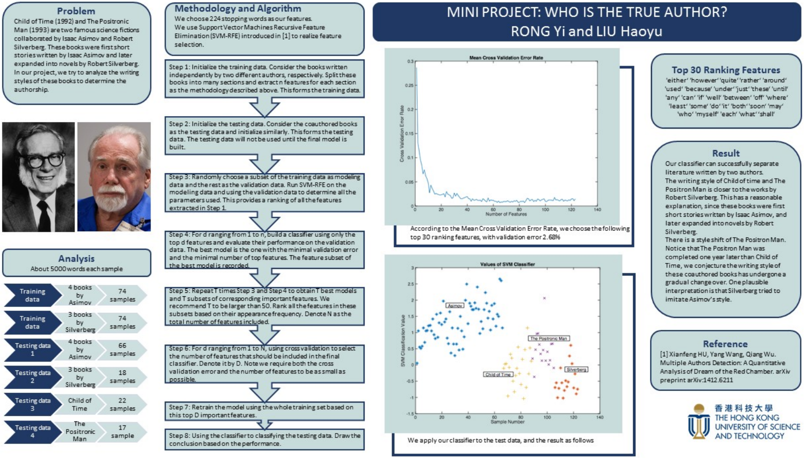
Mini Project
Who is the true author?

RONG Yi and LIU Haoyu

April 10, 2017 (Monday)

Vote for us



MINI PROJECT: WHO IS THE TRUE AUTHOR?

RONG Yi and LIU Haoyu

Problem

Child of Time (1992) and The Positronic Man (1993) are two famous science fictions collaborated by Isaac Asimov and Robert Silverberg. These books were first short stories written by Isaac Asimov and later expanded into novels by Robert Silverberg. In our project, we try to analyze the writing styles of these books to determine the authorship.

Analysis

About 5000 words each sample

Training data	4 books by Asimov	74 samples
Training data	3 books by Silverberg	74 samples
Testing data 1	4 books by Asimov	66 samples
Testing data 2	3 books by Silverberg	18 samples
Testing data 3	Child of Time	22 samples
Testing data 4	The Positronic Man	17 sample

Methodology and Algorithm

We choose 224 stopping words as our features.

We use Support Vector Machines Recursive Feature Elimination (SVM-RFE) introduced in [1] to realize feature selection.

Step 1: Initialize the training data. Consider the books written independently by two different authors, respectively. Split these books into many sections and extract n features for each section as the methodology described above. This forms the training data.

Step 2: Initialize the testing data. Consider the coauthored books as the testing data and initialize similarly. This forms the testing data. The testing data will not be used until the final model is built.

Step 3: Randomly choose a subset of the training data as modeling data and the rest as the validation data. Run SVM-RFE on the modeling data and using the validation data to determine all the parameters used. This provides a ranking of all the features extracted in Step 1.

Step 4: For d ranging from 1 to n, build a classifier using only the top d features and evaluate their performance on the validation data. The best model is the one with the minimal validation error and the minimal number of top features. The feature subset of the best model is recorded.

Step 5: Repeat T times Step 3 and Step 4 to obtain T best models and T subsets of corresponding important features. We recommend T to be larger than 50. Rank all the features in these subsets based on their appearance frequency. Denote N as the total number of features included.

Step 6: For d ranging from 1 to N, using cross validation to select the number of features that should be included in the final classifier. Denote it by D. Note we require both the cross validation error and the number of features to be as small as possible.

Step 7: Retrain the model using the whole training set based on this top D important features.

Step 8: Using the classifier to classifying the testing data. Draw the conclusion based on the performance.

According to the Mean Cross Validation Error Rate, we choose the following top 30 ranking features, with validation error 2.68%

We apply our classifier to the test data, and the result as follows

Top 30 Ranking Features

'either' 'however' 'quite' 'rather' 'around' 'used' 'because' 'under' 'just' 'these' 'until' 'any' 'can' 'if' 'well' 'between' 'off' 'where' 'least' 'some' 'do' 'it' 'both' 'soon' 'may' 'who' 'myself' 'each' 'what' 'shall'

Result

Our classifier can successfully separate literature written by two authors.

The writing style of Child of time and The Positron Man is closer to the works by Robert Silverberg. This has a reasonable explanation, since these books were first short stories written by Isaac Asimov, and later expanded into novels by Robert Silverberg.

