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

Motivation – Web corpora

- What is inside?
- Subcorpora for users' needs
- My interset: Genres & Topics
  - Genres determined by style vs. topic determined by words => should be easier

Defining Topics

- Top to bottom…Apriori definition: Wordnet, Wikipedia, web directories (dmoz.org, urlblacklist.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 $\Rightarrow$ English
- 14 level 1 topics
  - Arts, Business, Computers, Games, Health, Home, News, Recreation, Reference, Regional, Science, Shopping, Society, Sports
- 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

Downloading Pages from dmoz.org

- 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

FastText Is Magic! :-)

- New functions: test, test-label, quantize, nn, analogies
- “Newer” functions: being added to the Git repository: autotune
- DIY C++

Autotune Is Magic!

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:
currentScore = 0.688365
train took = 11980.6
```

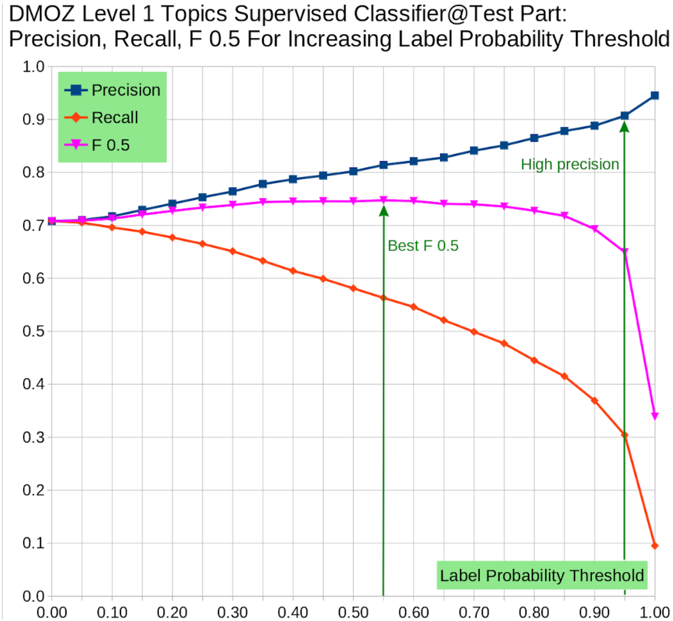
Best Level 2 Autotune (~1000 CPU-hours)

```
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:
currentScore = 0.567162
train took = 15250
```

Autotune Impolite Questions

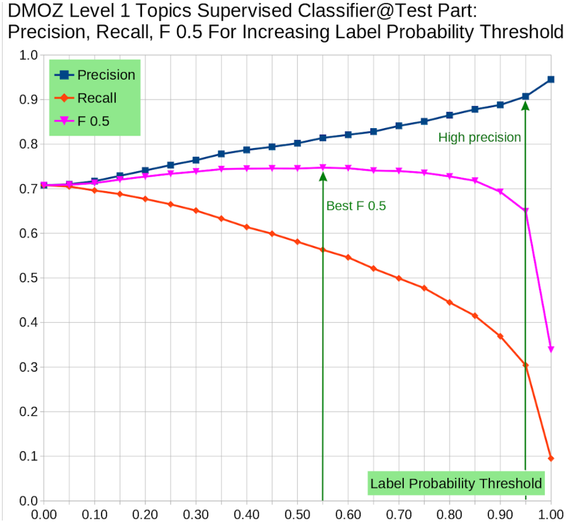
- More autotuning $\Rightarrow$ better result?
- Same algorithm, more CPUs for autotune $\Rightarrow$ competition winner?

Evaluation @Test



Setting the Label Probability Threshold

- High precision: 0.95
- Best F 0.5: 0.55 – Precision preferred at the cost of recall



Evaluation @Test, Threshold = 0.95

