Assignment 3

Distributional Semantics and POS Tagging

Submitted by:

Name: Adithya Avvaru

Roll No: 20162116

Latent Symantic Analysis

In [54]:

```
import io
from pprint import pprint
from time import time
def get_unigrams(file_name):
    unigrams = {}
    with io.open(file_name, encoding='utf8', errors='ignore') as f:
        for line in f:
            tokens = line.strip().split()

            # Added to extract only Verbs
            # verbs = nltk.pos_tag(tokens)
            # tokens = set([x for x,y in verbs if "VB" in y])

            for token in tokens:
                token = token.lower()
                try:
                    unigrams[token]
                except:
                    unigrams[token] = 0
                unigrams[token] += 1

    return unigrams

def index_unigrams(unigrams):
    new_unigrams = {}
    reverse_unigrams = {}
    for index, unigram in enumerate(unigrams):
        new_unigrams[unigram] = index
        reverse_unigrams[index] = unigram
    return new_unigrams, reverse_unigrams
```

In [55]:

```
file_name = "sample_corpus.txt"
unigrams = get_unigrams(file_name)
iunigrams,runigrams = index_unigrams(unigrams)
unigrams = sorted(unigrams.items(), key = lambda x: x[1], reverse = True )

#pprint(unigrams) # Figure out non-stop words
dimensions = [x[0] for x in unigrams[100:3100]]
idimensions = {x: index for index, x in enumerate(dimensions)}
```

In [56]:

```
import numpy
import nltk
from nltk import word_tokenize

def populate_cmatrix(file_name,iunigrams, dimensions, window, leftonly, rig
    e = 0
    s = 0
    cmatrix = numpy.memmap("lsa.cmatrix", dtype='float32', mode='w+', shape
    with open(file_name, encoding='utf-8', errors='ignore') as f:
        for index, line in enumerate(f):
            tokens = line.strip().split()
            for indexj, token in enumerate(tokens):
                token = token.lower()
                lcontext = rcontext = ""
                if leftonly :
                    lcontext = tokens[indexj - window:indexj]
                else:
                    lcontext = []
                if rightonly:
                    rcontext = tokens[indexj + 1:index + window]
                else:
                    rcontext = []
                context = [tok.lower() for tok in lcontext + rcontext]

                #verbs = nltk.pos_tag(context)
                #context = set([x for x,y in verbs if "VB" in y])

                try:
                    unigram_index = iunigrams[token]
                    for d in context:
                        #print(nltk.pos_tag([d]))

                        if d in dimensions:
                            j = dimensions[d]
                            cmatrix[unigram_index][j] += 1
                            s += 1
                except:
                    e += 1
    #print(e,s)
    return cmatrix
```

In [57]:

```
from scipy.spatial.distance import *
from numpy import linalg as LA
import numpy as np
from sklearn.decomposition import TruncatedSVD
```

In [61]:

```
def getDistance(twod_cmatrix):
    words = ["boy","sunday","eat","good","slowly","100"]
    for w1 in words:
        distance = {}
        for w2 in idimensions:
            if w1 == w2:
                continue
            id1 = iunigrams[w1]
            id2 = iunigrams[w2]
            v1, v2 = twod_cmatrix[id1], twod_cmatrix[id2]
            if np.linalg.norm(v2) == 0:
                continue
            distance[w2] = cosine(v1,v2)
        sortedDistance = sorted(distance.items(), key = lambda x : x[1], re

        temp = dict((x, y) for x, y in sortedDistance[:10])  # For top 10 e

        print(w1)
        print(list(temp.keys()))
```

In [62]:

```
def run(windowSize=2, noComp=10, leftOnly=True, rightOnly=True):
    s = time()
    cmatrix = populate_cmatrix(file_name, iunigrams, idimensions, window = 
                               leftonly = leftOnly, rightonly = rightOnly)
    svd = TruncatedSVD(n_components = noComp, random_state=42)
    svd.fit(cmatrix)
    twod_cmatrix = svd.transform(cmatrix)

    print("Window :",windowSize,"; No components :", noComp, "; Left only :
    print("------------------------------------------------------------------
    print("Time taken is ---- ", (time()-s))
    getDistance(twod_cmatrix)
```

