Description of notebook

A: Preprocessing demonstration

- Corpus 1: Jane Austen's pride and prejudice
- Corpus 2: 20 Newsgroups corpus

B: Preprocessing detailed (step by step)

C: Word2Vec model

A: Preprocessing

In [1]:
```
import os
import pandas as pd
import re
import string

import nltk
from nltk.tokenize import sent_tokenize
from nltk.tokenize import word_tokenize
import gensim
# from gensim.test.utils import common_texts, get_tmpfile
from gensim.models import Word2Vec
```

Corpus 1

Obtaining text

In [2]:
```
with open('../HW1/Pride and Prejudice - Jane Austen Chapter 1 to 20.txt'
) as f:
    corpus_1 = f.read()
```

In [3]:
```
print(corpus_1[0:500])
```

```
Chapter 1


      It is a truth universally acknowledged, that a single man in
      possession of a good fortune, must be in want of a wife.

      However little known the feelings or views of such a man may be
      on his first entering a neighbourhood, this truth is so well
      fixed in the minds of the surrounding families, that he is
      considered as the rightful property of some one or other of their
      daughters.

      “My dear Mr. Bennet,” said his lady to him one day, “have yo
```

Cleaning

In [4]:
```
def preprocessing_corpus_1(corpus):

    '''Takes string from novel, normalizes and tokenizes into sentence
s'''

    #Removal of white space
    step_1 = re.sub('\s+', ' ', corpus)
    # Removal of chapter tags
    step_2 = re.sub(r'Chapter \d* ', '', step_1)
    # Removal of digits
    step_3 = re.sub(r'\d+', '', step_2)
    # Remove other signs
    step_4 = re.sub(r'["_“”\'\`\-\*\(\)]','',step_3)
    #Tokenize to sentences while punctuation is still in place
    tokens_jane_austen = sent_tokenize(step_4)
    # Convert each token to lowercase
    lower_token = list(map(lambda token: token.lower(), tokens_jane_aust
en))
    # Remove punctuation from lowercase token
    punct_less_token = list(map(lambda token:
                                token.translate(str.maketrans('', '', string
.punctuation)), lower_token))

    return punct_less_token
```

In [5]:
```
processed_1 = preprocessing_corpus_1(corpus_1)
processed_1[0:10]
```

Out[5]:
```
['it is a truth universally acknowledged that a single man in possessio
n of a good fortune must be in want of a wife',
 'however little known the feelings or views of such a man may be on hi
s first entering a neighbourhood this truth is so well fixed in the min
ds of the surrounding families that he is considered as the rightful pr
operty of some one or other of their daughters',
 'my dear mr bennet said his lady to him one day have you heard that ne
therfield park is let at last',
 'mr bennet replied that he had not',
 'but it is returned she for mrs long has just been here and she told m
e all about it',
 'mr bennet made no answer',
 'do not you want to know who has taken it',
 'cried his wife impatiently',
 'you want to tell me and i have no objection to hearing it',
 'this was invitation enough']
```

Writing to text file - including spacing between each

In [6]:
```
textfile = open("corpus_1_text_file.txt", "w")
for element in processed_1:
    textfile.write(element + "\n" + "\n")
textfile.close()
```

Corpus 2:

Obtaining text

In [7]:
```
files_list = []
path = os.getcwd() + '/20news-bydate-train/'
for root, dirs, files in os.walk(path, topdown = False):
    for name in files:
        files_list.append(os.path.join(root, name))

corpus_2_list = []
for i in files_list:
    with open(i, 'r', encoding="utf8", errors="ignore") as f:
        file = f.read()
        corpus_2_list.append(file)
```

In [60]:
```
corpus_2_list[1000]
```

Out[60]:
```
'From: zowie@daedalus.stanford.edu (Craig "Powderkeg" DeForest)\nSubjec
t: Re: 5W30, 10W40, or 20W50\nArticle-I.D.: daedalus.ZOWIE.93Apr5215616
\nOrganization: Stanford Center for Space Science and Astrophysics\nLin
es: 37\nNNTP-Posting-Host: daedalus.stanford.edu\nIn-reply-to: Brad Tho
ne\'s message of Fri, 02 Apr 93 21:41:53 CST\n\nIn article <foo> Brad T
hone <C09615BT@WUVMD> writes:\nWell, there *is* a difference.\n\nI don
\'t happen to have my SAE manual handy, but oil viscosity in general\n_
decreases_ with temperature.  The SAE numbers are based on a `typical
\'\ncurve that oils used to all have, running from (say) the viscosity
of a\nroom-temperature 90-weight at 0C, down to (say) that of a room-te
mperature \n5-weight at 20C, for a typical 40-weight oil.\n\nOils that
are designed for operation in `normal\' temperatures just have\na weigh
t specification.  Oils that are designed for operation in exceedingly\n
cold temperatures have a `W\' tacked on the end, so in winter in a cold
\nplace, you\'d stick 10W in your car in the winter and 40 in it in the
summer,\nto approximate the appropriate viscosity throughout the yea
r.\n\nModern multi-viscosity oils change viscosity much less with tempe
rature.\nAs a result, their viscosity graphs cross over several curves.
A multi-vis\nspecification pegs the curve at two temperatures, a `norma
l\' operating\ntemperature and a `cold\' one (though I can\'t remember
the numbers...).\n\nIn any event, the weights do indicate a significant
difference.  Remember\nthat your engine is temperature-regulated (by th
e thermostat and\nradiator or air fins) most of the time -- unless you
overheat it or\nsomething.\n\nAny weight of oil is better than no oil,
or than very old, carbonized\noil.  Thin oil won\'t (in general) lubric
ate as well at temperature,\nthicker oil will (like a 20W50) will lubri
cate better at temperature, \nbut not as well during startup, when most
engine wear occurs.  \n\nIf you\'re planning on making long drives, the
20W50 is probably fine\n(esp. in the summer) in your 10W40 car.  But if
you\'re making short drives,\nstick to the 10W40.\n\n\n--\nDON\'T DRINK
SOAP! DILUTE DILUTE! OK!\n'
```