```
N 12205, P@1 0.907, R@1 0.304
F1 0.439  Precision 0.917  Recall 0.288  Arts
F1 0.206  Precision 0.853  Recall 0.117  Business
F1 0.515  Precision 0.912  Recall 0.359  Computers
F1 0.648  Precision 0.905  Recall 0.505  Games
F1 0.604  Precision 0.945  Recall 0.444  Health
F1 0.436  Precision 0.856  Recall 0.293  Home
F1 0.457  Precision 0.894  Recall 0.307  News
F1 0.428  Precision 0.890  Recall 0.282  Recreation
F1 0.396  Precision 0.914  Recall 0.252  Reference
F1 0.424  Precision 0.908  Recall 0.276  Regional
F1 0.413  Precision 0.895  Recall 0.268  Science
F1 0.394  Precision 0.880  Recall 0.254  Shopping
F1 0.361  Precision 0.907  Recall 0.225  Society
F1 0.619  Precision 0.927  Recall 0.464  Sports
```

Evaluation @Test, Threshold = 0.55

```
N 12205, P@1 0.814, R@1 0.563
F1 0.656  Precision 0.787  Recall 0.562  Arts
F1 0.505  Precision 0.710  Recall 0.392  Business
F1 0.722  Precision 0.804  Recall 0.655  Computers
F1 0.776  Precision 0.868  Recall 0.702  Games
F1 0.784  Precision 0.884  Recall 0.705  Health
F1 0.650  Precision 0.783  Recall 0.557  Home
F1 0.679  Precision 0.798  Recall 0.591  News
F1 0.641  Precision 0.798  Recall 0.536  Recreation
F1 0.639  Precision 0.831  Recall 0.519  Reference
F1 0.628  Precision 0.833  Recall 0.504  Regional
F1 0.645  Precision 0.798  Recall 0.541  Science
F1 0.631  Precision 0.806  Recall 0.519  Shopping
F1 0.580  Precision 0.775  Recall 0.464  Society
F1 0.770  Precision 0.873  Recall 0.689  Sports
```

Users' POW Evaluation @enTenTen15, Threshold = 0.55

173 random docs

Thr 0.95 => 14.9 % docs got a label

Thr 0.55 => 51.3 % docs got a label

	Threshold	0.95	0.55
Agreement	"That is the topic"	43	115
Weak Agreement	"That could be the topic"	6	25
Disagreement	"That is not the topic"	10	33
		83 %	81 %

Users' POW Evaluation @enTenTen15, Threshold = 0.55

Good	Bad	Class
14	4	Arts
6	3	Business
20	2	Computers
6	5	Games
9	0	Health
14	2	News
10	0	Recreation
10	2	Reference
12	3	Regional
14	4	Science
2	1	Shopping
16	4	Society
7	2	Sports

The Other Side of the Magic

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
```

The Other Side of the Magic

Still too much “Arts” after removing “Museum” and “Culture”:

```
echo "The Environmental Exposure Group is also part of th
MRC-PHE Centre for Environment and Health. For informatio
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 of Contents

Active Learning

Interactive machine learning procedure

- Used in supervised learning in multiple round annotation scheme
- Queries a source of truth (a human anontator) 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 benefitial 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 probabilites 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

According to Settles, Burr. “Active learning.” Synthesis Lectures on Artificial Intelligence and Machine Learning 6, no. 1 (2012): 1–114.

FastText + Active Learning Demo

- 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)
    $\Rightarrow$ Good result
  - No texts supporting user's rule in the corpus (yellow/non-yellow)
    $\Rightarrow$ Poor result

FastText + Active Learning Demo

```
Round 7: Consider the following sample: banán
Current prediction: VEGETABLE with 51% probability.
Enter the number of the bowl to put this sample in:
    [1] FRUIT, [2] VEGETABLE or [q] to quit: 1
Training a new model. Prediction of the new distribution:
| ==== FRUIT ====         | ==== VEGETABLE ====    |
| 100% BANÁN              | 100% BROKOLICE         |
| 100% DATLE              | 100% DÝNĚ              |
| 100% GRAPEFRUIT         | 100% MRKEV             |
| 100% MANGO              |  67% zelí              |
|  93% granátové jablko   |  67% paprika           |
|  81% mandarinka         |  65% květák            |
|  81% pomeranč           |  61% lilek             |
|  78% fík                |  60% okurka            |
|  65% kokos              |  56% rajče             |
|  61% ananas             |  55% brambor           |
```