In [63]:

```
run(windowSize=2, noComp=10)
```

```
Window : 2 ; No components : 10 ; Left only : True ; Right only : True
-----------------------------------------------------------------------
Time taken is ----  50.29086780548096
boy
['attend', 'adobe', 'minutes', 'nursing', 'forced', 'birds', 'miles', 'lei
sure', 'speaker', 'supplied']
sunday
['17th', 'seminar', 'routes', 'supported', 'ford', 'morning', 'saturday',
'evening', 'director', '1st']
eat
['am', 'sit', 'heavy', "'ll", "'ve", "'re", 'me', 'think', 'saying', 'sorr
y']
good
['though', 'him', 'thanks', 'often', 'god', 'see', 'always', 'take', 'went
', 'nor']
slowly
['parents', 'developed', 'clear', 'positive', 'brand', 'residents', 'meant
', 'helpful', 'understanding', 'illness']
100
['pp', 'hotels', 'column', '90', 'cm', 'white', '1000', '=', 'mm', 'black'
]
```

In [66]:

```
run(windowSize=2, noComp=50)
```

```
Window : 2 ; No components : 50 ; Left only : True ; Right only : True
-----------------------------------------------------------------------
Time taken is ----  58.586092472076416
boy
['him', 'dear', 'saw', 'green', 'dream', 'sought', 'sea', 'daughter', 'mot
her', 'wood']
sunday
['th', 'june', '11th', 'september', '4th', 'acts', 'morning', 'saturday',
'evening', '1st']
eat
['sleep', 'good', "'ll", 'just', 'did', 'try', 'looks', 'saying', 'fun', '
bit']
good
['yet', 'great', 'even', 'getting', 'because', 'said', 'way', 'really', 'w
atching', 'again']
slowly
['becomes', 'around', 'ice', 'red', 'far', 'bird', 'unfortunately', 'idea'
, 'exposure', 'brand']
100
['capacity', 'length', 'column', '1000', 'cm', 'document', '1', 'mm', '=',
'10']
```

In [67]:
```
run(windowSize=2, noComp=100)
```

```
Window : 2 ; No components : 100 ; Left only : True ; Right only : True
-----------------------------------------------------------------------
Time taken is ----  54.68529272079468
boy
['him', 'saw', 'dear', 'arms', 'green', 'dream', 'wood', 'sea', 'daughter'
, 'sought']
sunday
['bible', 'evening', 'saturday', '11th', '4th', 'acts', 'morning', 'john',
'june', '1st']
eat
['just', "'re", 'think', 'bit', 'did', 'try', 'looks', 'lot', 'fun', 'thin
g']
good
['yet', 'saw', "'ll", 'again', 'well', 'make', 'really', 'got', 'said', 'w
ay']
slowly
['properly', 'try', 'fact', 'unfortunately', 'fit', 'real', 'strength', 'w
atching', 'seem', 'enough']
100
['capacity', 'pp', 'length', '1000', 'document', 'cm', '1', 'mm', '=', '10
']
```

In [68]:
```
run(windowSize=2, noComp=200)
```

```
Window : 2 ; No components : 200 ; Left only : True ; Right only : True
-----------------------------------------------------------------------
Time taken is ----  60.26724600791931
boy
['true', 'him', 'green', 'dear', 'dream', 'wood', 'sea', 'daughter', 'moth
er', 'sought']
sunday
['bible', 'evening', '2nd', '11th', '4th', 'acts', 'morning', 'saturday',
'june', '1st']
eat
['bit', "'ll", 'think', 'just', 'did', 'feel', 'good', 'saying', 'fun', 't
hing']
good
['yet', 'saw', 'said', 'again', 'well', 'got', 'idea', 'really', 'true', "
'll"]
slowly
['try', 'around', 'fact', 'enough', 'near', 'unfortunately', 'another', 'h
owever', 'worker', 'take']
100
['capacity', 'volume', 'length', '1000', 'document', '1', 'mm', 'cm', '=',
'10']
```