There is a style shift of The Positron Man. Notice that The Positron Man was completed one year later than Child of Time, we conjecture the writing style of these coauthored books has undergone a gradual change over. One plausible interpretation is that Silverberg tried to imitate Asimov's style.

Reference

[1] Xianfeng HU, Yang Wang, Qiang Wu. Multiple Authors Detection: A Quantitative Analysis of Dream of the Red Chamber. arXiv preprint arXiv:1412.6211

香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

Figure: Our poster

Some of you may have heard of...

Figure: Bicentennial Man (You will see the name again.)

Most of you may have heard of...

Figure: I, Robot

Isaac Asimov: Our hero today

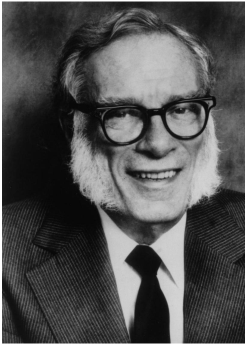


Figure: Isaac Asimov

*Wikipedia: Asimov was a prolific writer, and wrote or edited more than **500** books.*

500 books?!

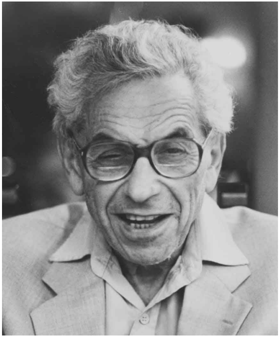


Figure: Paul Erdos

Wikipedia: Erdos published around 1500 mathematical papers during his lifetime.

Robert Silverberg: The other hero

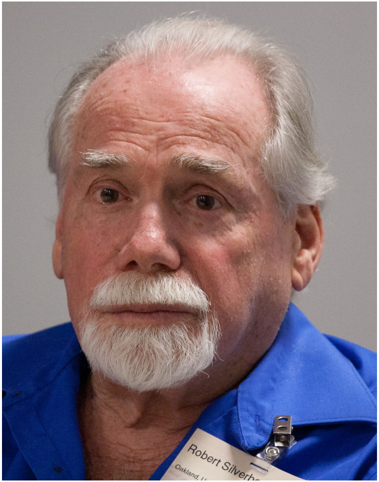
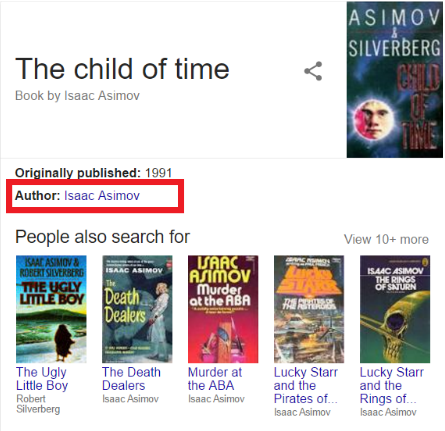


Figure: Robert Silverberg

Child of Time: Our story starts from here



The child of time

Book by Isaac Asimov

Originally published: 1991

Author: Isaac Asimov

People also search for

View 10+ more

The Ugly Little Boy
Robert Silverberg

The Death Dealers
Isaac Asimov

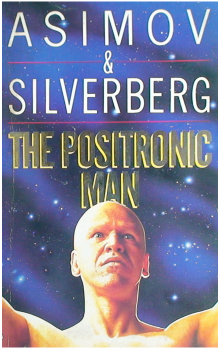
Murder at the ABA
Isaac Asimov

Lucky Starr and the Pirates of...
Isaac Asimov

Lucky Starr and the Rings of...
Isaac Asimov

Figure: Asimov or Silverberg? (Remember: Originally published in 1991.)

The Positronic Man: Our story continues

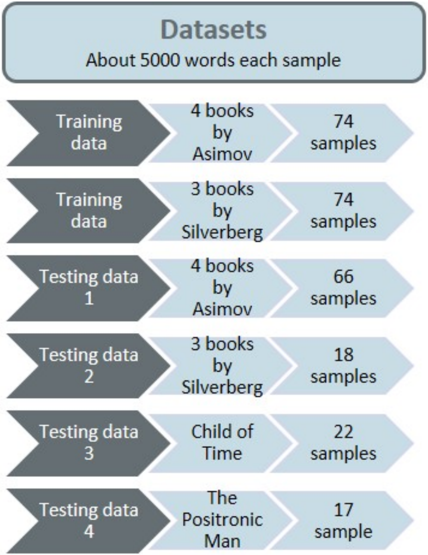


Wikipedia: The Positronic Man is a 1992 novel by Isaac Asimov and Robert Silverberg, based on Asimov's novelette ***The Bicentennial Man****. The film Bicentennial Man, starring Robin Williams, was based both on the original story and the novel.*

