Classification of Topics of Web Documents Using Fasttext’s Supervised Learning on Classes and Data from dmoz.org
And Active Learning Demo Shown at Night of Scientists

Vít Suchomel

Oct 8, 2019
NLP Seminar
Brno
Table of Contents

1. Topic Classification

2. FastText + Active Learning Demo
Motivation – Web corpora

- What is inside?
- Subcorpora for users’ needs
- My interset: Genres & Topics
  - Genres determined by style vs. topic determined by words $\Rightarrow$ should be easier
Defining Topics

- Top to bottom...Apriori definition: Wordnet, Wikipedia, web directories (dmoz.org, url blacklist.com, curlie.org)
- Bottom to top...Data driven: vector representations, Gensim topics
Web Directory dmoz.org

- Shut down on 2017-03-17
- We got the last directory
- Now curlie.org
Web Directory dmoz.org

- Multiple languages ⇒ English
- 14 level 1 topics
- Hundreds of level 2 topics
  - E.g. Arts|Movies, Society|History, Sports|Track and Field
- Directory depth: 1 to 10
  - E.g. Recreation|Theme Parks|Individual Parks
  - Business|Mining and Drilling|Tools and Equipment|Mining
  - Sports|Water Sports|Swimming and Diving|Regional|Europe|United Kingdom|England
  - Society|Issues|Warfare and Conflict|Specific Conflicts|War on Terrorism|News and Media|September 11, 2001|BBC News
Wget 49 million URLs
  - 532,095 pages
  - 1,375,816 sites

2,178,334,898 tokens in 3,797,798 docs after processing
Balanced Level 1 Topics

1220530 docs, 14 level 1 topics, single label
Level 1 counts:

98320  Arts
98694  Business
98259  Computers
53828  Games
98826  Health
45942  Home
44722  News
98673  Recreation
93176  Reference
97322  Regional
96994  Science
99378  Shopping
97399  Society
98997  Sports
Balanced Level 2 Topics

1648085 docs, 355 level 2 topics, single label (rarely multilabel)
Level 1 counts:

146497  Arts
277332  Business
119920  Computers
 43674   Games
124461  Health
 42186   Home
 40911   News
129144  Recreation
 41175   Reference
 79223   Regional
108254  Science
185554  Shopping
147733  Society
162021  Sports
Issue: Documents in Multiple Categories

Example document:
https://www.liveabout.com/love-and-romance-4145433

0012288-17 Arts|Bodyart|Articles
0020909-209 Arts|Directories
0022507-68 Arts|Genres|Horror
0452960-19 Health|Beauty|Advice
0535415-76 Recreation|Humor|Jokes|Tasteless
0573222-52 Recreation|Tobacco|Cigars
Solution: Documents in Multiple Categories

- 2% multiclass level 1 docs removed
- 2% level 2 docs with multiclass level 1 removed
- 1% level 2 docs with multiclass level 2 kept $\Rightarrow$ Multilabel
Data Split

- 1% test set
- 2% evaluation set
- 97% training set
FastText Is...

- Learn text representations and text classifiers
- Vector representation of words
- Mikolov, now Facebook Research
• New functions: test, test-label, quantize, nn, analogies
• “Newer” functions: being added to the Git repository: autotune
• DIY C++
Best F1 topic level 1 autotune:

ns/ws5/neg5  0.677525 after 13 trials
ns/ws5/neg10 0.680365 after 13 trials
ns/ws10/neg5 0.677678 after 13 trials
ns/ws10/neg10 0.683732 after 13 trials
ns/ws5/neg15 0.684625 after 16 trials
ns/ws10/neg15 0.680162 after 16 trials
Best Level 1 Autotune (~3000 CPU-hours)

Trial = 8
ws = 5
neg = 15
epoch = 50
lr = 0.139148
dim = 100
minCount = 5
wordNgrams = 1
minn = 3
maxn = 6
bucket = 5000000
dsub = 2
loss = ns

Progress: 3.391% Trials: 8 Best score: 0.687341 ETA: 0

currentScore = 0.688365
train took = 11980.6
Best Level 2 Autotune (~1000 CPU-hours)

Trial = 5
ws = 5
neg = 15
epoch = 50
lr = 0.44913
dim = 100
minCount = 5
wordNgrams = 1
minn = 3
maxn = 6
bucket = 2590739
dsub = 2
loss = ns
Progress: 2.927% Trials: 5 Best score: 0.574681 ETA: 698h55m33s
currentScore = 0.567162
train took = 15250
• More autotuning ⇒ better result?
• Same algorithm, more CPUs for autotune ⇒ competition winner?
Evaluation Test

