In [1]:

```
from pycorenlp import StanfordCoreNLP
nlp = StanfordCoreNLP('http://localhost:9000')
```

In [2]:

```
import numpy as np
import time
import word2vec
import string
import cPickle as pickle
```

In [3]:

```
in_data = None
with open('data/news_reuters_5.csv', 'r') as infile:
    in_data = infile.read().split('\n')
print len(in_data)
```

```
16787
```

In [4]:

```
title_data = []
for article in in_data:
    fields = article.split(',')
    if len(fields) < 4:
        continue
    title_data += [(fields[0], fields[2], fields[3])]
#     if len(title_data) >= 10:
#         break
```

In [5]:

```
print title_data[0]
```

```
('GOOG', '20171107', "Alphabet's Waymo to launch robotaxis with no hu
man in driver's seat ")
```

In [6]:

```
titles = []
for t in title_data:
    title = t[2].strip()
    title = ''.join(i for i in title if ord(i)<128)
    title = title.replace('  ', ' ')
    title = title.replace('\'s', '')
    for i in range(3):
        if '-' in title and (title[title.find('-') - 1].isupper() or
'UPDATE' in title):
#             print 'REEE'
#             print title
#             print title[title.find('-') + 1:]
            title = title[title.find('-') + 1:]
        else:
            break

    title = title.translate(None, string.punctuation)
    title = title + '.'
    title = title.lower()
    if len(titles) > 0 and title in titles[-1][2]:
        continue
    titles += [(t[0], t[1], title)]
#     if len(titles) >= 10:
#         break
print titles[:10]
print len(titles)
```

```
[('GOOG', '20171107', 'alphabet waymo to launch robotaxis with no hum
an in driver seat.'), ('GOOG', '20171107', 'indonesia to summon messe
nger search engine providers over content.'), ('GOOG', '20171102', 'a
utonation announces waymo fleet repair deal shares jump.'), ('GOOG',
'20171102', 'us lawmakers release sample of russianbought facebook ad
s.'), ('GOOG', '20171031', 'google ditched autopilot driving feature
after test user napped behind wheel.'), ('GOOG', '20171027', 'alphabe
t mobile ad revenue surges shares jump.'), ('GOOG', '20171027', 'no e
nd in sight for tech giant share gains.'), ('GOOG', '20171026', 'alph
abet posts qtrly earnings per share of 957.'), ('GOOG', '20171026',
'alphabet revenue rises 24 pct on mobile advertising growth.'), ('GOO
G', '20171026', 'alphabet looks to snowy michigan to test selfdriving
cars.')]
12680
```

In [7]:

```
text = '. '.join([t[2] for t in titles])
print len(text)
print text[:1000]
```

```
756721
alphabet waymo to launch robotaxis with no human in driver seat.. ind
onesia to summon messenger search engine providers over content.. aut
onation announces waymo fleet repair deal shares jump.. us lawmakers 
release sample of russianbought facebook ads.. google ditched autopil
ot driving feature after test user napped behind wheel.. alphabet mob
ile ad revenue surges shares jump.. no end in sight for tech giant sh
are gains.. alphabet posts qtrly earnings per share of 957.. alphabet
revenue rises 24 pct on mobile advertising growth.. alphabet looks to
snowy michigan to test selfdriving cars.. alphabet balloon project to
provide limited internet in puerto rico.. alphabet capitalg leads lyf
t 1 billion funding round.. new york times business news  oct 20.. al
phabet capitalg leads lyft 1 bln funding round.. lyft says alphabet l
eads latest 1 billion round of funding.. alphabet to develop hightech
waterfront site in toronto.. google launches advanced gmail security 
features for highrisk users..
```

In [8]:

```
with open('data/texts/text.txt', 'w') as outfile:
    outfile.write(text)
word2vec.word2phrase('data/texts/text.txt', 'data/texts/text-phrases.
txt', verbose=True)
```

```
Starting training using file data/texts/text.txt
Words processed: 100K     Vocab size: 59K
Vocab size (unigrams + bigrams): 37586
Words in train file: 117829
```

