Semi-supervised text classification using topic models

Miha Pavlinek

15th Workshop on “Software Engineering Education and Reverse Engineering”
Bohinj, Slovenia August 23rd – 30th 2015
Topic modeling

• Topic modeling is a process for finding semantically related clusters of words in text corpora – topics
  • Latent Semantic Analysis - LSA
  • Probabilistic LSA - pLSA
  • Latent Dirichlet Allocation (LDA)
    • Most effective generative statistical model, which combines semantically related concepts.
Intuition behind LDA

Documents

For many years, films about football were a bit of a joke. Having John Huston behind the camera meant that 1981’s Escape to Victory - with Pelé up front, Ipswich’s Russell Osman at the back and Sylvester Stallone in goal - was one of the best of them. But, as Huston’s son Danny admitted to me recently, his father had never watched a game in his life and didn’t even know how many players a football team should have on each side. Despite advances in special effects, the spontaneity and skill of a real game is still extremely hard to stage, yet films with football at their core have become increasingly respectable of late. In successive years, the Cannes Film Festival, under the auspices of its Olympic Lyonnais-supporting director Thierry Frémaux, has welcomed films about Zinedine Zidane, Diego Maradona and, only last month, Eric Cantona. Douglas Gordon and Philippe Parreno’s Zidane - a tightly focused study of the former French captain playing a match for Real Madrid - remains, for me, the best work about actually playing the game. Director Emir Kusturica’s documentary Maradona is more about the iconography of the Argentinian star. In Looking for Eric, Loach, working with regular photographer Barry Ackroyd, stays true to his usual realist vision of working Britain, yet also manages to make one of his least typical films in that for much of its first hour, it plays as a comedy. Steve Evets gives a gruff but tender performance as Eric Bishop, the struggling postie whose life has gradually fallen apart. Living with two stoned stepsons, he can’t summon the courage to reconnect with his ex-wife, Lily.
Intuition behind LDA
Intuition behind LDA

Topics

Topic 1
- football 0.04
- goal 0.03
- game 0.02
- play 0.02
- team 0.01

Topic 2
- movie 0.06
- actor 0.03
- play 0.01
- film 0.01
- camera 0.01

Topic 3
- father 0.02
- son 0.02
- life 0.02
- wife 0.01
- parent 0.01

Documents
Intuition behind LDA

Topics

Topic 1
- football 0.04
- goal 0.03
- game 0.02
- play 0.02
- team 0.01
...

Topic 2
- movie 0.06
- actor 0.03
- play 0.01
- film 0.01
- camera 0.01
...

Topic 3
- father 0.02
- son 0.02
- life 0.02
- wife 0.01
- parent 0.01
...

Documents

Topic proportions

...
Intuition behind LDA

Topics

Topic 1
- football 0.04
- goal 0.03
- game 0.02
- play 0.02
- team 0.01
...

Topic 2
- movie 0.06
- actor 0.03
- play 0.01
- film 0.01
- camera 0.01
...

Topic 3
- father 0.02
- son 0.02
- life 0.02
- wife 0.01
- parent 0.01
...

Documents

Latent (hidden) variables

Topic proportions
Intuition behind LDA

Topic 1
- football 0.04
- goal 0.03
- game 0.02
- play 0.02
- team 0.01
...

Topic 2
- movie 0.06
- actor 0.03
- play 0.01
- film 0.01
- camera 0.01
...

Topic 3
- father 0.02
- son 0.02
- life 0.02
- wife 0.01
- parent 0.01
...

Intuition behind LDA

For many years, films about football were a bit of a joke. Having John Huston behind the camera meant that 1981’s Escape to Victory - with Pelé up front, Ipswich’s Russell Osman at the back and Sylvester Stallone in goal - was one of the best of them. But, as Huston’s son Danny admitted to me recently, his father had never watched a game in his life and didn't even know how many players a football team should have on each side.

Despite advances in special effects, the spontaneity and skill of a real game is still extremely hard to stage, yet films with football at their core have become increasingly respectable of late. In successive years, the Cannes Film Festival, under the auspices of its Olympic Lyonnais-supporting director Thierry Frémaux, has welcomed films about Zinedine Zidane, Diego Maradona and, only last month, Eric Cantona.

Douglas Gordon and Philippe Parreno's Zidane - a tightly focused study of the former French captain playing a match for Real Madrid - remains, for me, the best work about actually playing the game. Director Emir Kusturica's documentary Maradona is more about the iconography of the Argentinian star.

