

Brody_Vogel_NLP_Assignment4

September 29, 2018

```
In [1]: # Brody Vogel NLP Homework 4 #
        from gensim.models import KeyedVectors

        vecfile = '/Users/brodyvogel/Downloads/GoogleNews-vectors-negative300.bin'

        vecs = KeyedVectors.load_word2vec_format(vecfile, binary = True)

In [4]: # Part 1
        # The dimensionality of the word embeddings is (300,)
        # There are 3 million word embeddings; each is a 1-D, 300-item vector
        print(len(vecs.vectors))
        dog = vecs['dog']
        dog.shape

3000000

Out[4]: (300,)
```

```
In [5]: result = vecs.most_similar(positive = ['picnic'], topn = 5)

In [6]: # The five most-similar words are 'picnics', 'picnic-lunch', 'Picnic',
        # 'potluck-picnic', and 'picnic-supper'
        result

Out[6]: [('picnics', 0.7400875091552734),
         ('picnic_lunch', 0.721373975276947),
         ('Picnic', 0.7005339860916138),
         ('potluck_picnic', 0.6683274507522583),
         ('picnic_supper', 0.6518912315368652)]

In [7]: # 'tissue' doesn't match the rest of the words
        vecs.doesnt_match(['tissue', 'papyrus', 'manila', 'newsprint', 'parchment', 'gazette'])

Out[7]: 'tissue'

In [8]: # 'leg' is to 'jump' as 'forearm' is to 'throw'
        vecs.most_similar(positive = ['leg', 'throw'], negative = ['jump'])[0]

Out[8]: ('forearm', 0.482946515083313)
```