In [69]:
```
run(windowSize=5, noComp=10)
```

```
Window : 5 ; No components : 10 ; Left only : True ; Right only : True
-----------------------------------------------------------------------
Time taken is ----  54.401859283447266
boy
['reader', 'minutes', 'forced', 'birds', 'miles', 'rain', 'nursing', 'spea
ker', 'chair', 'ian']
sunday
['component', '17th', 'seminar', 'routes', 'ford', 'morning', 'saturday',
'director', 'supplied', '1st']
eat
['happen', 'worry', 'think', 'wanted', "'ve", "'re", 'me', 'saying', 'sit'
, "'ll"]
good
['though', 'him', 'getting', 'god', 'always', 'might', 'miss', 'right', 'w
ent', 'fear']
slowly
['encouraged', 'surely', 'becomes', 'turning', 'vast', 'unfortunately', 'm
eant', 'confidence', 'helpful', 'understanding']
100
['n', 'hotels', 'column', 'mm', 'black', 'cm', 'white', '1000', '=', '90']
```

In [70]:
```
run(windowSize=5, noComp=50)
```

```
Window : 5 ; No components : 50 ; Left only : True ; Right only : True
-----------------------------------------------------------------------
Time taken is ----  58.407567262649536
boy
['him', 'drink', 'dear', 'saw', 'green', 'dream', 'wood', 'saying', 'mothe
r', 'sought']
sunday
['bible', 'th', 'june', '11th', '4th', 'acts', 'morning', 'saturday', 'eve
ning', '1st']
eat
['just', 'worry', 'did', "'ll", 'lot', 'try', "'re", 'saying', 'moment', '
again']
good
['because', 'even', 'getting', 'might', 'just', 'idea', 'really', 'watchin
g', 'again', 'way']
slowly
['means', 'simply', 'right', 'point', 'unfortunately', 'idea', 'takes', 'a
tmosphere', 'situation', 'probably']
100
['capacity', 'length', 'column', '1000', 'cm', 'document', '1', 'mm', '=',
'10']
```

In [71]:

```
run(windowSize=5, noComp=100)
```

```
Window : 5 ; No components : 100 ; Left only : True ; Right only : True
------------------------------------------------------------------------
Time taken is ----  59.312278032302856
boy
['him', 'green', 'arms', 'saw', 'dear', 'dream', 'wood', 'saying', 'daught
er', 'sought']
sunday
['bible', 'evening', 'saturday', '11th', '4th', 'acts', 'morning', 'john',
'june', '1st']
eat
['just', "'re", 'think', 'try', 'did', 'got', 'lot', 'again', 'thing', 'bi
t']
good
['yet', 'make', 'little', "'ll", 'well', 'just', 'got', 'way', 'said', 'ag
ain']
slowly
['try', 'feeling', 'fact', 'watching', 'fit', 'real', 'enough', 'takes', '
right', 'take']
100
['capacity', 'volume', 'length', '1000', 'document', '1', 'mm', 'cm', '=',
'10']
```

