Learning from Data - Final project

Author Profiling

Mart Busger op Vollenbroek
S2174634

Introduction

The task we had to perform for the final project of Learning from Data was to develop a system that could profile the author of a tweet (e.g. distinguish the gender and age category from the author of a tweet).

Developing the system

Intro

The data used for training was separated into 4 different languages: English, Dutch, Italian and Spanish. The approach chosen for the problem is one where there is one system that should classify all the languages. The reason for this is mainly because of time management issues.
The versions of the packages used for the project are as follows:
Python: *3.4.3*
Nltk: *3.1*
Scikit-learn: *0.17*

Reading in the data

For reading in the the training and test data, the fileRead() function is used. The function takes the setting (training / test) and language as parameters and is therefore used 8 times, twice for every language. If the function is used to read in training data, the truth.txt file is also checked to read in the gold-standard data. Dependent on the setting, the documents and its author are returned (in case of test data) or the documents, genders, age categories and its author are returned (in case of training data).

Feature selection

Several features have been selected to be used and some of them were thrown out because they influenced the scores badly instead of increasing them. In the first row of table 1 the initial results for gender classification for English can be found, using a tf-idf vectorizer, a multinomialNB classifier and no preprocessing whatsoever. The following rows add several features such as preprocessing, tokenizing and the use of bigrams.

Settings English gender classification	Precision	Recall	F-score
TfidfVectorizer MultinomialNB	0.60	0.58	0.58
TfidfVectorizer(preprocessing) MultinomialNB	0.59	0.57	0.57
TfidfVectorizer(using bigrams) MultinomialNB	0.58	0.56	0.56
TfidfVectorizer Support Vector Machine(linear kernel, C=1.0)	0.62	0.61	0.61

Table 1: Initial results

After seeing that most of the features I tried to add actually harmed the system performance, I tried a different classifier. Because the SVM seemed to work better, I continued testing using that classifier. The results from this are found in table 2.

Settings English gender & age classification	Precision	Recall	F-score	Precision	Recall	F-score
TfidfVectorizer Support Vector Machine (linear kernel, C=1.5)	0.62	0.61	0.62	0.59	0.60	0.59
TfidfVectorizer Support Vector Machine (rbf kernel, C=1.5)	0.19	0.43	0.26	0.24	0.49	0.32
TfidfVectorizer TfidfVectorizer (using bigrams) Support Vector Machine (linear kernel, C=1.5)	0.62	0.62	0.62	0.63	0.63	0.61
TfidfVectorizer TfidfVectorizer (using bigrams) CountVectorizer Support Vector Machine (linear kernel, C=1.5)	0.63	0.62	0.62	0.64	0.64	0.63
TfidfVectorizer TfidfVectorizer (using trigrams) CountVectorizer Support Vector Machine (linear kernel, C=1.5)	0.61	0.61	0.61	0.66	0.65	0.65

Table 2: Further results English gender & age classification

After combining several vector as features, the scores went up a little but they are still only at highest 6% above the baseline which is the most frequent class baseline of 56% because that percentage is the amount of tweets placed by women. The first three columns of precision, recall and f-score are still the results for English gender classification and the second second set of columns containing precision, recall and f-score are the results for the English age classification. While trigrams are performing worse for gender classification, they actually work well for age classification. Another thing worth noting is that the multi class classification is actually performing better than the binary classification. Because of lack of time, no further feature selection has been done and these settings will be used on the test data.

Final results on training data

After training the system on the English data, several accuracy scores were calculated. These can be found in table 3:

Language	Gender accuracy	Gender precision	Gender recall	Gender f-score	Age accuracy	Age precision	Age recall	Age f-score
English	0.619	0.63	0.62	0.62	0.651	0.66	0.65	0.65
Dutch	0.584	0.58	0.58	0.58				
Italian	0.640	0.64	0.64	0.64				
Spanish	0.596	0.63	0.60	0.61	0.436	0.50	0.44	0.43

Table 3: Final results on training data

For calculating these scores, the same gender classifier and age classifier were used for every language.

Truth files

Finally, using the classifiers obtained from the genderClassifier() and ageClassifier() functions the predictions are made in the makePredictions() function and are written to a truth.txt file with the makeTruthFile() function in the corresponding language directory with the test data.