Cleaning

In [61]:
```
def preprocessing_corpus_2(corpus):

    '''Takes string from news corpus, normalizes and tokenizes into sent
ences'''

    #Removal of white space
    step_1 = re.sub('\s+', ' ', corpus)
    # RFrom/To and subject line of email
    step_2 = re.sub(r'\bFrom: .*? writes: ', '', step_1)
    # Removal of digits
    step_3 = re.sub(r'\d+', '', step_2)
    # Remove other signs
    step_4 = re.sub(r'["_“”\'\`\-\*\(\)]','',step_3)
    #Tokenize to sentences while punctuation is still in place
    tokens_news = sent_tokenize(step_4)
    # Convert each token to lowercase
    lower_token = list(map(lambda token: token.lower(), tokens_news))
    # Remove punctuation from lowercase token
    punct_less_token = list(map(lambda token:
                                token.translate(str.maketrans('', '', string
.punctuation)), lower_token))

    return punct_less_token
```

In [62]:
```
preprocessing_corpus_2(corpus_2_list[1000])
```

Out[62]:
```
['well there is a difference',
 'i dont happen to have my sae manual handy but oil viscosity in genera
l decreases with temperature',
 'the sae numbers are based on a typical curve that oils used to all ha
ve running from say the viscosity of a roomtemperature weight at c down
to say that of a roomtemperature weight at c for a typical weight oil',
 'oils that are designed for operation in normal temperatures just have
a weight specification',
 'oils that are designed for operation in exceedingly cold temperatures
have a w tacked on the end so in winter in a cold place youd stick w in
your car in the winter and  in it in the summer to approximate the appr
opriate viscosity throughout the year',
 'modern multiviscosity oils change viscosity much less with temperatur
e',
 'as a result their viscosity graphs cross over several curves',
 'a multivis specification pegs the curve at two temperatures a normal
operating temperature and a cold one though i cant remember the number
s',
 'in any event the weights do indicate a significant difference',
 'remember that your engine is temperatureregulated by the thermostat a
nd radiator or air fins most of the time  unless you overheat it or som
ething',
 'any weight of oil is better than no oil or than very old carbonized o
il',
 'thin oil wont in general lubricate as well at temperature thicker oil
will like a w will lubricate better at temperature but not as well duri
ng startup when most engine wear occurs',
 'if youre planning on making long drives the w is probably fine esp',
 'in the summer in your w car',
 'but if youre making short drives stick to the w  dont drink soap',
 'dilute dilute',
 'ok']
```

Writing to text file - including spacing between each

In [63]:
```
textfile = open("corpus_2_text_file.txt", "w")
for news_story in corpus_2_list:
    tokenized_story = preprocessing_corpus_2(news_story)
    for sentence in tokenized_story:
        textfile.write(sentence + "\n" + "\n")
textfile.close()
```

B: Preprocessing detailed

Corpus 1

Book before cleaning

In [64]:
```
corpus_1[0:1000]
```

Out[64]:
```
'Chapter 1\n\n      It is a truth universally acknowledged, that a sing
le man in\n      possession of a good fortune, must be in want of a wif
e.\n\n      However little known the feelings or views of such a man ma
y be\n      on his first entering a neighbourhood, this truth is so wel
l\n      fixed in the minds of the surrounding families, that he is\n
considered as the rightful property of some one or other of their\n
daughters.\n\n      “My dear Mr. Bennet,” said his lady to him one day,
“have you\n      heard that Netherfield Park is let at last?”\n\n
Mr. Bennet replied that he had not.\n\n      “But it is,” returned she;
“for Mrs. Long has just been here, and\n      she told me all about i
t.”\n\n      Mr. Bennet made no answer.\n\n      “Do not you want to kn
ow who has taken it?” cried his wife\n      impatiently.\n\n      “_You
_ want to tell me, and I have no objection to hearing it.”\n\n      Thi
s was invitation enough.\n\n      “Why, my dear, you must know, Mrs. Lo
ng says that Netherfield is\n '
```

Removal of white space

In [65]:
```
# Remove white space and new lines
step_1 = re.sub('\s+', ' ', corpus_1)
step_1[0:1000]
```

Out[65]:
```
'Chapter 1 It is a truth universally acknowledged, that a single man in
possession of a good fortune, must be in want of a wife. However little
known the feelings or views of such a man may be on his first entering
a neighbourhood, this truth is so well fixed in the minds of the surrou
nding families, that he is considered as the rightful property of some
one or other of their daughters. “My dear Mr. Bennet,” said his lady to
him one day, “have you heard that Netherfield Park is let at last?” Mr.
Bennet replied that he had not. “But it is,” returned she; “for Mrs. Lo
ng has just been here, and she told me all about it.” Mr. Bennet made n
o answer. “Do not you want to know who has taken it?” cried his wife im
patiently. “_You_ want to tell me, and I have no objection to hearing i
t.” This was invitation enough. “Why, my dear, you must know, Mrs. Long
says that Netherfield is taken by a young man of large fortune from the
north of England; that he came down on Monday in a chaise and four to s
e'
```