In [72]:

```
run(windowSize=5, noComp=200)
```

```
Window : 5 ; No components : 200 ; Left only : True ; Right only : True
------------------------------------------------------------------------
Time taken is ----  63.86641335487366
boy
['him', 'green', 'saw', 'dear', 'dream', 'wood', 'sea', 'daughter', 'mothe
r', 'sought']
sunday
['bible', 'evening', '2nd', '11th', '4th', 'acts', 'morning', 'saturday',
'june', '1st']
eat
["'ll", 'thing', 'lot', 'again', 'just', 'did', 'got', 'saying', 'seemed',
'getting']
good
['yet', 'saw', 'again', 'well', 'though', 'really', 'got', 'idea', 'said',
"'ll"]
slowly
['make', 'move', 'even', 'fact', 'either', 'want', 'however', 'enough', 'b
roken', 'take']
100
['capacity', 'volume', 'length', '1000', 'document', '1', 'mm', 'cm', '=',
'10']
```

```
In [73]: run(windowSize=10, noComp=10)
```

```
Window : 10 ; No components : 10 ; Left only : True ; Right only : True
-----------------------------------------------------------------------
Time taken is ----  57.13647532463074
boy
['week', 'tv', 'little', 'mr.', 'finds', 'see', 'sea', 'headed', 'private'
, 'tea']
sunday
['component', 'seminar', 'routes', 'ford', 'psychology', 'morning', 'satur
day', '11th', 'dates', '1st']
eat
['thing', 'worry', "'ve", 'think', "'re", 'saying', 'someone', 'me', 'tell
', 'paying']
good
['though', 'him', 'saw', 'little', 'see', 'god', 'always', 'something', 'e
veryone', 'went']
slowly
['statutory', 'certain', 'understanding', 'confidence', 'developed', 'spea
k', 'respond', 'entirely', 'managers', 'compensation']
100
['black', 'n', 'column', 'mm', 'cm', 'irish', 'white', '1000', '=', '90']
```

```
In [74]: run(windowSize=10, noComp=50)
```

```
Window : 10 ; No components : 50 ; Left only : True ; Right only : True
-----------------------------------------------------------------------
Time taken is ----  64.65309500694275
boy
['him', 'green', 'took', 'dear', 'dream', 'sought', 'sea', 'daughter', 'mo
ther', 'wood']
sunday
['bible', 'th', 'june', '11th', '4th', 'acts', 'morning', 'saturday', 'eve
ning', '1st']
eat
['just', 'worry', "'ll", 'stop', 'bit', "'re", 'good', 'lot', 'fun', 'agai
n']
good
['something', 'yet', 'even', 'getting', 'long', 'idea', 'really', 'watchin
g', 'while', 'way']
slowly
['means', 'becomes', 'fact', 'whole', 'positive', 'negative', 'real', 'get
s', 'atmosphere', 'situation']
100
['capacity', 'length', 'column', '1000', 'cm', 'document', '1', 'mm', '=',
'10']
```

In [75]:

```
run(windowSize=10, noComp=100)
```

```
Window : 10 ; No components : 100 ; Left only : True ; Right only : True
-----------------------------------------------------------------------
Time taken is ----  63.01830506324768
boy
['him', 'green', 'saw', 'dear', 'dream', 'wood', 'sea', 'daughter', 'mothe
r', 'sought']
sunday
['bible', 'evening', 'saturday', '11th', '4th', 'acts', 'morning', 'john',
'june', '1st']
eat
["'re", 'just', 'again', 'think', 'bit', 'did', 'too', 'lot', 'fun', 'gett
ing']
good
['yet', 'make', 'again', 'well', 'little', 'just', 'long', 'said', 'while'
, 'way']
slowly
['try', 'feeling', 'fact', 'whole', 'positive', 'fit', 'real', 'spot', 'di
fficulties', 'take']
100
['capacity', 'volume', 'length', '1000', 'document', '1', 'mm', 'cm', '=',
'10']
```

In [76]:

```
run(windowSize=10, noComp=200)
```

```
Window : 10 ; No components : 200 ; Left only : True ; Right only : True
-----------------------------------------------------------------------
Time taken is ----  70.83616185188293
boy
['saw', 'green', 'dear', 'him', 'mother', 'dream', 'wood', 'sea', 'daughte
r', 'sought']
sunday
['bible', 'evening', '11th', 'pub', '4th', 'acts', 'morning', 'saturday',
'june', '1st']
eat
['just', 'got', 'again', 'think', 'bit', 'did', 'feel', 'saying', 'fun', '
because']
good
['yet', 'saw', 'again', 'well', 'though', 'really', 'got', 'idea', 'said',
"'ll"]
slowly
['through', 'feeling', 'move', 'fact', 'whole', 'way', 'difficulties', 'br
ing', 'take', 'enough']
100
['capacity', 'volume', 'length', '1000', '1', 'document', 'mm', 'cm', '=',
'10']
```

In [64]:
```
run(windowSize=2, noComp=10, leftOnly=True, rightOnly=False)
```

```
Window : 2 ; No components : 10 ; Left only : True ; Right only : False
-----------------------------------------------------------------------
Time taken is ----  13.201389074325562
boy
['complex', 'leadership', 'affected', 'cities', 'tv', 'farm', 'customers',
'lies', 'private', 'income']
sunday
['initial', 'contains', 'except', 'discussed', 'gender', 'managed', 'host'
, 'indeed', 'involves', 'writers']
eat
['encouraged', 'carried', 'support', 'influence', 'recognise', 'white', 'r
elatively', 'run', 'require', 'owned']
good
['anything', 'doing', 'gone', 'going', 'safe', 'got', 'seems', 'getting',
'presented', 'another']
slowly
['transfer', 'close', 'control', 'foot', 'hundreds', 'thinking', 'rest', '
down', 'draft', 'routes']
100
['evolution', 'fax', 'africa', 'west', 'asia', 'g.', 'september', 'ireland
', 'japan', 'circuit']
```

