CIS 530 Fall 2015 Project Report

Mao-Hsu Chen

December 16, 2015

1 Introduction

This project involves a semi-supervised authorship attribution task with some training data labeled as 1 if the author is Gina Kolata and as 0 otherwise. The binary classifier trained on the given data is then used to distinguish the authorship of new excerpts in the testing data. The key point here is to identify some *stylometric* features that separates Kolata's style of writing from those of other authors. As a single author can write articles about various topics, these personal stylistic traits should generally be topic-independent and remain consistent across different genres. Unlike what we did in Homework 2, where a bigram language model was used to predict the type of new excerpts, the content information is of little importance. What matters more for literary style lies in the contextual or structural information.

Among the previously-proposed potential style markers, lexical and character features such as word frequencies and character ngrams are relatively inexpensive in terms of computations whereas syntactic and semantic features often require robust and accurate NLP tools to perform the parsing, as summarized by Stamatatos (2009). With the above mentioned issues taken into account, a baseline SVM with the top-1000 most common words in the training data as features was first implemented and was later supplemented with other stylometric features. First of all, as reported in many works (e.g. Chung and Pennebaker, 2007), *function words*, although with little or no lexical meaning, are promising style indicators that reveal the links between language and personality. The list of the most frequent function words in the training set was thus considered as part of our feature space. Secondly, character-level measures such as character *n*-grams, on the one hand, can provide hints of not only lexical but also contextual information, and use of punctuation as well; on the other hand, they are more tolerant to noise than full word forms. In addition, frequencies of different part-of-speech may reflect one's writing style in that more adjectives may be used in romantic description while verbs and nouns may be preferred in rational reasoning and argument.

Stamatatos (2009) points out that the first dozens of most common words of a corpus are usually dominated by closed class words while open class words become majority after a few hundred words. Since all function words belong to the group of closed class words, one can expect a great overlap between the list of function words and the top-1000 most common words in the training data, and it would be interesting to find out how the number of non-function words included in the feature space affects the classification result. Meanwhile, it is uncertain as to how many *n*-grams should be included as feature and what number *n* should be. These three parameters would therefore be manipulated with the validation data.

When SVM is adopted in the system, one can expect that the more features are included the better the performance would be, which is clearly exemplified in our experimental results. The achieved high performance (over 90%) here only implies an integral success of all features used but fails to identify the contribution of individual feature set. Another thing to be noticed is that this project, as a practical case study, shows that there always exists a difference between the training and the testing data sets, especially when the latter is unlabeled, which is often the case in real-world application.

2 Method

2.1 Features

Four sets of features were utilized in this project, including function words, most frequent lexical words, character *n*-grams, and part-of-speech tags.

As some short function words are often included in the list of stop words, the file `stopwords.txt` from Homework 4 consisting of 593 tokens was adopted here as a source of function word list. A union set taken from the set of the stop words and the set of the top-1000 most frequent contained about 830 tokens (the exact number varied depending on the training set). For each term in the feature space, the logarithmic relative frequency was calculated and then weighted by its inverse document frequency. The resulting feature vector for each document, or excerpt here, was normalized to unit length with respect to the Euclidean norm.

The computation of character *n*-grams was computationally simplistic. A set of top-1000 most common *n*-grams was generated for $n = 3$, 4, and 5 respectively. Other things being equal, the comparison among the results of the three *n*-gram sets showed that 4-grams (average accuracy = 88.2936%) outperformed than 3-grams (average accuracy = 87.5011%) and 5-grams (average accuracy = 87.6413%); thus 4-grams were used throughout the rest of the experiments. Assume that the lengths of the character-level and the lexical features are independent, the same transformation of frequency vectors, idf weighting, and unit length normalization were applied to the 4-grams feature vector separately from the lexical word feature vector mentioned above.

The 12 POS tags formed the final set of features, which presumably was independent of the other two aforementioned feature sets. The values of the POS feature set denoted the percentage of each POS tag within an excerpt. The three sets of features were combined into the feature space to train the classifier.

