CIS 530 Fall 2015 Project

Instructor: Ani Nenkova
Head TA: Anne Cocos

Released: November 20, 2015
Due: 9:00am, December 16, 2015. No late days allowed.

Overview

For the final project you will develop a author recognition system, which recognizes the writing of an author among samples from a group of authors. Your task will be, given a piece of writing, to determine whether or not Gina Kolata was the author. You will develop a supervised binary classifier to make your predictions. The focus of the project is on feature engineering, so the challenge is to come up with features that distinguish Kolata and non-Kolata articles well.

Our evaluation metric will be overall accuracy in classifying the test set. The teams that achieve the highest accuracy will be awarded extra credit, so unlike your previous homework assignments, the goal here is to develop the best possible system for the task. The labels for the entire test dataset will not be released but you can submit **up to five** preliminary predictions and receive a score to check your progress. Your sixth submission is your final submission, and it should be based on the model you describe and evaluate in your final report. We will set up a leaderboard so that you can compare your results to the other teams and the baseline.

The baseline system is an SVM that uses the top-1000 most common words in the training data as features. You may build this system first and improve upon it, or start your own from scratch. For inspiration you may wish to read one or more of the following:

- Stamatatos, A survey of modern authorship attribution methods (2009)
- Diederich et al., Authorship attribution with support vector machines (2003)
- Segarra et al., Authorship attribution through function word adjacency networks (2014)

1 Data

The labeled data for the project are excerpts from New York Times articles. To compile the dataset we found 343 articles written by Gina Kolata; these are the positive examples for the classification task. Each of these was matched with the four most similar non-Kolata NYT articles; these are the negative examples. All of the articles were broken into excerpts of roughly 150-200 words and combined them into training and test sets.

There are a total of 12,113 excerpts in the training set. Excerpts from articles by Kolata are labeled 1 and excerpts from other authors are labeled 0. There are 1346 excerpts in the test set. We will not provide the labels for the test set, but you will submit your predictions in a specified format (see below) and receive your accuracy score. You will only be able to evaluate your performance on the test set five times before your final submission, so you will need to plan carefully which settings of your model you want to compare.

All project data are stored on `biglab` under `/home1/c/cis530/project/data`. The files include:

- `project_articles_train`, which contains the training excerpts and labels, tab-separated, one per line.
- `project_articles_test`, which contains the test excerpts without labels, one per line.

2 Evaluation

To evaluate your system, you will submit your predictions for each test excerpt. The predictions file should contain one line for each excerpt, consisting of the predicted label (1 for Kolata, 0 for other), *in the same order that the excerpts appear in* `project_articles_test`, as shown below:

```
1
0
0
1
0
...
```

The test set was randomly drawn, so the distribution of Kolata- and non-Kolata articles is roughly, but not exactly, the same in the test set as in the training set.
We will use accuracy – the number of correct predictions divided by the number of samples– to evaluate your predictions. The accuracy of your submission will be automatically updated on the class leaderboard. Updates will happen automatically every 30 minutes.

3 Project guidelines

- Projects can be done individually or in teams of two.
- You cannot manually label the data in the test set in order to get better performance.
- In your project report you should report the accuracy of the system submitted last to the leaderboard. You should submit the code that generated the final leaderboard results.
- The five best systems among all groups get extra credit for the final project. First place gets 25%, second gets 20%, third gets 15%, fourth gets 10%, and fifth gets 5%.

4 Your own system

Your task is the develop a supervised system to distinguish Kolata from non-Kolata articles. You can reuse code from earlier class assignments as you implement features for your classifier. You may also use any off-the-shelf tools (i.e. tokenizers, POS taggers, dependency parsers, etc) that you find useful. You should describe all tools you use for pre-processing in your report.

Note that we did not release a dedicated development set, so to avoid over-fitting you should perform model validation on your own using data within the training set. You can do this either by explicitly separating a validation set from the training set data, or using cross-validation.

Your final submission should include the code for the final result you submitted, which is also recorded on the leaderboard.

5 Report requirement

For the project report, please use the templates provided at the top of the project resource page. The project report should be no more than 3 pages long. It should include the following sections:

- **Introduction**. The general idea/method for your system: which features you chose to implement and why you expect them to be predictive.
- **Method**. What resources or tools you have used and how they are included in your model.
- **Final system**. Give a brief description of your final system.
- **Experiments**. The accuracy of your system and its variants on the test set. The final accuracy you report should be that for your last submission on the leaderboard. You should record any preliminary results before the final submission so that you can include them in your writeup.
- **Discussion and Analysis**. Discuss your results and any interesting comparisons/conclusions that you can draw from them.

Each of the Background, Methods and Analysis results will be scored separately for grading. The background section will be receive full scores if the proposed approach logically matches the task. The methods section will receive full scores if you use at least two types of text analysis drawn from the topics covered in class. The discussion will receive full grade if your provide an insight about why the system does or does not work well and what would be additional ways for improvement you could have explored if you had unlimited time for the task.

6 Submission

6.1 Register your team

Before getting results on the leaderboard, you need to submit your team name. **Every team member has to submit an individual registration**. The deadline for registration will be **11:59PM, December 3**. You simply need to submit a plain text file called `team.txt` that contains a single line with your team name.

```
% turnin -c cis530 -p proj_groups team.txt
```

6.2 Submit to the leaderboard

You are allowed to submit five preliminary test predictions to the leaderboard. The fifth submission will be your final one – the one you need to submit code for and describe in your final report.

To get your prediction accuracy on the test set, submit a single file called `test.txt` to the leaderboard.

```
% turnin -c cis530 -p leaderboard test.txt
```

The format of the file `test.txt` is described in Section 2. Your accuracy on the test set will appear at the leaderboard after the next scheduled update. The leaderboard will be automatically updated every 30 minutes, so you may not see your score immediately. The leaderboard will be hosted here: `http://www.cis.upenn.edu/~cis530/leaderboard.htm`. **The leaderboard will close at 11:59PM, December 15**.

6.3 Submit your code and write-up

Please submit the code for your system and a short README describing your submitted files and how to run your code. Your final submission is due by **09:00AM, December 16**.

```
% turnin -c cis530 -p project project_yourpennkey.zip
```

Only one group member needs to make the final submission. Your submitted .zip should include the following:

- Code for your system
- A short README describing how to run your code
- Your final project report (.pdf format)

6.4 Milestones

- December 3rd: Register your team
- December 4th: Leaderboard opens
- December 15th: Leaderboard closes
- December 16th: Final submission due
- December 16th-17th: Project presentations