In [77]:
```
run(windowSize=2, noComp=50, leftOnly=True, rightOnly=False)
```

```
Window : 2 ; No components : 50 ; Left only : True ; Right only : False
-----------------------------------------------------------------------
Time taken is ----  15.383311986923218
boy
['instance', 'exercise', 'party', 'case', 'weekly', 'day', 'farm', 'movie'
, 'aspect', 'player']
sunday
['simply', 'organised', 'appointed', 'qualified', 'added', 'henry', 'crisi
s', 'planned', 'established', 'brought']
eat
['affect', 'soul', 'default', 'mean', 'require', 'expect', 'draw', 'exist'
, 'seem', 'tell']
good
['great', 'little', 'perfect', 'much', 'extra', 'unique', 'own', 'watching
', 'clear', 'another']
slowly
['formal', 'reading', 'thanks', 'through', 'advice', 'quick', 'teach', 'sp
end', 'down', 'touch']
100
['pp', 'k', 'march', '5', '3', 'december', '7.', 'assessed', 'april', '10'
]
```

In [78]:
```
run(windowSize=2, noComp=100, leftOnly=True, rightOnly=False)
```

```
Window : 2 ; No components : 100 ; Left only : True ; Right only : False
-----------------------------------------------------------------------
Time taken is ----  17.163907289505005
boy
['truly', 'instance', 'piece', 'manufacturer', 'day', 'nation', 'linux', '
vision', 'sequence', 'aspect']
sunday
['simply', 'henry', 'qualified', 'placed', 'sometimes', 'maintained', 'hel
ping', 'organised', 'brought', 'crew']
eat
['come', 'affect', 'exist', 'default', 'mean', 'recognise', 'expect', 'dra
w', 'tell', 'seem']
good
['great', 'useful', 'particularly', 'quite', 'warm', 'much', 'happy', 'lon
g', 'strong', 'busy']
slowly
['mortgage', 'capital', 'feeling', 'killed', 'price', 'lost', 'chance', 't
hanks', 'touch', 'assist']
100
['pp', 'assessed', 'cm', 'philosophy', '3', 'legislation', '1000', '2', '1
0', '7.']
```

In [79]:
```
run(windowSize=2, noComp=200, leftOnly=True, rightOnly=False)
```

```
Window : 2 ; No components : 200 ; Left only : True ; Right only : False
-----------------------------------------------------------------------
Time taken is ----  21.108182668685913
boy
['truly', 'instance', 'piece', 'vision', 'knows', 'song', 'nation', 'day',
'sequence', 'aspect']
sunday
['talks', '17th', 'june', 'october', '19th', 'november', 'morning', 'septe
mber', 'july', 'january']
eat
['draw', 'scott', 'aged', 'young', 'speak', 'fantastic', 'publishing', 'ca
nnot', 'adults', 'expect']
good
['slightly', 'nice', 'accurate', 'remain', 'relatively', 'unique', 'useful
', 'excellent', 'significant', 'behind']
slowly
['feeling', 'cup', 'touch', 'killed', 'involved', 'summit', 'bank', 'thank
s', 'heritage', 'assist']
100
['pp', 'assessed', '1000', 'philosophy', 'cm', '2007', 'december', '2', '1
0', '7.']
```