In Looking for Eric, Loach, working with regular photographer Barry Ackroyd, stays true to his usual realist vision of working Britain, yet also manages to make one of his least typical films in that for much of its first hour, it plays as a comedy. Steve Evets gives a gruff but tender performance as Eric Bishop, the struggling postie whose life has gradually fallen apart. Living with two stoned stepsons, he can't summon the courage to reconnect with his ex-wife, Lily.
For many years, films about football were a bit of a joke. Having John Huston behind the camera meant that 1981’s Escape to Victory - with Pelé up front, Ipswich’s Russell Osman at the back and Sylvester Stallone in goal - was one of the best of them. But, as Huston’s son Danny admitted to me recently, his father had never watched a game in his life and didn’t even know how many players a football team should have on each side. Despite advances in special effects, the spontaneity and skill of a real game is still extremely hard to stage, yet films with football at their core have become increasingly respectable of late. In successive years, the Cannes Film Festival, under the auspices of its Olympic Lyonnais-supporting director Thierry Frémaux, has welcomed films about Zinedine Zidane, Diego Maradona and, only last month, Eric Cantona. Douglas Gordon and Philippe Parreno’s Zidane - a tightly focused study of the former French captain playing a match for Real Madrid - remains, for me, the best work about actually playing the game. Director Emir Kusturica’s documentary Maradona is more about the iconography of the Argentinian star.

In Looking for Eric, Loach, working with regular photographer Barry Ackroyd, stays true to his usual realist vision of working Britain, yet also manages to make one of his least typical films in that for much of its first hour, it plays as a comedy. Steve Evets gives a gruff but tender performance as Eric Bishop, the struggling postie whose life has gradually fallen apart. Living with two stoned stepsons, he can’t summon the courage to reconnect with his ex-wife, Lily.
## Example of real topics

<table>
<thead>
<tr>
<th>computer</th>
<th>health</th>
<th>espn</th>
<th>imdb</th>
<th>movie</th>
</tr>
</thead>
<tbody>
<tr>
<td>security</td>
<td>heart</td>
<td>sports</td>
<td>database</td>
<td>film</td>
</tr>
<tr>
<td>network</td>
<td>disease</td>
<td>news</td>
<td>title</td>
<td>actor</td>
</tr>
<tr>
<td>virus</td>
<td>cholesterol</td>
<td>baseball</td>
<td>movie</td>
<td>reviews</td>
</tr>
<tr>
<td>spam</td>
<td>food</td>
<td>scores</td>
<td>celebs</td>
<td>episode</td>
</tr>
<tr>
<td>spyware</td>
<td>life</td>
<td>nba</td>
<td>internet</td>
<td>scripts</td>
</tr>
<tr>
<td>home</td>
<td>nutrition</td>
<td>stats</td>
<td>management</td>
<td>cinema</td>
</tr>
<tr>
<td>anti</td>
<td>conditions</td>
<td>game</td>
<td>spielberg</td>
<td>character</td>
</tr>
<tr>
<td>internet</td>
<td>living</td>
<td>basketball</td>
<td>search</td>
<td>dvd</td>
</tr>
<tr>
<td>users</td>
<td>medical</td>
<td>college</td>
<td>character</td>
<td>scene</td>
</tr>
<tr>
<td>guide</td>
<td>healthy</td>
<td>standings</td>
<td>festival</td>
<td>star</td>
</tr>
<tr>
<td>information</td>
<td>risk</td>
<td>team</td>
<td>award</td>
<td>action</td>
</tr>
<tr>
<td>email</td>
<td>tips</td>
<td>player</td>
<td>board</td>
<td>news</td>
</tr>
</tbody>
</table>
Semi-Supervised learning

• Between unsupervised and supervised
• Learning with labeled and unlabeled data
  • Labeled instances are difficult and expensive to obtain
  • Unlabeled data may be easy to collect
• How it works?
  • Similar distributions across labeled and unlabeled instances
Topic distributions for classes

- European union
- Environment
- Travelling
- Crime
- Sport
Self-training

• Semi-supervised algorithm
• Learning process uses its own predictions to teach itself
• Repeat:
  1. Train $f$ on labeled data set
  2. Use $f$ to predict labels for unlabeled data
  3. Unlabeled instances with most confident predictions are added to labeled data set
Self-training with topics

Train set

Test set
Self-training with topics

Train set

Test set

Unlabeled set

Labeled set
Self-training with topics

Train set

Test set

Unlabeled set

Labeled set
Self-training with topics
Self-training with topics

Centroids

<table>
<thead>
<tr>
<th></th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>...</th>
<th>Tn</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.34</td>
<td>0.01</td>
<td>0.04</td>
<td>...</td>
<td>0.13</td>
</tr>
<tr>
<td>2</td>
<td>0.01</td>
<td>0.46</td>
<td>0.11</td>
<td>...</td>
<td>0.01</td>
</tr>
<tr>
<td>3</td>
<td>0.14</td>
<td>0.04</td>
<td>0.42</td>
<td>...</td>
<td>0.21</td>
</tr>
</tbody>
</table>

