Dickens, Jane, and Austen

Using NLP to Classify Authors by Diction

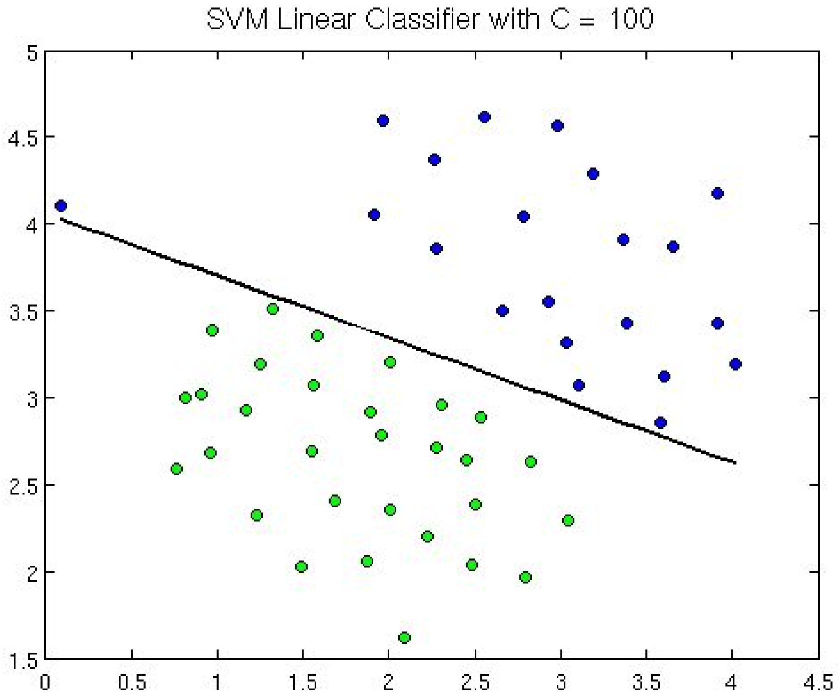
by Jason Hortsch

Why?

- Identifying unknown authors (Shakespeare)
- Helping to aid fan fiction or other writing based on mimicry
- Get to a point where an AI could write a book?

How?

- Used a technique called an ‘SVM’
- Turned text into numbers, & had it split into similar groups
- Example:
  - [the dog ran]
  - **[22, 3, 55]**
  - **[18, 7, 63]**
  - **--------------**
  - [5, 98, 16]
  - [6, 87, 13]

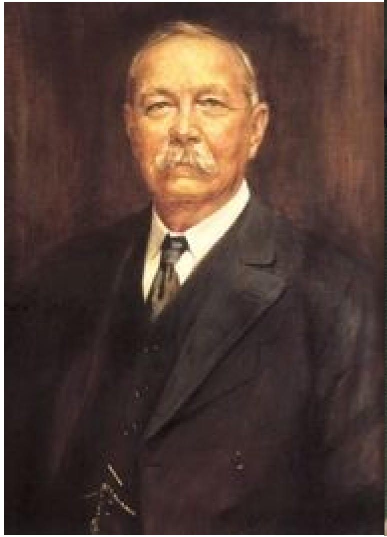


Process

- Started with just two authors

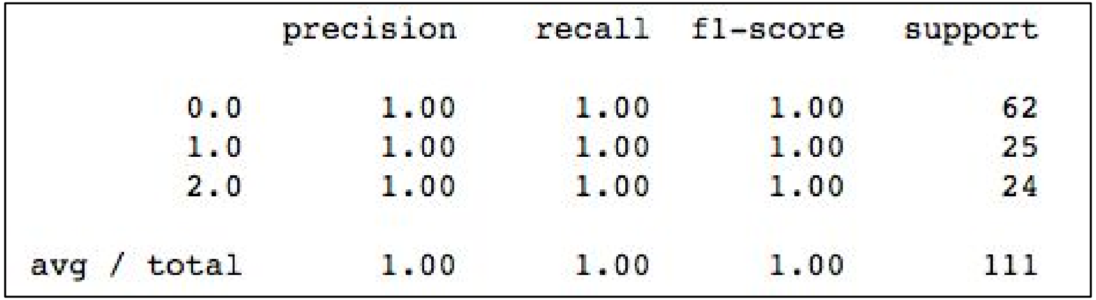
- Used 5000 word chunks, then used real chapters

- Went up to three authors



Results

- Perfection!



```
             precision    recall  f1-score   support

        0.0       1.00      1.00      1.00        62
        1.0       1.00      1.00      1.00        25
        2.0       1.00      1.00      1.00        24

avg / total       1.00      1.00      1.00       111
```

- Not a good thing
- Probably split on character names

Moving Forward

- Probably not focus on diction so much

- Turn more towards structural tendencies:
  - Sentence length
  - Paragraph length
  - Chapter length
  - Vocabulary difficulty
  - Part of speech use
  - Part of speech pairs

- Make it easy to use and accessible