In [80]:
```
run(windowSize=5, noComp=10, leftOnly=True, rightOnly=False)
```

```
Window : 5 ; No components : 10 ; Left only : True ; Right only : False
------------------------------------------------------------------------
Time taken is ----  16.049750804901123
boy
['yet', 'itself', 'rather', 'passed', 'green', 'gave', 'complaint', 'acces
sible', 'playing', 'gets']
sunday
['bible', 'soul', 'miles', 'wonderful', 'built', 'fell', 'opposite', 'rema
ining', 'match', 'sentence']
eat
['challenge', 'anyway', 'conduct', 'until', 'again', 'towards', '–', 'ahea
d', 'close', 'rest']
good
['stay', 'even', 'ever', 'like', 'test', 'entirely', "'d", 'right', 'need'
, 'look']
slowly
['place', 'until', 'eyes', 'planned', 'deep', 'behalf', 'rest', 'extra', '
forward', 'break']
100
['coast', 'conference', 'minutes', '4.', '2003', 'october', 'september', '
north', 'million', '2000']
```

In [81]:
```
run(windowSize=5, noComp=50, leftOnly=True, rightOnly=False)
```

```
Window : 5 ; No components : 50 ; Left only : True ; Right only : False
------------------------------------------------------------------------
Time taken is ----  19.888325214385986
boy
['length', 'usual', 'stay', 'hold', 'show', 'regard', 'came', 'eat', 'same
', 'onto']
sunday
['2003', 'friday', '2004', 'october', '19th', '5th', '4th', 'saturday', '2
001', 'january']
eat
['come', 'consistent', 'seem', 'usual', 'too', 'speak', 'dog', 'regard', '
really', 'same']
good
["'ll", 'longer', 'hard', 'best', 'look', 'like', 'much', 'sense', 'fun',
'another']
slowly
['find', 'mode', 'sir', 'follow', 'even', 'combination', 'make', 'without'
, 'hand', 'rest']
100
['entries', 'column', '1000', 'cm', '1', 'motor', '2', '=', 'million', '10
']
```

In [82]:
```
run(windowSize=5, noComp=100, leftOnly=True, rightOnly=False)
```

```
Window : 5 ; No components : 100 ; Left only : True ; Right only : False
------------------------------------------------------------------------
Time taken is ----  21.17309880256653
boy
['every', 'term', 'length', 'since', 'until', 'day', 'summer', 'go', 'same
', 'however']
sunday
['2002', '2003', '2004', 'october', 'september', '2001', 'saturday', '2005
', 'january', 'jack']
eat
['top', 'tree', 'back', 'flash', 'does', '3.', 'really', 'down', 'fairly',
'onto']
good
["'ll", 'great', 'best', 'even', 'look', 'like', 'much', 'strong', 'every'
, 'way']
slowly
['correct', 'mode', 'truth', 'even', 'get', 'stop', 'along', 'without', 'm
e', 'rest']
100
['pp', 'column', '1000', '1', 'cm', 'mm', '2', '=', '10', '7.']
```

In [83]:
```
run(windowSize=5, noComp=200, leftOnly=True, rightOnly=False)
```

```
Window : 5 ; No components : 200 ; Left only : True ; Right only : False
------------------------------------------------------------------------
Time taken is ----  25.457983016967773
boy
['aspect', 'praise', 'until', 'day', 'boys', 'however', 'book', 'every', '
crime', 'object']
sunday
['july', 'april', 'friday', 'october', 'september', 'morning', 'saturday',
'june', 'january', '1st']
eat
['homes', 'top', '3.', 'aged', 'young', 'does', 'varied', 'loan', 'onto',
'down']
good
['hard', 'little', "'ll", 'best', 'think', 'like', 'fit', 'nice', 'lot', '
look']
slowly
['song', 'truth', 'correct', 'mode', 'owner', 'summit', 'chance', 'easily'
, 'touch', 'advance']
100
['entries', 'column', '1000', '1', 'cm', 'mm', '2', '=', '10', '12']
```

```
In [84]: run(windowSize=10, noComp=10, leftOnly=True, rightOnly=False)
```

```
Window : 10 ; No components : 10 ; Left only : True ; Right only : False
-----------------------------------------------------------------------
Time taken is ----  19.988296508789062
boy
['messages', 'actually', 'function', 'exactly', 'mention', 'eyes', 'record
ing', 'dream', 'goal', 'trying']
sunday
['la', 'confirmed', 'followed', 'venue', 'kingdom', '3rd', 'opened', 'mile
s', 'perfect', 'sugar']
eat
['secure', 'makes', 'around', 'even', 'possibly', 'far', 'noise', 'multipl
e', 'movement', 'effect']
good
['opportunity', 'might', 'things', 'suffered', 'done', 'believe', 'everyon
e', 'come', 'could', 'right']
slowly
['putting', 'speak', 'along', 'phase', 'advance', 'victory', 'conservative
', 'upon', 'paid', 'made']
100
['column', 'mm', '150', 'irish', 'nine', 'select', 'edition', 'beauty', 'q
', 'location']
```

```
In [85]: run(windowSize=10, noComp=50, leftOnly=True, rightOnly=False)
```

```
Window : 10 ; No components : 50 ; Left only : True ; Right only : False
-----------------------------------------------------------------------
Time taken is ----  23.705852031707764
boy
['mum', 'her', 'hit', 'she', 'eyes', 'mrs', 'favourite', 'daughter', 'marr
ied', 'movie']
sunday
['station', 'great', '1st', 'venue', 'october', 'september', 'st', 'saturd
ay', 'after', 'town']
eat
['deep', 'whole', 'until', 'alone', 'once', 'negative', 'grow', 'behind',
'real', 'understand']
good
['everyone', 'just', 'things', 'look', 'like', 'much', 'make', 'take', 'co
uld', 'way']
slowly
['weight', 'tables', 'turn', 'mass', 'river', 'character', 'hold', 'ball',
'several', 'shows']
100
['capacity', 'column', '1000', '1', 'cm', 'motor', 'mm', '2', '=', '10']
```