Removal of chapter tags

The word chapter appears 20 times

In [66]:
```
len(re.findall(r'Chapter ', step_1))
```

Out[66]:
```
20
```

All of these 20 times are for new chapters (therefore the word does not appear in the text itself)

In [67]:
```
re.findall(r'Chapter \d* ', step_1)[18:20]
```

Out[67]:
```
['Chapter 19 ', 'Chapter 20 ']
```

In [68]:
```
# Remove chapter tags
step_2 = re.sub(r'Chapter \d* ', '', step_1)
step_2[0:1000]
```

Out[68]:
```
'It is a truth universally acknowledged, that a single man in possessio
n of a good fortune, must be in want of a wife. However little known th
e feelings or views of such a man may be on his first entering a neighb
ourhood, this truth is so well fixed in the minds of the surrounding fa
milies, that he is considered as the rightful property of some one or o
ther of their daughters. “My dear Mr. Bennet,” said his lady to him one
day, “have you heard that Netherfield Park is let at last?” Mr. Bennet
replied that he had not. “But it is,” returned she; “for Mrs. Long has
just been here, and she told me all about it.” Mr. Bennet made no answe
r. “Do not you want to know who has taken it?” cried his wife impatient
ly. “_You_ want to tell me, and I have no objection to hearing it.” Thi
s was invitation enough. “Why, my dear, you must know, Mrs. Long says t
hat Netherfield is taken by a young man of large fortune from the north
of England; that he came down on Monday in a chaise and four to see the
plac'
```

Removal of digits

Digits appear twice in the text

In [69]:
```
re.findall(r'\d+', step_2)
```

Out[69]:
```
['15', '18']
```

Here is one example

In [70]:
```
re.findall(r'Monday, November 18th, by four o’clock, and shall probably
 trespass on your hospitality ', step_2)
```

Out[70]:
```
['Monday, November 18th, by four o’clock, and shall probably trespass o
n your hospitality ']
```

In [71]:
```
# Remove all digits
step_3 = re.sub(r'\d+', '', step_2)
step_3[0:1000]
```

Out[71]:
```
'It is a truth universally acknowledged, that a single man in possessio
n of a good fortune, must be in want of a wife. However little known th
e feelings or views of such a man may be on his first entering a neighb
ourhood, this truth is so well fixed in the minds of the surrounding fa
milies, that he is considered as the rightful property of some one or o
ther of their daughters. “My dear Mr. Bennet,” said his lady to him one
day, “have you heard that Netherfield Park is let at last?” Mr. Bennet
replied that he had not. “But it is,” returned she; “for Mrs. Long has
just been here, and she told me all about it.” Mr. Bennet made no answe
r. “Do not you want to know who has taken it?” cried his wife impatient
ly. “_You_ want to tell me, and I have no objection to hearing it.” Thi
s was invitation enough. “Why, my dear, you must know, Mrs. Long says t
hat Netherfield is taken by a young man of large fortune from the north
of England; that he came down on Monday in a chaise and four to see the
plac'
```

Remove futile symbols (non punctuation)

In [72]:
```
# Remove other signs
step_4 = re.sub(r'["_“”\'\`\-\*\(\)]','',step_3)
step_4[0:1000]
```

Out[72]:
```
'It is a truth universally acknowledged, that a single man in possessio
n of a good fortune, must be in want of a wife. However little known th
e feelings or views of such a man may be on his first entering a neighb
ourhood, this truth is so well fixed in the minds of the surrounding fa
milies, that he is considered as the rightful property of some one or o
ther of their daughters. My dear Mr. Bennet, said his lady to him one d
ay, have you heard that Netherfield Park is let at last? Mr. Bennet rep
lied that he had not. But it is, returned she; for Mrs. Long has just b
een here, and she told me all about it. Mr. Bennet made no answer. Do n
ot you want to know who has taken it? cried his wife impatiently. You w
ant to tell me, and I have no objection to hearing it. This was invitat
ion enough. Why, my dear, you must know, Mrs. Long says that Netherfiel
d is taken by a young man of large fortune from the north of England; t
hat he came down on Monday in a chaise and four to see the place, and w
as so m'
```

Tokenize to sentences while punctuation is still in place

In [73]:
```
# Tokenize
tokens_jane_austen = sent_tokenize(step_4)
tokens_jane_austen[0:10]
```

Out[73]:
```
['It is a truth universally acknowledged, that a single man in possessi
on of a good fortune, must be in want of a wife.',
 'However little known the feelings or views of such a man may be on hi
s first entering a neighbourhood, this truth is so well fixed in the mi
nds of the surrounding families, that he is considered as the rightful
property of some one or other of their daughters.',
 'My dear Mr. Bennet, said his lady to him one day, have you heard that
Netherfield Park is let at last?',
 'Mr. Bennet replied that he had not.',
 'But it is, returned she; for Mrs. Long has just been here, and she to
ld me all about it.',
 'Mr. Bennet made no answer.',
 'Do not you want to know who has taken it?',
 'cried his wife impatiently.',
 'You want to tell me, and I have no objection to hearing it.',
 'This was invitation enough.']
```