DMOZ Level 1 Topics Supervised Classifier Test Part: Precision, Recall, F 0.5 For Increasing Label Probability Threshold
Setting the Label Probability Threshold

- High precision: 0.95
- Best F 0.5: 0.55 – Precision preferred at the cost of recall
<table>
<thead>
<tr>
<th>Category</th>
<th>F1</th>
<th>Precision</th>
<th>Recall</th>
<th>Relevant Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arts</td>
<td>0.439</td>
<td>0.917</td>
<td>0.288</td>
<td></td>
</tr>
<tr>
<td>Business</td>
<td>0.206</td>
<td>0.853</td>
<td>0.117</td>
<td></td>
</tr>
<tr>
<td>Computers</td>
<td>0.515</td>
<td>0.912</td>
<td>0.359</td>
<td></td>
</tr>
<tr>
<td>Games</td>
<td>0.648</td>
<td>0.905</td>
<td>0.505</td>
<td></td>
</tr>
<tr>
<td>Health</td>
<td>0.604</td>
<td>0.945</td>
<td>0.444</td>
<td></td>
</tr>
<tr>
<td>Home</td>
<td>0.436</td>
<td>0.856</td>
<td>0.293</td>
<td></td>
</tr>
<tr>
<td>News</td>
<td>0.457</td>
<td>0.894</td>
<td>0.307</td>
<td></td>
</tr>
<tr>
<td>Recreation</td>
<td>0.428</td>
<td>0.890</td>
<td>0.282</td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>0.396</td>
<td>0.914</td>
<td>0.252</td>
<td></td>
</tr>
<tr>
<td>Regional</td>
<td>0.424</td>
<td>0.908</td>
<td>0.276</td>
<td></td>
</tr>
<tr>
<td>Science</td>
<td>0.413</td>
<td>0.895</td>
<td>0.268</td>
<td></td>
</tr>
<tr>
<td>Shopping</td>
<td>0.394</td>
<td>0.880</td>
<td>0.254</td>
<td></td>
</tr>
<tr>
<td>Society</td>
<td>0.361</td>
<td>0.907</td>
<td>0.225</td>
<td></td>
</tr>
<tr>
<td>Sports</td>
<td>0.619</td>
<td>0.927</td>
<td>0.464</td>
<td></td>
</tr>
<tr>
<td>Category</td>
<td>F1</td>
<td>Precision</td>
<td>Recall</td>
<td>Description</td>
</tr>
<tr>
<td>-------------</td>
<td>------</td>
<td>-----------</td>
<td>--------</td>
<td>----------------</td>
</tr>
<tr>
<td>Arts</td>
<td>0.656</td>
<td>0.787</td>
<td>0.562</td>
<td></td>
</tr>
<tr>
<td>Business</td>
<td>0.505</td>
<td>0.710</td>
<td>0.392</td>
<td></td>
</tr>
<tr>
<td>Computers</td>
<td>0.722</td>
<td>0.804</td>
<td>0.655</td>
<td></td>
</tr>
<tr>
<td>Games</td>
<td>0.776</td>
<td>0.868</td>
<td>0.702</td>
<td></td>
</tr>
<tr>
<td>Health</td>
<td>0.784</td>
<td>0.884</td>
<td>0.705</td>
<td></td>
</tr>
<tr>
<td>Home</td>
<td>0.650</td>
<td>0.783</td>
<td>0.557</td>
<td></td>
</tr>
<tr>
<td>News</td>
<td>0.679</td>
<td>0.798</td>
<td>0.591</td>
<td></td>
</tr>
<tr>
<td>Recreation</td>
<td>0.641</td>
<td>0.798</td>
<td>0.536</td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>0.639</td>
<td>0.831</td>
<td>0.519</td>
<td></td>
</tr>
<tr>
<td>Regional</td>
<td>0.628</td>
<td>0.833</td>
<td>0.504</td>
<td></td>
</tr>
<tr>
<td>Science</td>
<td>0.645</td>
<td>0.798</td>
<td>0.541</td>
<td></td>
</tr>
<tr>
<td>Shopping</td>
<td>0.631</td>
<td>0.806</td>
<td>0.519</td>
<td></td>
</tr>
<tr>
<td>Society</td>
<td>0.580</td>
<td>0.775</td>
<td>0.464</td>
<td></td>
</tr>
<tr>
<td>Sports</td>
<td>0.770</td>
<td>0.873</td>
<td>0.689</td>
<td></td>
</tr>
</tbody>
</table>
173 random docs
Thr 0.95 => 14.9 % docs got a label
Thr 0.55 => 51.3 % docs got a label

Agreement "That is the topic" 43 115
Weak Agreement "That could be the topic" 6 25
Disagreement "That is not the topic" 10 33