2.2 Other improvements

K-fold cross-validation was used to avoid overfitting the training data. k = 9 was chosen such that the number of the validation data and that of the testing data could match.

2.3 Resources and tools

The package *sklearn.cross_validation* in *scikit-learn* was adopted for the *k*-fold cross-validation. The `libsvm` library was used for the implementation of a binary-class prediction SVM with a linear kernel. Stanford Part of Speech tagger was employed to perform the POS tagging, and instead of using the 45 tags in the PTB tagset as features, the parsed POStags were further mapped onto the 12 tags in the Google Universal tagset to reduce the noise of mislabeling by the POS tagger. The mapping was borrowed from `cis530/hw3/data/en-ptb.map`.

3 Final System

The team name for the leaderboard is *lingolingoling*.

With the capability of managing a large number of dimensions, support vector machine was ideal for an authorship attribution classifier when the number of stylometric features was unknown. The final system contained three separate feature sets: 1) a union of the stop words list and the top-1000 most frequent words, 2) the top-15,000 4-grams, and 3) the twelve Google Universal POS tags. The frequencies of the former two sets underwent logarithmic relativity transformation and were further weighted by their individual idf, and at last L2 normalization was applied to each of the two sets. The combination of the three feature sets formed the final feature space. Each excerpt in the training and testing data was represented with such a feature vector. The ones for the training data along with the known labels (1 for Kolata's work and 0 otherwise) were inputted to the SVM to build a binary-class classifier, which was later used to predict the labels for the testing data.

4 Experiments

4.1 Final submission system

The accuracy of the final system on the test set is 93.4621099554%, and the average accuracy of the training set using 9-fold cross validation is 93.8413863435%.

4.2 Other system alternatives

Accuracy	Train	Test	Parameter anipulation					
			norm	n-gram	# of n-grams	# of mfw	stop words	POS
other 1	86.6011%	-	L2	-	-	100	-	-
other 2	88.5412%	-	L2	-	-	1000	-	-
other 3	87.5011%	-	L2	3	1000	-	-	-
other 4	**88.2936%**	-	L2	**4**	1000	-	-	-
other 5	87.6413%	-	L2	5	1000	-	-	-
other 6	79.1381%	-	na	4	1000	1000	-	-
other 7	79.1877%	-	L1	4	1000	1000	-	-
1st sub.	**88.970%**	87.7415%	**L2**	4	1000	1000	-	-
other 8	88.7890%	-	L2	4	1000	-	v	-
other 9	89.0614%	-	L2	4	1000	500	v	-
other 10	**89.3586%**	-	L2	4	1000	**1000**	v	-
other 11	89.6805%	-	L2	4	1500	1000	v	-
other 12	90.5142%	-	L2	4	2000	1000	v	-
2nd sub.	91.3398%	90.9361%	L2	4	3000	1000	v	-
3rd sub.	-	93.7593%	L2	4	15000	1000	v	-
4th sub.	-	93.4621%	L2	4	20000	1000	v	-
5th sub.	-	94.4279%	L2	4	20000	1000	v	v
Final	93.8414%	93.4621%	L2	4	15000	1000	v	v

Table 1: Performance and parameter manipulation of all the experiments.

5 Discussion and Analysis

Table 1 shows the performance of all the experiments along with the parameter settings. Comparing the result of attempts "other 3, 4, and 5", one can conclude that 4-grams served as a better stylometric feature set than 3-grams or 5-grams. With the Euclidean norm (L2), the SVM classifier outperformed those either without normalization or with L1 normalization, as exemplified by the results of attempts 6 & 7 and 1st submission. The performance of the rest of the experiments revealed a general tendency that the SVM system would improve as the feature space enlarges.

References

Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556, 2009.

Cindy Chung and James W Pennebaker. The psychological functions of function words. *Social communication*, pages 343–359, 2007.