Convert each token to lowercase

In [74]:
```
# Lowercase
lower_token = list(map(lambda token: token.lower(), tokens_jane_austen))
lower_token[0:5]
```

Out[74]:
```
['it is a truth universally acknowledged, that a single man in possessi
on of a good fortune, must be in want of a wife.',
 'however little known the feelings or views of such a man may be on hi
s first entering a neighbourhood, this truth is so well fixed in the mi
nds of the surrounding families, that he is considered as the rightful
property of some one or other of their daughters.',
 'my dear mr. bennet, said his lady to him one day, have you heard that
netherfield park is let at last?',
 'mr. bennet replied that he had not.',
 'but it is, returned she; for mrs. long has just been here, and she to
ld me all about it.']
```

Remove punctuation from lowercase token

In [75]:
```
punct_less_token = list(map(lambda token:
                            token.translate(str.maketrans('', '', string
.punctuation)), lower_token))
punct_less_token[0:5]
```

Out[75]:
```
['it is a truth universally acknowledged that a single man in possessio
n of a good fortune must be in want of a wife',
 'however little known the feelings or views of such a man may be on hi
s first entering a neighbourhood this truth is so well fixed in the min
ds of the surrounding families that he is considered as the rightful pr
operty of some one or other of their daughters',
 'my dear mr bennet said his lady to him one day have you heard that ne
therfield park is let at last',
 'mr bennet replied that he had not',
 'but it is returned she for mrs long has just been here and she told m
e all about it']
```

Corpus 2

Sample string before treatment

In [76]:
```
sample_string = corpus_2_list[1000]
sample_string
```

Out[76]:
```
'From: zowie@daedalus.stanford.edu (Craig "Powderkeg" DeForest)\nSubjec
t: Re: 5W30, 10W40, or 20W50\nArticle-I.D.: daedalus.ZOWIE.93Apr5215616
\nOrganization: Stanford Center for Space Science and Astrophysics\nLin
es: 37\nNNTP-Posting-Host: daedalus.stanford.edu\nIn-reply-to: Brad Tho
ne\'s message of Fri, 02 Apr 93 21:41:53 CST\n\nIn article <foo> Brad T
hone <C09615BT@WUVMD> writes:\nWell, there *is* a difference.\n\nI don
\'t happen to have my SAE manual handy, but oil viscosity in general\n_
decreases_ with temperature.  The SAE numbers are based on a `typical
\'\ncurve that oils used to all have, running from (say) the viscosity
of a\nroom-temperature 90-weight at 0C, down to (say) that of a room-te
mperature \n5-weight at 20C, for a typical 40-weight oil.\n\nOils that
are designed for operation in `normal\' temperatures just have\na weigh
t specification.  Oils that are designed for operation in exceedingly\n
cold temperatures have a `W\' tacked on the end, so in winter in a cold
\nplace, you\'d stick 10W in your car in the winter and 40 in it in the
summer,\nto approximate the appropriate viscosity throughout the yea
r.\n\nModern multi-viscosity oils change viscosity much less with tempe
rature.\nAs a result, their viscosity graphs cross over several curves.
A multi-vis\nspecification pegs the curve at two temperatures, a `norma
l\' operating\ntemperature and a `cold\' one (though I can\'t remember
the numbers...).\n\nIn any event, the weights do indicate a significant
difference.  Remember\nthat your engine is temperature-regulated (by th
e thermostat and\nradiator or air fins) most of the time -- unless you
overheat it or\nsomething.\n\nAny weight of oil is better than no oil,
or than very old, carbonized\noil.  Thin oil won\'t (in general) lubric
ate as well at temperature,\nthicker oil will (like a 20W50) will lubri
cate better at temperature, \nbut not as well during startup, when most
engine wear occurs.  \n\nIf you\'re planning on making long drives, the
20W50 is probably fine\n(esp. in the summer) in your 10W40 car.  But if
you\'re making short drives,\nstick to the 10W40.\n\n\n--\nDON\'T DRINK
SOAP! DILUTE DILUTE! OK!\n'
```

Removing blank spaces and new lines

In [77]:
```
step_1 = re.sub('\s+', ' ', sample_string)
step_1
```

Out[77]:
```
'From: zowie@daedalus.stanford.edu (Craig "Powderkeg" DeForest) Subjec
t: Re: 5W30, 10W40, or 20W50 Article-I.D.: daedalus.ZOWIE.93Apr5215616
Organization: Stanford Center for Space Science and Astrophysics Lines:
37 NNTP-Posting-Host: daedalus.stanford.edu In-reply-to: Brad Thone\'s
message of Fri, 02 Apr 93 21:41:53 CST In article <foo> Brad Thone <C09
615BT@WUVMD> writes: Well, there *is* a difference. I don\'t happen to
have my SAE manual handy, but oil viscosity in general _decreases_ with
temperature. The SAE numbers are based on a `typical\' curve that oils
used to all have, running from (say) the viscosity of a room-temperatur
e 90-weight at 0C, down to (say) that of a room-temperature 5-weight at
20C, for a typical 40-weight oil. Oils that are designed for operation
in `normal\' temperatures just have a weight specification. Oils that a
re designed for operation in exceedingly cold temperatures have a `W\'
tacked on the end, so in winter in a cold place, you\'d stick 10W in yo
ur car in the winter and 40 in it in the summer, to approximate the app
ropriate viscosity throughout the year. Modern multi-viscosity oils cha
nge viscosity much less with temperature. As a result, their viscosity
graphs cross over several curves. A multi-vis specification pegs the cu
rve at two temperatures, a `normal\' operating temperature and a `cold
\' one (though I can\'t remember the numbers...). In any event, the wei
ghts do indicate a significant difference. Remember that your engine is
temperature-regulated (by the thermostat and radiator or air fins) most
of the time -- unless you overheat it or something. Any weight of oil i
s better than no oil, or than very old, carbonized oil. Thin oil won\'t
(in general) lubricate as well at temperature, thicker oil will (like a
20W50) will lubricate better at temperature, but not as well during sta
rtup, when most engine wear occurs. If you\'re planning on making long
drives, the 20W50 is probably fine (esp. in the summer) in your 10W40 c
ar. But if you\'re making short drives, stick to the 10W40. -- DON\'T D
RINK SOAP! DILUTE DILUTE! OK! '
```