Methodology

- Wording and phrasing reflect writing habits
- Feature selection with support vector machine recursive feature elimination (SVM-RFE)
- Train the model using books written independently by two different authors
- Classify the testing datasets and draw the conclusion based on the performance.
- Similar idea: Dream of the Red Chamber

Datasets



Datasets

About 5000 words each sample

Training data	4 books by Asimov	74 samples
Training data	3 books by Silverberg	74 samples
Testing data 1	4 books by Asimov	66 samples
Testing data 2	3 books by Silverberg	18 samples
Testing data 3	Child of Time	22 samples
Testing data 4	The Positronic Man	17 sample

Top 30 Ranking Features

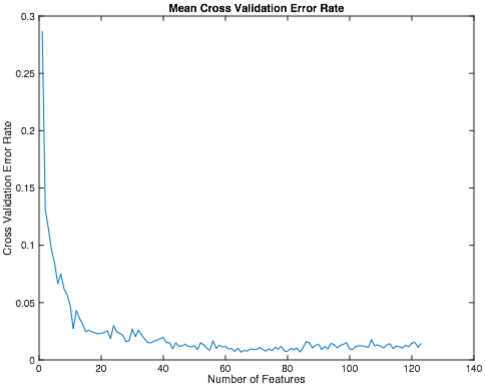


Figure: According to the Mean Cross Validation Error Rate, we choose the following top 30 ranking features, with validation error 2.68%.

Top 30 Ranking Features

either, however, quite, rather, around, used, because, under, just, these, until, any, can, if, well, between, off, where, least, some, do, it, both, soon, may, who, myself, each, what, shall

Results

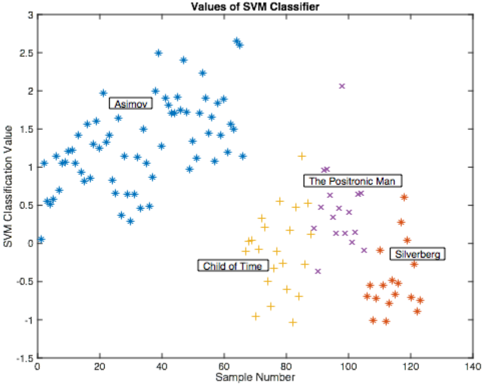


Figure: We apply our classifier to the test data, and the result as above.

Results

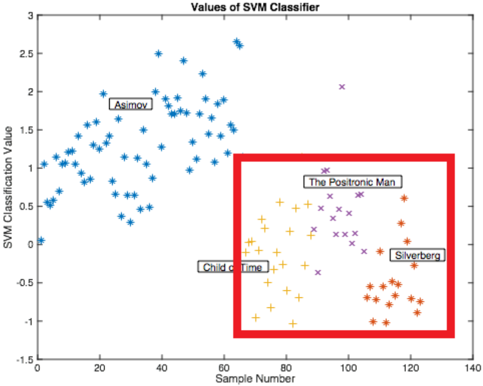


Figure: The writing style of Child of Time and The Positron Man is closer to the works by Robert Silverberg.

Results

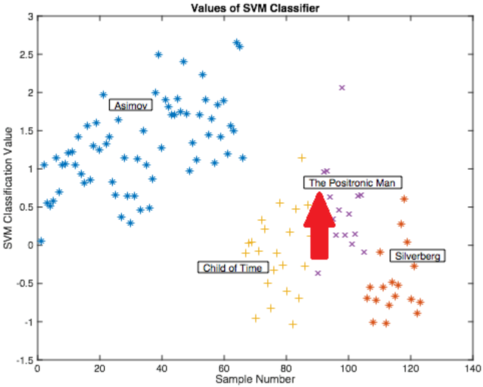
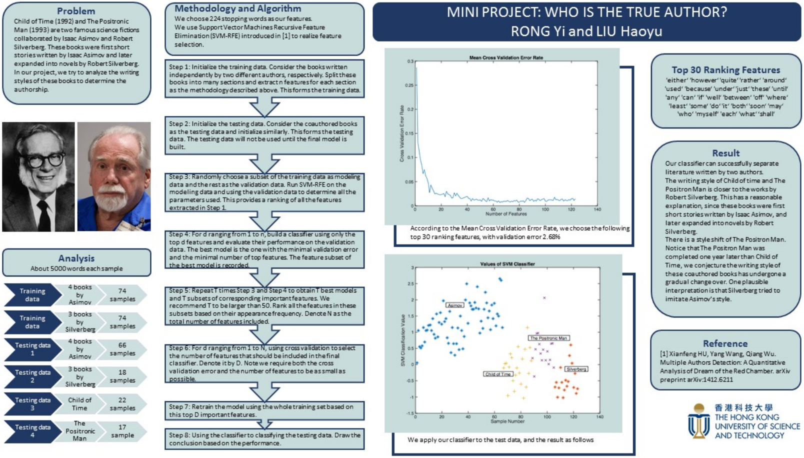