In [86]:
```
run(windowSize=10, noComp=100, leftOnly=True, rightOnly=False)
```

```
Window : 10 ; No components : 100 ; Left only : True ; Right only : False
------------------------------------------------------------------------
Time taken is ----  25.24770212173462
boy
['mum', 'her', 'hit', 'she', 'eyes', 'mrs', 'favourite', 'returned', 'daug
hter', 'husband']
sunday
['2005', 'october', 'morning', 'november', 'september', 'st', 'thursday',
'june', 'after', 'january']
eat
['differences', 'until', 'whole', 'used', 'real', 'negative', 'grow', 'abl
e', 'secondary', 'another']
good
['everyone', 'make', 'look', 'like', 'much', 'way', 'enough', 'right', 'ju
st', 'take']
slowly
['through', 'resulting', 'force', 'song', 'character', 'ball', 'exposure',
'putting', 'while', 'door']
100
['capacity', 'x', '1000', '1', 'cm', 'tours', 'mm', '2', '=', '10']
```

In [87]:
```
run(windowSize=10, noComp=200, leftOnly=True, rightOnly=False)
```

```
Window : 10 ; No components : 200 ; Left only : True ; Right only : False
------------------------------------------------------------------------
Time taken is ----  29.54141855239868
boy
['mum', 'her', 'favourite', 'she', 'returned', 'mother', 'eyes', 'daughter
', 'husband', 'movie']
sunday
['monday', '11th', 'october', 'morning', 'november', 'september', 'few', '
saturday', 'june', 'january']
eat
['best', 'both', 'meals', 'whole', 'used', 'negative', 'secondary', 'grow'
, 'young', 'parents']
good
['everyone', 'little', 'get', 'known', 'look', 'like', "'m", 'thought', 't
ake', 'quite']
slowly
['through', 'song', 'mail', 'advance', 'resulting', 'western', 'job', 'bal
l', 'exposure', 'door']
100
['capacity', 'pp', '1000', '1', 'cm', 'tours', 'mm', '2', '=', '10']
```

In [65]:
```
run(windowSize=2, noComp=10, leftOnly=False, rightOnly=True)
```

```
Window : 2 ; No components : 10 ; Left only : False ; Right only : True
-----------------------------------------------------------------------
Time taken is ----  46.580514430999756
boy
['adobe', 'attend', 'final', 'green', 'location', 'forced', 'birds', 'loca
tions', 'speaker', 'supplied']
sunday
['literature', '17th', 'seminar', 'saturday', 'pub', 'morning', '1st', 'di
rector', 'dates', 'routes']
eat
['happen', 'something', 'heavy', 'stop', "'ve", "'re", 'am', 'sorry', 'say
ing', 'me']
good
['though', 'him', 'thanks', 'treated', 'father', 'taking', 'take', 'moveme
nt', 'nor', 'men']
slowly
['amounts', 'gives', 'parties', 'rather', 'lower', 'temperature', 'particu
lar', 'exposure', 'aimed', 'considered']
100
['tour', 'hotel', 'hotels', 'height', '1000', '1', 'white', '90', '=', 'mm
']
```

