08 - Word2Vec & Doc2Vec

April 28, 2020

1 Word2Vec & Doc2Vec (gensim)

Analysis of Social Media Contents

Alessandro Ortis - University of Catania

Partially based on: Dive Into NLTK, Part X: Play with Word2Vec Models based on NLTK Corpus by TextMiner

1.1 Exploring the `gutenburg` corpus

Project Gutenberg (PG) is a volunteer effort to digitize and archive cultural works. Most of the items in its collection are full texts of public domain books.

```
In [1]: from nltk.corpus import gutenberg
        #gutenberg.readme().replace('\n', ' ')
```

```
In [2]: gutenberg.fileids()
```

```
Out[2]: ['austen-emma.txt',
         'austen-persuasion.txt',
         'austen-sense.txt',
         'bible-kjv.txt',
         'blake-poems.txt',
         'bryant-stories.txt',
         'burgess-busterbrown.txt',
         'carroll-alice.txt',
         'chesterton-ball.txt',
         'chesterton-brown.txt',
         'chesterton-thursday.txt',
         'edgeworth-parents.txt',
         'melville-moby_dick.txt',
         'milton-paradise.txt',
         'shakespeare-caesar.txt',
         'shakespeare-hamlet.txt',
         'shakespeare-macbeth.txt',
         'whitman-leaves.txt']
```

```
In [3]: bible_kjv_sents = gutenberg.sents('bible-kjv.txt')
        len(bible_kjv_sents)
```

```
Out[3]: 30103
```

1.2 Implementing Word2Vec

In [4]:
```
from string import punctuation

discard_punctuation_and_lowercased_sents = [[word.lower() for word in sent if word not
                                             for sent in bible_kjv_sents]
discard_punctuation_and_lowercased_sents[3]
```

Out[4]:
```
['in',
 'the',
 'beginning',
 'god',
 'created',
 'the',
 'heaven',
 'and',
 'the',
 'earth']
```

In [5]:
```
from gensim.models import word2vec

bible_kjv_word2vec_model = word2vec.Word2Vec(discard_punctuation_and_lowercased_sents,
#bible_kjv_word2vec_model = word2vec.Word2Vec(discard_punctuation_and_lowercased_sents
bible_kjv_word2vec_model.save('bible_word2vec_gensim')
# model = Word2Vec.load(fname) # To load a model
word_vectors = bible_kjv_word2vec_model.wv
#del bible_kjv_word2vec_model # When we finish training the model, we can only delete
word_vectors.save_word2vec_format('bible_word2vec_org', 'bible_word2vec_vocabulary')
len(word_vectors.vocab)
```

Out[5]: 5279

In [6]:
```
word_vectors.most_similar(['god']) # Most similar as in closest in the word graph. Wor
```

Out[6]:
```
[('hosts', 0.7622880339622498),
 ('truth', 0.7604811191558838),
 ('lord', 0.757301926612854),
 ('glory', 0.7538343667984009),
 ('salvation', 0.7431286573410034),
 ('spirit', 0.7370332479476929),
 ('christ', 0.7303678393363953),
 ('faith', 0.7278977632522583),
 ('righteousness', 0.6934632062911987),
 ('fear', 0.6776974201202393)]
```

In [7]:
```
word_vectors.most_similar(['heaven'], topn=3)
```

Out[7]:
```
[('earth', 0.7155585289001465),
 ('heavens', 0.659392237663269),
 ('darkness', 0.6375309228897095)]
```

```
In [8]: word_vectors.most_similar(["moses"], topn=5)

Out[8]: [('joshua', 0.8303318619728088),
         ('jeremiah', 0.8232778310775757),
         ('hezekiah', 0.7776656150817871),
         ('samuel', 0.772584080696106),
         ('balaam', 0.7700286507606506)]

In [9]: word_vectors.most_similar(positive=['woman', 'king'], negative=['man'], topn=1)

Out[9]: [('daughter', 0.6467757225036621)]

In [27]: # The `_cosmul` variant uses a slightly-different comparison when using multiple posi
         word_vectors.most_similar_cosmul(positive=['woman', 'king'], negative=['man'], topn=1)

Out[27]: [('chamberlains', 0.987970769405365)]

In [28]: word_vectors.similarity('lord', 'god')

Out[28]: 0.74958503

In [36]: word_vectors.doesnt_match("lord god salvation food spirit".split())

C:\Program Files (x86)\Microsoft Visual Studio\Shared\Anaconda3_64\lib\site-packages\gensim\mod
  vectors = vstack(self.word_vec(word, use_norm=True) for word in used_words).astype(REAL)


Out[36]: 'food'
```

2 Doc2Vec

2.0.1 Exercise

Try to explore the Doc2vec model implemented in genism.

```
In [4]: from gensim.models.doc2vec import Doc2Vec, TaggedDocument
        from nltk.tokenize import word_tokenize
```