Removing From/To and subject line of email

Finding from to section

In [78]:
```
re.findall(r'\bFrom: .*? writes: ', step_1)
```

Out[78]:
```
['From: zowie@daedalus.stanford.edu (Craig "Powderkeg" DeForest) Subjec
t: Re: 5W30, 10W40, or 20W50 Article-I.D.: daedalus.ZOWIE.93Apr5215616
Organization: Stanford Center for Space Science and Astrophysics Lines:
37 NNTP-Posting-Host: daedalus.stanford.edu In-reply-to: Brad Thone\'s
message of Fri, 02 Apr 93 21:41:53 CST In article <foo> Brad Thone <C09
615BT@WUVMD> writes: ']
```

In [79]:
```
step_2 = re.sub(r'\bFrom: .*? writes: ', '', step_1)
step_2
```

Out[79]:
```
"Well, there *is* a difference. I don't happen to have my SAE manual ha
ndy, but oil viscosity in general _decreases_ with temperature. The SAE
numbers are based on a `typical' curve that oils used to all have, runn
ing from (say) the viscosity of a room-temperature 90-weight at 0C, dow
n to (say) that of a room-temperature 5-weight at 20C, for a typical 40
-weight oil. Oils that are designed for operation in `normal' temperatu
res just have a weight specification. Oils that are designed for operat
ion in exceedingly cold temperatures have a `W' tacked on the end, so i
n winter in a cold place, you'd stick 10W in your car in the winter and
40 in it in the summer, to approximate the appropriate viscosity throug
hout the year. Modern multi-viscosity oils change viscosity much less w
ith temperature. As a result, their viscosity graphs cross over several
curves. A multi-vis specification pegs the curve at two temperatures, a
`normal' operating temperature and a `cold' one (though I can't remembe
r the numbers...). In any event, the weights do indicate a significant
difference. Remember that your engine is temperature-regulated (by the
thermostat and radiator or air fins) most of the time -- unless you ove
rheat it or something. Any weight of oil is better than no oil, or than
very old, carbonized oil. Thin oil won't (in general) lubricate as well
at temperature, thicker oil will (like a 20W50) will lubricate better a
t temperature, but not as well during startup, when most engine wear oc
curs. If you're planning on making long drives, the 20W50 is probably f
ine (esp. in the summer) in your 10W40 car. But if you're making short
drives, stick to the 10W40. -- DON'T DRINK SOAP! DILUTE DILUTE! OK! "
```

Removal of digits

In [80]:
```
# Remove all digits
step_3 = re.sub(r'\d+', '', step_2)
step_3
```

Out[80]:
```
"Well, there *is* a difference. I don't happen to have my SAE manual ha
ndy, but oil viscosity in general _decreases_ with temperature. The SAE
numbers are based on a `typical' curve that oils used to all have, runn
ing from (say) the viscosity of a room-temperature -weight at C, down t
o (say) that of a room-temperature -weight at C, for a typical -weight
oil. Oils that are designed for operation in `normal' temperatures just
have a weight specification. Oils that are designed for operation in ex
ceedingly cold temperatures have a `W' tacked on the end, so in winter
in a cold place, you'd stick W in your car in the winter and  in it in
the summer, to approximate the appropriate viscosity throughout the yea
r. Modern multi-viscosity oils change viscosity much less with temperat
ure. As a result, their viscosity graphs cross over several curves. A m
ulti-vis specification pegs the curve at two temperatures, a `normal' o
perating temperature and a `cold' one (though I can't remember the numb
ers...). In any event, the weights do indicate a significant differenc
e. Remember that your engine is temperature-regulated (by the thermosta
t and radiator or air fins) most of the time -- unless you overheat it
or something. Any weight of oil is better than no oil, or than very ol
d, carbonized oil. Thin oil won't (in general) lubricate as well at tem
perature, thicker oil will (like a W) will lubricate better at temperat
ure, but not as well during startup, when most engine wear occurs. If y
ou're planning on making long drives, the W is probably fine (esp. in t
he summer) in your W car. But if you're making short drives, stick to t
he W. -- DON'T DRINK SOAP! DILUTE DILUTE! OK! "
```

Remove futile symbols (non punctuation)

In [81]:
```
# Remove other signs
step_4 = re.sub(r'["_“”\'\`\-\*\(\)]','',step_3)
step_4
```