In [9]:

```
texts = ['']
title2info = {}
index = 0
i = 0
# with open('data/texts/text-phrases.txt', 'r') as infile:
with open('data/texts/text.txt', 'r') as infile:
    for t in infile.read().split('.. '):
        if len(texts[index]) + len(t) > 1e4:
            index += 1
            texts += ['']
        texts[index] += t + ". "
        title2info[t] = (titles[i])
        i += 1
print len(texts)
print title2info.items()[:3]
print len(title2info), len(titles), len(titles) - len(title2info)
```

```
75
[('qualcomm talks up future toptier smartphone chip', ('QCOM', '20140
407', 'qualcomm talks up future toptier smartphone chip.')), ('paulso
n  co inc takes share stake in dish network monsanto', ('GOOGL', '201
70515', 'paulson  co inc takes share stake in dish network monsant
o.')), ('intel pledges 125 mln for startups that back women minoritie
s', ('INTC', '20150609', 'intel pledges 125 mln for startups that bac
k women minorities.'))]
9921 12680 2759
```

In [10]:

```
# print texts[0]
```

In [11]:

```
start_time = time.time()

relations = [0 for i in range(100)]
outputs = []
for i in range(len(texts)):
    output = nlp.annotate(texts[i], properties={
      'annotators': 'openie',
      'outputFormat': 'json'
      })

    outputs += [output]
    # print output
#     print len(output['sentences'])
    for j in range(len(output['sentences'])):
        relations[len(output['sentences'][j]['openie'])] += 1
#     print 'text {}: {} sec'.format(i+1, time.time() - start_time)

print relations
```

```
[4132, 2096, 2417, 1329, 1193, 382, 445, 116, 209, 76, 51, 20, 50, 8,
13, 4, 18, 4, 5, 2, 2, 2, 1, 1, 10, 1, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
```

In [12]:

```
print sum(relations[1:]), sum(relations)
```

```
8458 12590
```

In [13]:

```
print outputs[0]['sentences'][3]['openie']
```

```
[{u'subjectSpan': [0, 1], u'relationSpan': [2, 3], u'objectSpan': [3,
4], u'object': u'sample', u'relation': u'release', u'subject': u'u
s'}, {u'subjectSpan': [0, 1], u'relationSpan': [2, 3], u'objectSpan':
[3, 8], u'object': u'sample of russianbought facebook ads', u'relatio
n': u'release', u'subject': u'us'}, {u'subjectSpan': [1, 2], u'relati
onSpan': [2, 3], u'objectSpan': [3, 8], u'object': u'sample of russia
nbought facebook ads', u'relation': u'release', u'subject': u'lawmake
rs'}, {u'subjectSpan': [1, 2], u'relationSpan': [2, 3], u'objectSpa
n': [3, 4], u'object': u'sample', u'relation': u'release', u'subjec
t': u'lawmakers'}]
```

In [14]:

```
print sum([len(outputs[i]['sentences']) for i in range(len(outputs
))])
```

```
12590
```

In [15]:

```
data = {}
bad = 0
good = 0
weird = 0
for o in outputs:
    for s in o['sentences']:
        if len(s['openie']) > 0:
            try:
                info = title2info[' '.join([s['tokens'][i]['word'] fo
r i in range(len(s['tokens']))][:-1])]
    #             print s['openie']
                if info in data:
                    weird += 1
                data[info] = s['openie']
                good += 1
            except:
#                 print ' '.join([s['tokens'][i]['word'] for i in ran
ge(len(s['tokens']))])
                bad += 1
#                 print info
#                 print s['openie']
print len(data), bad, good, weird

with open('openie.p', 'w') as outfile:
    pickle.dump(data, outfile)
```

```
6193 425 8033 1840
```

In [ ]:

```

```

In [ ]:

```

```

```
In [ ]:
```