Threshold 0.95 0.55
83 % 81 %
<table>
<thead>
<tr>
<th>Good</th>
<th>Bad</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>4</td>
<td>Arts</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>Business</td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>Computers</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>Games</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>Health</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>News</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>Recreation</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>Reference</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>Regional</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>Science</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Shopping</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
<td>Society</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>Sports</td>
</tr>
</tbody>
</table>
Why is this text labelled “Arts”?

echo "The Environmental Exposure Group is also part of the MRC-PHE Centre for Environment and Health. For information on the Centre please visit their website. MRC-PHE Centre for Environment and Health In the Chair: Professor Paul Elliott King's College London (Room TBC) Science Museum, South Kensington, London – first floor, entry via Cosmos & Culture. From planning your work and applying for funding to getting and writing up results, project management affects every part of your research work." | /
./fasttext predict-prob models/dmoz_lvl1.bin - 14 0.1
__label__Arts  0.976
__label__Reference  0.905
__label__Science  0.539
Still too much “Arts” after removing “Museum” and “Culture”:

echo "The Environmental Exposure Group is also part of the MRC-PHE Centre for Environment and Health. For information on the Centre please visit their website. MRC-PHE Centre for Environment and Health In the Chair: Professor Paul Elliott King's College London (Room TBC) Science, South Kensington, London – first floor, entry via Cosmos. From planning your work and applying for funding, to getting and writing up results, project management affects every part of your research work." | \
./fasttext predict-prob models/dmoz_lvl1.bin - 14 0.1
__label__Arts 0.910
__label__Reference 0.743
__label__Science 0.492
Improvements Todo

- Decide how to deal with level 2 topics (E.g. Arts|Animation, Arts|Movies)
- Category overlaps
- Bad pages: bad page at a good web, bad page at an old/hijacked web
- Do not classify short documents (< 50 words) \( \Rightarrow \) Precision increase
<table>
<thead>
<tr>
<th></th>
<th>Table of Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Topic Classification</td>
</tr>
<tr>
<td>2</td>
<td>FastText + Active Learning Demo</td>
</tr>
</tbody>
</table>
Active Learning

Interactive machine learning procedure

- Used in supervised learning in multiple round annotation scheme
- Queries a source of truth (a human annotator) in the process of learning
- Aims to select the samples to improve the classifier the most in the next round rather than selecting samples randomly
Active Learning Case

Active Learning is beneficial when the following conditions are met:

- A lot of samples (web corpus documents)
- Training a classifier is cheap and fast (FastText)
- Annotation by a human is expensive (genre annotation takes 90 seconds per document in average in my annotation scheme)
Active Learning Approaches

Various approaches to selecting samples to annotate

• Uncertainty sampling (e.g. samples with the highest entropy of probdist over classes)
  • FastText gives probabilities of class labels ⇒ I am using this
  • Working directly with vector representations would give more options
• Reducing the hypothesis space (e.g. query by disagreement)
• Minimizing expected error and variance

- Data: 150 to 5000 sentences from csTenTen12 for fruits/vegetables
- Pre-trained skipgram models from csTenTen12
- Classes: User defined, e.g. fruit/vegetable, yellow/non-yellow
- Active Learning: User queried for each round of training a classifier
- Shows limits of using corpus samples
  - User’s rule matches the bias of the corpus (fruit/vegetable)  
    ⇒ Good result
  - No texts supporting user’s rule in the corpus (yellow/non-yellow)  
    ⇒ Poor result
Round 7: Consider the following sample: banán
Current prediction: VEGETABLE with 51% probability.
Enter the number of the bowl to put this sample in:


Training a new model. Prediction of the new distribution:

<table>
<thead>
<tr>
<th>FRUIT</th>
<th>VEGETABLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>BANÁN</td>
<td>BROKOLICE</td>
</tr>
<tr>
<td>DATLE</td>
<td>DÝNĚ</td>
</tr>
<tr>
<td>GRAPEFRUIT</td>
<td>MRKEV</td>
</tr>
<tr>
<td>MANGO</td>
<td>67% zelí</td>
</tr>
<tr>
<td>93% granátové jablko</td>
<td>67% paprika</td>
</tr>
<tr>
<td>81% mandarinka</td>
<td>65% květák</td>
</tr>
<tr>
<td>81% pomeranč</td>
<td>61% lilek</td>
</tr>
<tr>
<td>78% fík</td>
<td>60% okurka</td>
</tr>
<tr>
<td>65% kokos</td>
<td>56% rajče</td>
</tr>
<tr>
<td>61% ananas</td>
<td>55% brambor</td>
</tr>
</tbody>
</table>