Out[81]:
```
'Well, there is a difference. I dont happen to have my SAE manual hand
y, but oil viscosity in general decreases with temperature. The SAE num
bers are based on a typical curve that oils used to all have, running f
rom say the viscosity of a roomtemperature weight at C, down to say tha
t of a roomtemperature weight at C, for a typical weight oil. Oils that
are designed for operation in normal temperatures just have a weight sp
ecification. Oils that are designed for operation in exceedingly cold t
emperatures have a W tacked on the end, so in winter in a cold place, y
oud stick W in your car in the winter and  in it in the summer, to appr
oximate the appropriate viscosity throughout the year. Modern multivisc
osity oils change viscosity much less with temperature. As a result, th
eir viscosity graphs cross over several curves. A multivis specificatio
n pegs the curve at two temperatures, a normal operating temperature an
d a cold one though I cant remember the numbers.... In any event, the w
eights do indicate a significant difference. Remember that your engine
is temperatureregulated by the thermostat and radiator or air fins most
of the time  unless you overheat it or something. Any weight of oil is
better than no oil, or than very old, carbonized oil. Thin oil wont in
general lubricate as well at temperature, thicker oil will like a W wil
l lubricate better at temperature, but not as well during startup, when
most engine wear occurs. If youre planning on making long drives, the W
is probably fine esp. in the summer in your W car. But if youre making
short drives, stick to the W.  DONT DRINK SOAP! DILUTE DILUTE! OK! '
```

Tokenize to sentences while punctuation is still in place

In [82]:
```
# Tokenize
tokens_news = sent_tokenize(step_4)
tokens_news
```

Out[82]:
```
['Well, there is a difference.',
 'I dont happen to have my SAE manual handy, but oil viscosity in gener
al decreases with temperature.',
 'The SAE numbers are based on a typical curve that oils used to all ha
ve, running from say the viscosity of a roomtemperature weight at C, do
wn to say that of a roomtemperature weight at C, for a typical weight o
il.',
 'Oils that are designed for operation in normal temperatures just have
a weight specification.',
 'Oils that are designed for operation in exceedingly cold temperatures
have a W tacked on the end, so in winter in a cold place, youd stick W
in your car in the winter and  in it in the summer, to approximate the
appropriate viscosity throughout the year.',
 'Modern multiviscosity oils change viscosity much less with temperatur
e.',
 'As a result, their viscosity graphs cross over several curves.',
 'A multivis specification pegs the curve at two temperatures, a normal
operating temperature and a cold one though I cant remember the number
s....',
 'In any event, the weights do indicate a significant difference.',
 'Remember that your engine is temperatureregulated by the thermostat a
nd radiator or air fins most of the time  unless you overheat it or som
ething.',
 'Any weight of oil is better than no oil, or than very old, carbonized
oil.',
 'Thin oil wont in general lubricate as well at temperature, thicker oi
l will like a W will lubricate better at temperature, but not as well d
uring startup, when most engine wear occurs.',
 'If youre planning on making long drives, the W is probably fine es
p.',
 'in the summer in your W car.',
 'But if youre making short drives, stick to the W.  DONT DRINK SOAP!',
 'DILUTE DILUTE!',
 'OK!']
```

Convert each token to lowercase

In [83]:
```
# Lowercase
lower_token = list(map(lambda token: token.lower(), tokens_news))
lower_token[0:5]
```

Out[83]:
```
['well, there is a difference.',
 'i dont happen to have my sae manual handy, but oil viscosity in gener
al decreases with temperature.',
 'the sae numbers are based on a typical curve that oils used to all ha
ve, running from say the viscosity of a roomtemperature weight at c, do
wn to say that of a roomtemperature weight at c, for a typical weight o
il.',
 'oils that are designed for operation in normal temperatures just have
a weight specification.',
 'oils that are designed for operation in exceedingly cold temperatures
have a w tacked on the end, so in winter in a cold place, youd stick w
in your car in the winter and  in it in the summer, to approximate the
appropriate viscosity throughout the year.']
```

Remove punctuation from lowercase token

In [84]:
```
punct_less_token = list(map(lambda token:
                            token.translate(str.maketrans('', '', string
.punctuation)), lower_token))
punct_less_token[0:5]
```

Out[84]:
```
['well there is a difference',
 'i dont happen to have my sae manual handy but oil viscosity in genera
l decreases with temperature',
 'the sae numbers are based on a typical curve that oils used to all ha
ve running from say the viscosity of a roomtemperature weight at c down
to say that of a roomtemperature weight at c for a typical weight oil',
 'oils that are designed for operation in normal temperatures just have
a weight specification',
 'oils that are designed for operation in exceedingly cold temperatures
have a w tacked on the end so in winter in a cold place youd stick w in
your car in the winter and  in it in the summer to approximate the appr
opriate viscosity throughout the year']
```

Word2Vec model

Take clean sentence tokens list, and convert to word tokens list of list

In [85]:
```
word_tokens_list_of_list = []

for news_story in corpus_2_list:
    tokenized_story = preprocessing_corpus_2(news_story)
    for sentence in tokenized_story:
        word_tokens_list = word_tokenize(sentence)
        word_tokens_list_of_list.append(word_tokens_list)

word_tokens_list_of_list[0:2]
```

Out[85]:
```
[['lebanese',
  'resistance',
  'forces',
  'detonated',
  'a',
  'bomb',
  'under',
  'an',
  'israeli',
  'occupation',
  'patrol',
  'in',
  'lebanese',
  'territory',
  'two',
  'days',
  'ago'],
 ['three', 'soldiers', 'were', 'killed', 'and', 'two', 'wounded']]
```

