

Our model for Gender relies both in statistical features, such as Tf-Idf and word count, as in linguistic features, such us Part Of Speech tags.

Punctuation marks have revelaed to be an important factor to distinguish between man and women.

Conclussions (II)

BoW Baseline, accuracy:

- Variety: 0.7721
- Gender: 0.6643

Our best results, accuracy:

- Variety: 0.9179, 18.9% improvement.
- Gender: 0.7743, 16.6% improvement.

Future work.

We do consider that an interesting work can be done in exploding the URLs. We think we could get important features that could help in classifying either in Variety and Gender.

An URL can have a national domain (.es, .ar, .cl, etc) which would help for Variety. Some domains are known to be gender specific (with many exception, but we try to generalize). For example, vogue, enfemenino.com, could be female URLs. While playboy.com, sportive newspapers and other could be male related URL.

PAN 2017.

We would like to remark that our results, if considered in PAN 2017 context, would have given us the 10th best score in Variety classification, and 8th in Gender classification.

We kindly request the instructors to invite us to the next PAN edition, since this task has been challenging, hard but very enjoyable.