Author Identification

Uras Mutlu & E. Burak Dündar

2017

Boğaziçi University
Introduction

● Problem:
  ○ Text -----> Author
  ○ Gender, age
  ○ Text to Text similarity
● Features: word frequencies, characters
● Plays of Shakespeare, 19th century
● “Federalist Papers”
  ○ The most influential work
● Stylometric features, until the 1990s
Features

- Character features ? ^ ; –
  - Character n-grams
- Lexical features
  - Bag of words, word n-grams, frequencies of words, functional words
- Syntactic features
  - Syntactic patterns for capturing style
- Semantic features
  - Meaning of a text
- Application-specific features
  - Authors on same theme, different keywords
Methods

● Bayesian approach

  ○ \[ P(v|a_1, \ldots, a_n) = \frac{P(v) \cdot P(a_1, \ldots, a_n|v)}{P(a_1, \ldots, a_n)} \]

● Compression approach:

  ○ Similarity = \( C(\text{text+unseenText}) - C(\text{text}) \)

● SVM outperforms Decision Tree and Neural Networks

  ○ 70–90% accuracy

● Convolutional Neural Network

  ○ State-of-the-art
Methods - State of the art

- \[ n \times k \text{ representation of sentence with static and non-static channels} \]
- Convolutional layer with multiple filter widths and feature maps
- Max-over-time pooling
- Fully connected layer with dropout and softmax output

wait for the video and do n't rent it
EXAMPLE USE-CASES

- **Intelligence**
  - Authorship identification for online communication messages of terrorist organizations.

- **Crime investigation**
  - Verifying the authenticity of digital evidence. Such as e-mails, suicide notes, electronic journals.

- **Cyber-crimes**
  - Identifying the developers who wrote malicious software by capturing their coding style.
Tools and Systems

- Turnitin, 1997
  - online plagiarism prevention system
- Online authorship attribution system
  - AICBT: Canadian R&D company
  - Supports 2 Authors
  - NLTK and Sci-kit Python Libraries
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Emails</th>
<th>IMDb</th>
<th>Blogs</th>
<th>Twitter</th>
<th>reddit</th>
<th>Ave</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>50</td>
<td>62</td>
<td>10</td>
<td>50</td>
<td>10</td>
</tr>
<tr>
<td>SVM+Stems</td>
<td>83.0</td>
<td>72.9</td>
<td>88.3</td>
<td>36.5</td>
<td>29.7</td>
<td>91.7</td>
</tr>
<tr>
<td>SCAP</td>
<td>83.1</td>
<td>69.0</td>
<td>94.8</td>
<td>48.6</td>
<td>41.6</td>
<td>91.3</td>
</tr>
<tr>
<td>Imposters</td>
<td>52.0</td>
<td>32.9</td>
<td>76.9</td>
<td>35.4</td>
<td>22.6</td>
<td>71.4</td>
</tr>
<tr>
<td>LDAH-S</td>
<td>82.0</td>
<td>39.1</td>
<td>72.0</td>
<td>52.5</td>
<td>18.3</td>
<td>90.0</td>
</tr>
<tr>
<td>CNN-word</td>
<td>89.7</td>
<td>82.5</td>
<td>84.3</td>
<td>59.0</td>
<td>43.0</td>
<td>96.2</td>
</tr>
<tr>
<td>CNN-word-word</td>
<td>90.3</td>
<td>81.0</td>
<td>82.0</td>
<td>58.8</td>
<td>41.4</td>
<td>95.7</td>
</tr>
<tr>
<td>CNN-char</td>
<td>93.1</td>
<td>88.1</td>
<td>91.7</td>
<td>59.7</td>
<td>48.1</td>
<td>97.5</td>
</tr>
<tr>
<td>CNN-word-char</td>
<td>90.3</td>
<td>84.9</td>
<td>90.2</td>
<td>61.2</td>
<td>49.4</td>
<td>97.2</td>
</tr>
<tr>
<td>CNN-word-word-char</td>
<td>90.1</td>
<td>83.8</td>
<td>88.4</td>
<td>61.2</td>
<td>47.0</td>
<td>95.9</td>
</tr>
</tbody>
</table>
**Conclusion**

- In last 15-20 years, with the advancements in information retrieval, machine learning, and natural language processing, there has been a lot of studies involving authorship identification.
- Most systems work well with closed sets and small number of authors, but there is still some progress needed for identifying the works of large number of authors.
- Better NLP tools that give less noisy data when analyzing texts syntactically and semantically can be very helpful to capture writing styles and achieve better accuracies.
References


Thank you!