Figure: There is a style shift of The Positron Man.

Thanks



MINI PROJECT: WHO IS THE TRUE AUTHOR?

RONG Yi and LIU Haoyu

Problem

Child of Time (1992) and The Positronic Man (1993) are two famous science fictions collaborated by Isaac Asimov and Robert Silverberg. These books were first short stories written by Isaac Asimov and later expanded into novels by Robert Silverberg. In our project, we try to analyze the writing styles of these books to determine the authorship.

Analysis

About 5000 words each sample

Training data	4 books by Asimov	74 samples
Training data	3 books by Silverberg	74 samples
Testing data 1	4 books by Asimov	66 samples
Testing data 2	3 books by Silverberg	18 samples
Testing data 3	Child of Time	22 samples
Testing data 4	The Positronic Man	17 sample

Methodology and Algorithm

We choose 224 stopping words as our features.
We use Support Vector Machines Recursive Feature Elimination (SVM-RFE) introduced in [1] to realize feature selection.

Step 1: Initialize the training data. Consider the books written independently by two different authors, respectively. Split these books into many sections and extract n features for each section as the methodology described above. This forms the training data.

Step 2: Initialize the testing data. Consider the coauthored books as the testing data and initialize similarly. This forms the testing data. The testing data will not be used until the final model is built.

Step 3: Randomly choose a subset of the training data as modeling data and the rest as the validation data. Run SVM-RFE on the modeling data and using the validation data to determine all the parameters used. This provides a ranking of all the features extracted in Step 1.

Step 4: For d ranging from 1 to n, build a classifier using only the top d features and evaluate their performance on the validation data. The best model is the one with the minimal validation error and the minimal number of top features. The feature subset of the best model is recorded.

Step 5: Repeat T times Step 3 and Step 4 to obtain T best models and T subsets of corresponding important features. We recommend T to be larger than 50. Rank all the features in these subsets based on their appearance frequency. Denote N as the total number of features included.

Step 6: For d ranging from 1 to N, using cross validation to select the number of features that should be included in the final classifier. Denote it by D. Note we require both the cross validation error and the number of features to be as small as possible.

Step 7: Retrain the model using the whole training set based on this top D important features.

Step 8: Using the classifier to classifying the testing data. Draw the conclusion based on the performance.

According to the Mean Cross Validation Error Rate, we choose the following top 30 ranking features, with validation error 2.68%

We apply our classifier to the test data, and the result as follows

Top 30 Ranking Features

'either' 'however' 'quite' 'rather' 'around' 'used' 'because' 'under' 'just' 'these' 'until' 'any' 'can' 'if' 'well' 'between' 'off' 'where' 'least' 'some' 'do' 'it' 'both' 'soon' 'may' 'who' 'myself' 'each' 'what' 'shall'

Result

Our classifier can successfully separate literature written by two authors.
The writing style of Child of time and The Positron Man is closer to the works by Robert Silverberg. This has a reasonable explanation, since these books were first short stories written by Isaac Asimov, and later expanded into novels by Robert Silverberg.
There is a style shift of The Positron Man. Notice that The Positron Man was completed one year later than Child of Time, we conjecture the writing style of these coauthored books has undergone a gradual change over. One plausible interpretation is that Silverberg tried to imitate Asimov's style.

Reference

[1] Xianfeng HU, Yang Wang, Qiang Wu. Multiple Authors Detection: A Quantitative Analysis of Dream of the Red Chamber. arXiv preprint arXiv:1412.6211

香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

Reference

Xianfeng HU, Yang Wang, Qiang Wu. Multiple Authors Detection: A Quantitative Analysis of Dream of the Red Chamber. arXiv:1412.6211.