02_using_trained_vectors

September 29, 2021

0.1 Imports & Settings

[1]:
```
from time import time
import warnings
from collections import Counter
from pathlib import Path
import pandas as pd
import numpy as np

import matplotlib.pyplot as plt

from gensim.models import Word2Vec, KeyedVectors
from gensim.scripts.glove2word2vec import glove2word2vec
```

[2]:
```
warnings.filterwarnings('ignore')
```

[3]:
```
analogies_path = Path('data', 'analogies', 'analogies-en.txt')
```

0.2 Convert GloVE Vectors to gensim format

The various GloVE vectors are available here. Download link for the wikipedia version. Unzip and store in `data/glove`.

0.2.1 WikiPedia

[4]:
```
glove_path = Path('data/glove')
glove_wiki_file= glove_path / 'glove.6B.300d.txt'
word2vec_wiki_file = glove_path / 'glove.wiki.gensim.txt'
```

[ ]:
```
glove2word2vec(glove_input_file=glove_wiki_file,␣
↪word2vec_output_file=word2vec_wiki_file)
```

0.2.2 Twitter Data

[18]:
```
glove_twitter_file= glove_path / 'glove.twitter.27B.200d.txt'
word2vec_twitter_file = glove_path / 'glove.twitter.gensim.txt'
```

[19]:
```
glove2word2vec(glove_input_file=glove_twitter_file,␣
↪word2vec_output_file=word2vec_twitter_file)
```

[19]:
```
(1193517, 200)
```

0.2.3 Common Crawl

[26]:
```
glove_crawl_file= glove_path / 'glove.840B.300d.txt'
word2vec_crawl_file = glove_path / 'glove.crawl.gensim.txt'
```

[27]:
```
glove2word2vec(glove_input_file=glove_crawl_file,␣
↪word2vec_output_file=word2vec_crawl_file)
```

[27]:
```
(2196018, 300)
```

0.3 Evaluate embeddings

[37]:
```
def eval_analogies(file_name, vocab=30000):
    model = KeyedVectors.load_word2vec_format(file_name, binary=False)
    accuracy = model.wv.accuracy(analogies_path,
                                 restrict_vocab=vocab,
                                 case_insensitive=True)
    return (pd.DataFrame([[c['section'],
                           len(c['correct']),
                           len(c['incorrect'])] for c in accuracy],
                         columns=['category', 'correct', 'incorrect'])
            .assign(samples=lambda x: x.correct.add(x.incorrect))
            .assign(average=lambda x: x.correct.div(x.samples))
            .drop(['correct', 'incorrect'], axis=1))
```

[40]:
```
result = eval_analogies(word2vec_twitter_file, vocab=100000)
```

0.3.1 twitter result

[41]:
```
result
```

[41]:

	category	samples	average
0	capital-common-countries	462	0.701299
1	capital-world	930	0.690323
2	city-in-state	3644	0.350714
3	currency	268	0.018657
4	family	342	0.824561
5	gram1-adjective-to-adverb	650	0.143077
6	gram2-opposite	342	0.365497
7	gram3-comparative	1260	0.757937
8	gram4-superlative	930	0.686022

9	gram5-present-participle	702	0.750712
10	gram6-nationality-adjective	870	0.750575
11	gram7-past-tense	1190	0.576471
12	gram8-plural	1122	0.811052
13	gram9-plural-verbs	600	0.655000
14	total	13312	0.564228

0.3.2 wiki result

[39]:
```
result
```

[39]:

	category	samples	average
0	capital-common-countries	506	0.948617
1	capital-world	8372	0.964644
2	city-in-state	4242	0.599953
3	currency	752	0.174202
4	family	506	0.881423
5	gram1-adjective-to-adverb	992	0.225806
6	gram2-opposite	756	0.285714
7	gram3-comparative	1332	0.882132
8	gram4-superlative	1056	0.746212
9	gram5-present-participle	1056	0.699811
10	gram6-nationality-adjective	1640	0.925000
11	gram7-past-tense	1560	0.611538
12	gram8-plural	1332	0.780781
13	gram9-plural-verbs	870	0.585057
14	total	24972	0.754445

0.3.3 Common Crawl result

[33]:
```
result
```

[33]:

	category	samples	average
0	capital-common-countries	506	0.946640
1	capital-world	4290	0.917483
2	city-in-state	4242	0.706742
3	currency	206	0.184466
4	family	420	0.978571
5	gram1-adjective-to-adverb	992	0.388105
6	gram2-opposite	702	0.363248
7	gram3-comparative	1332	0.876877
8	gram4-superlative	1122	0.919786
9	gram5-present-participle	1056	0.827652
10	gram6-nationality-adjective	1406	0.948791
11	gram7-past-tense	1560	0.621154
12	gram8-plural	1332	0.864114
13	gram9-plural-verbs	870	0.672414

```
14                        total    20036  0.779347
```

[16]:
```
result
```

[16]:

	category	correct	incorrect	average
0	capital-common-countries	482	24	0.952569
1	capital-world	6093	227	0.964082
2	city-in-state	2472	1646	0.600291
3	currency	112	390	0.223108
4	family	392	28	0.933333
5	gram1-adjective-to-adverb	228	764	0.229839
6	gram2-opposite	205	497	0.292023
7	gram3-comparative	1175	157	0.882132
8	gram4-superlative	737	193	0.792473
9	gram5-present-participle	686	306	0.691532
10	gram6-nationality-adjective	1445	37	0.975034
11	gram7-past-tense	954	606	0.611538
12	gram8-plural	1016	244	0.806349
13	gram9-plural-verbs	472	340	0.581281
14	total	16469	5459	0.751049

[17]:
```
result.to_csv(glove_path / 'accuracy.csv', index=False)
```

[ ]:
```

```