Build models

In [86]:
```
def build_word2vec(list_of_list, dimension_size, window_size, min_obs, m
odel_type, model_name):

    """
    Creates a model object
    Args:
        list_of_list (list): preprocessed text corpus
        dimension_size (int): size of dimensions in model
        window_size (int): window size used for training
        min_count (int): minimum observed instances of a word to be cons
idered
        model_type (binary 1 or 0): 1 = skipgram, 0 = CBOW
        model_name: name of object
    Returns:
        A trained word2vec model, that is saved as an object
    """

    new_model = gensim.models.Word2Vec(list_of_list, vector_size = dimen
sion_size, window = window_size,
                                       sg = model_type, min_count = min_
obs)
    new_model.save(model_name)

    return new_model
```

Create Skipgram model - with different vector and window parameters

In [87]:
```
model_sg_50_3 = build_word2vec(word_tokens_list_of_list, 50, 3, 5, 1, "m
odel_sg_50_3")
model_sg_50_5 = build_word2vec(word_tokens_list_of_list, 50, 5, 5, 1, "m
odel_sg_50_5")
model_sg_50_7 = build_word2vec(word_tokens_list_of_list, 50, 7, 5, 1, "m
odel_sg_50_7")
model_sg_100_3 = build_word2vec(word_tokens_list_of_list, 100, 3, 5, 1,
"model_sg_100_3")
model_sg_100_5 = build_word2vec(word_tokens_list_of_list, 100, 5, 5, 1,
"model_sg_100_5")
model_sg_100_7 = build_word2vec(word_tokens_list_of_list, 100, 7, 5, 1,
"model_sg_100_7")
model_sg_200_3 = build_word2vec(word_tokens_list_of_list, 200, 3, 5, 1,
"model_sg_200_3")
model_sg_200_5 = build_word2vec(word_tokens_list_of_list, 200, 5, 5, 1,
"model_sg_200_5")
model_sg_200_7 = build_word2vec(word_tokens_list_of_list, 200, 7, 5, 1,
"model_sg_200_7")
```

Create CBOW model - with different vector and window parameters

In [88]:
```
model_cbow_50_3 = build_word2vec(word_tokens_list_of_list, 50, 3, 5, 1,
"model_cbow_50_3")
model_cbow_50_5 = build_word2vec(word_tokens_list_of_list, 50, 5, 5, 1,
"model_cbow_50_5")
model_cbow_50_7 = build_word2vec(word_tokens_list_of_list, 50, 7, 5, 1,
"model_cbow_50_7")
model_cbow_100_3 = build_word2vec(word_tokens_list_of_list, 100, 3, 5, 1
, "model_cbow_100_3")
model_cbow_100_5 = build_word2vec(word_tokens_list_of_list, 100, 5, 5, 1
, "model_cbow_100_5")
model_cbow_100_7 = build_word2vec(word_tokens_list_of_list, 100, 7, 5, 1
, "model_cbow_100_7")
model_cbow_200_3 = build_word2vec(word_tokens_list_of_list, 200, 3, 5, 1
, "model_cbow_200_3")
model_cbow_200_5 = build_word2vec(word_tokens_list_of_list, 200, 5, 5, 1
, "model_cbow_200_5")
model_cbow_200_7 = build_word2vec(word_tokens_list_of_list, 200, 7, 5, 1
, "model_cbow_200_7")
```

In [89]:
```
print(str(model_cbow_200_7.__repr__()))
```

```
<gensim.models.word2vec.Word2Vec object at 0x7f953f3cc890>
```

Evaluate models based on identified nearest neighbors words

In [90]:
```
evaluation_list = ['government', 'army', 'happy', 'food', 'pride',
                   'wealth', 'overwhelming', 'education','family','compu
ter']
```

In [91]:
```
def evaluate_model_nn(model_object, index_model_name, evaluation_word_li
sts, top_n):

    """
    Identifies top n nearest words for words in a lit
    Args:
        model_object (obj): word2vec model
        index_model_name (str): name of model for row indices naming
        evaluation_word_lists (list): words to be used as reference to i
dentify nearest neighbors
        top_n (int): top n number of nearest neighbors
    Returns:
        A pd dataframe summary
    """


    # Create blank df
    results_df = pd.DataFrame()

    for i, word in enumerate(evaluation_word_lists):

        result = model_object.wv.most_similar(word, topn=top_n)
        results_df[word] = [result]

    results_df.index = [str(index_model_name)]

    return results_df
```

Evaluating all models at once

In [92]:
```
list_model_objects = [model_sg_50_3, model_sg_50_5, model_sg_50_7,
                      model_sg_100_3, model_sg_100_5, model_sg_100_7,
                      model_sg_200_3, model_sg_200_5, model_sg_200_7,
                     model_cbow_50_3, model_cbow_50_5, model_cbow_50_7,
                      model_cbow_100_3, model_cbow_100_5, model_cbow_100
_7,
                      model_cbow_200_3, model_cbow_200_5, model_cbow_200
_7]
list_index_names = ['sg_50_3', 'sg_50_5', 'sg_50_7',
                      'sg_100_3', 'sg_100_5', 'sg_100_7',
                      'sg_200_3', 'sg_200_5', 'sg_200_7',
                     'cbow_50_3', 'cbow_50_5', 'cbow_50_7',
                      'cbow_100_3', 'cbow_100_5', 'cbow_100_7',
                      'cbow_200_3', 'cbow_200_5', 'cbow_200_7']

agg_results = evaluate_model_nn(list_model_objects[0], list_index_names[
0], evaluation_list, 3)

for i in range(1,18):
    # Add row to summary table
    new_row = pd.DataFrame(evaluate_model_nn(list_model_objects[i], list
_index_names[i], evaluation_list, 3))
    agg_results = agg_results.append(new_row)
```