Topic distributions for unlabeled instances

<table>
<thead>
<tr>
<th></th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>...</th>
<th>Tn</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.07</td>
<td>0.01</td>
<td>0.04</td>
<td>...</td>
<td>0.11</td>
</tr>
<tr>
<td>2</td>
<td>0.01</td>
<td>0.32</td>
<td>0.15</td>
<td>...</td>
<td>0.33</td>
</tr>
<tr>
<td>3</td>
<td>0.82</td>
<td>0.02</td>
<td>0.02</td>
<td>...</td>
<td>0.12</td>
</tr>
<tr>
<td>4</td>
<td>0.37</td>
<td>0.14</td>
<td>0.04</td>
<td>...</td>
<td>0.19</td>
</tr>
<tr>
<td>5</td>
<td>0.17</td>
<td>0.42</td>
<td>0.01</td>
<td>...</td>
<td>0.03</td>
</tr>
<tr>
<td>6</td>
<td>0.01</td>
<td>0.07</td>
<td>0.41</td>
<td>...</td>
<td>0.11</td>
</tr>
<tr>
<td>7</td>
<td>0.11</td>
<td>0.09</td>
<td>0.23</td>
<td>...</td>
<td>0.17</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Proposed model

1. Generate topic model on entire train set
2. Split train set to labeled and unlabeled instances
3. While similar instances exists
   a. Take topic distributions from labeled set and calculate centroids for each class
   b. Measure distance between unlabeled instances and centroids (cosine similarity)
   c. Label most similar instances from unlabeled set with class from closest centroid and move instances to labeled set
## Results

**RTV SLO (NBMN: 81.8, SVM: 84.95)**

<table>
<thead>
<tr>
<th>Labeled instances</th>
<th>Baseline result</th>
<th>Simple self-training</th>
<th>Our method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NBMN</td>
<td>SVM</td>
<td>NBMN</td>
</tr>
<tr>
<td>1%</td>
<td>58.3</td>
<td>36.64</td>
<td>67.59</td>
</tr>
<tr>
<td>5%</td>
<td>73.74</td>
<td>74.8</td>
<td>75.62</td>
</tr>
<tr>
<td>10%</td>
<td>76.5</td>
<td>77.65</td>
<td>76.4</td>
</tr>
</tbody>
</table>

**20 Newsgroups (NBMN: 79.77, SVM: 75.6)**

<table>
<thead>
<tr>
<th>Labeled instances</th>
<th>Baseline result</th>
<th>Simple self-training</th>
<th>Our method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NBMN</td>
<td>SVM</td>
<td>NBMN</td>
</tr>
<tr>
<td>1%</td>
<td>32.26</td>
<td>24.88</td>
<td>45.15</td>
</tr>
<tr>
<td>5%</td>
<td>59.09</td>
<td>43.36</td>
<td>65.27</td>
</tr>
<tr>
<td>10%</td>
<td>67.06</td>
<td>56.68</td>
<td>71.3</td>
</tr>
</tbody>
</table>
## Results

### Reuters R8 (NBMN: 90.64, SVM: 96.09)

<table>
<thead>
<tr>
<th>Labeled instances</th>
<th>Baseline result</th>
<th>Simple self-training</th>
<th>Our method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NBMN</td>
<td>SVM</td>
<td>NBMN</td>
</tr>
<tr>
<td>1%</td>
<td>84.11*</td>
<td>81.32</td>
<td>75.92</td>
</tr>
<tr>
<td>5%</td>
<td>88.54</td>
<td>90.13*</td>
<td>83.7</td>
</tr>
<tr>
<td>10%</td>
<td>90.96</td>
<td>93.76*</td>
<td>85.56</td>
</tr>
</tbody>
</table>

### Google snippets – short texts (NBMN: 81.36, SVM: 64.61)

<table>
<thead>
<tr>
<th>Labeled instances</th>
<th>Baseline result</th>
<th>Simple self-training</th>
<th>Our method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NBMN</td>
<td>SVM</td>
<td>NBMN</td>
</tr>
<tr>
<td>1%</td>
<td>36.27</td>
<td>29.65</td>
<td>77.89*</td>
</tr>
<tr>
<td>5%</td>
<td>61.09</td>
<td>57.15</td>
<td>80.79*</td>
</tr>
<tr>
<td>10%</td>
<td>71.71</td>
<td>63.16</td>
<td>82.02*</td>
</tr>
</tbody>
</table>
Conclusions

• Representation with topics yields good results in semi-supervised settings with few labeled instances

• In some cases our approach outperforms other methods

• Future work
  • Improve our model with testing different parameters
  • Implement model on real example
  • Include texts from third corpora (e.g. Wikipedia)
Thank you for your attention!

QUESTIONS? COMMENTS?

(miha.pavlinek@um.si)