In [93]:
```
agg_results
```

Out[93]:

	government	army	happy	fo
sg_50_3	[(greek, 0.804038941860199), (policies, 0.7906...	[(ottoman, 0.8858816623687744), (forces, 0.885...	[(surprised, 0.9102444648742676), (proud, 0.90...	[(trucl 0.927154660224914 (minimal, 0.917
sg_50_5	[(greek, 0.8535125255584717), (genocide, 0.847...	[(russian, 0.8825209140777588), (dictatorship,...	[(careful, 0.8871726989746094), (comfortable, ...	[(frienc 0.907289803028106 (wives, 0.903
sg_50_7	[(xsoviet, 0.836116373538971), (removing, 0.81...	[(dictatorship, 0.9182860851287842), (organize...	[(careful, 0.8877199292182922), (confusing, 0....	[(wive 0.888133764266967 (clothing, 0.885
sg_100_3	[(greek, 0.7724157571792603), (xsoviet, 0.7724...	[(forces, 0.8722518086433411), (azeri, 0.87006...	[(surprised, 0.9026287794113159), (glad, 0.893...	[(trucl 0.933631002902984 (jobs, 0.915132
sg_100_5	[(xsoviet, 0.7944872975349426), (greek, 0.7880...	[(dictatorship, 0.9032689929008484), (moslem, ...	[(confusing, 0.8958973288536072), (careful, 0....	[(babi 0.892612159252166 (clothing, 0.88
sg_100_7	[(fascist, 0.8196161985397339), (xsoviet, 0.81...	[(dictatorship, 0.8783590197563171), (slaughte...	[(bet, 0.879052996635437), (glad, 0.8663113117...	[(clothir 0.876994669437408 (tenant, 0.87
sg_200_3	[(genocide, 0.7920501828193665), (forces, 0.77...	[(moslem, 0.8863533735275269), (azeri, 0.88231...	[(confusing, 0.9133730530738831), (lucky, 0.90...	[(babi 0.91150486469268 (clothing, 0.906
sg_200_5	[(greek, 0.8036022186279297), (xsoviet, 0.7999...	[(moslem, 0.896993100643158), (dictatorship, 0...	[(confusing, 0.8820159435272217), (careful, 0....	[(bu 0.893142938613891 (howling, 0.88812
sg_200_7	[(xsoviet, 0.807157576084137), (crime, 0.78833...	[(dictatorship, 0.8981454372406006), (moslem, ...	[(careful, 0.8562300205230713), (honest, 0.854...	[(clothir 0.901617705821999 (babies, 0.87
cbow_50_3	[(greek, 0.7928274869918823), (territories, 0....	[(forces, 0.8862621188163757), (azeri, 0.88594...	[(confusing, 0.9049172401428223), (comfortable...	[(trucl 0.91998636722564 (minimal, 0.913
cbow_50_5	[(xsoviet, 0.8273578882217407), (fascist, 0.80...	[(dictatorship, 0.9162772297859192), (organize...	[(willing, 0.8881098628044128), (comfortable, ...	[(wive 0.909685790538787 (knives, 0.9078
cbow_50_7	[(xsoviet, 0.8323005437850952), (greek, 0.8192...	[(dictatorship, 0.8901408910751343), (forces, ...	[(careful, 0.8757197856903076), (enjoy, 0.8647...	[(wive 0.899152994155883 (clothing, 0.897
cbow_100_3	[(xsoviet, 0.7732653617858887), (greek, 0.7655...	[(azeri, 0.8809871673583984), (ottoman, 0.8758...	[(surprised, 0.9032120108604431), (confusing, ...	[(trucl 0.919551074504852 (jobs, 0.90845
cbow_100_5	[(xsoviet, 0.8036972880363464), (greek, 0.7835...	[(dictatorship, 0.8955327868461609), (forces, ...	[(comfortable, 0.8726111054420471), (glad, 0.8...	[(axe 0.894328355789184 (wives, 0.893800
cbow_100_7	[(xsoviet, 0.7987973690032959), (removing, 0.7...	[(dictatorship, 0.8958802819252014), (moslem, ...	[(careful, 0.8588906526565552), (glad, 0.85569...	[(clothir 0.881793498992919 (burn, 0.878
cbow_200_3	[(genocide, 0.7881043553352356), (forces, 0.78...	[(russian, 0.8871954083442688), (forces, 0.869...	[(careful, 0.9035475850105286), (lucky, 0.9031...	[(few 0.928062796592712 (jobs, 0.927287
cbow_200_5	[(xsoviet, 0.8066669702529907), (movement, 0.7...	[(moslem, 0.890616774559021), (dictatorship, 0...	[(comfortable, 0.9116938710212708), (careful, ...	[(howlir 0.90565955638885 (clothing, 0.89

	government	army	happy	fo
cbow_200_7	[(xsoviet, 0.8384032249450684), (greece, 0.814...	[(dictatorship, 0.8854653239250183), (moslem, ...	[(confusing, 0.8708259463310242), (strongly, 0...	[(defenceles 0.877611339092254 (passports

In [94]:
```
agg_results.to_csv('results_test.csv')
```