In [88]:
```
run(windowSize=2, noComp=50, leftOnly=False, rightOnly=True)
```

```
Window : 2 ; No components : 50 ; Left only : False ; Right only : True
-----------------------------------------------------------------------
Time taken is ----  49.639742374420166
boy
['sea', 'him', 'dear', 'green', 'dream', 'wood', 'father', 'daughter', 'mo
ther', 'sought']
sunday
['bible', 'evening', 'john', '11th', '4th', 'acts', 'morning', 'saturday',
'june', '1st']
eat
["'re", 'stay', "'ll", 'just', 'did', 'lot', 'know', 'looks', 'saying', 'b
it']
good
['yet', 'true', 'around', 'even', 'again', 'life', 'saw', 'really', 'right
', 'way']
slowly
['properly', 'impossible', 'red', 'shown', 'majority', 'suggests', 'facts'
, 'takes', 'brand', 'move']
100
['capacity', 'length', 'column', '1000', 'document', 'cm', '1', 'mm', '=',
'10']
```

In [89]:
```
run(windowSize=2, noComp=100, leftOnly=False, rightOnly=True)
```

```
Window : 2 ; No components : 100 ; Left only : False ; Right only : True
-----------------------------------------------------------------------
Time taken is ----  51.35407614707947
boy
['him', 'dear', 'arms', 'green', 'dream', 'wood', 'sea', 'daughter', 'moth
er', 'sought']
sunday
['bible', 'evening', 'saturday', '11th', '4th', 'acts', 'morning', 'john',
'june', '1st']
eat
['just', 'got', "'re", 'think', 'bit', 'did', 'feel', 'looks', 'lot', 'thi
ng']
good
['yet', 'saw', "'ll", 'well', 'though', 'got', 'make', 'way', 'said', 'aga
in']
slowly
['properly', 'however', 'fact', 'surrounding', 'therapy', 'necessary', 'fa
cts', 'understood', 'pressure', 'climate']
100
['capacity', 'length', 'score', '1000', 'document', 'cm', '1', 'mm', '=',
'10']
```

In [90]:
```
run(windowSize=2, noComp=200, leftOnly=False, rightOnly=True)
```

```
Window : 2 ; No components : 200 ; Left only : False ; Right only : True
-----------------------------------------------------------------------
Time taken is ----  56.907262086868286
boy
['him', 'saying', 'dear', 'green', 'dream', 'wood', 'sea', 'daughter', 'mo
ther', 'sought']
sunday
['bible', 'evening', '2nd', '11th', '4th', 'acts', 'morning', 'saturday',
'june', '1st']
eat
['bit', 'feel', "'ll", 'think', 'just', 'did', 'got', 'saying', 'seemed',
"'d"]
good
['yet', 'saw', "'ll", 'well', 'though', 'got', 'dream', 'said', 'true', 'a
gain']
slowly
['eligible', 'qualification', 'social', 'climate', 'take', 'facts', 'regis
tered', 'pressure', 'worker', 'fact']
100
['capacity', 'volume', 'length', '1000', 'document', '1', 'mm', 'cm', '=',
'10']
```

In [91]:

```
run(windowSize=5, noComp=10, leftOnly=False, rightOnly=True)
```

```
Window : 5 ; No components : 10 ; Left only : False ; Right only : True
-----------------------------------------------------------------------
Time taken is ----  45.56382417678833
boy
['adobe', 'attend', 'final', 'green', 'location', 'forced', 'birds', 'loca
tions', 'speaker', 'supplied']
sunday
['literature', '17th', 'seminar', 'saturday', 'pub', 'morning', '1st', 'di
rector', 'dates', 'routes']
eat
['happen', 'something', 'heavy', 'stop', "'ve", "'re", 'am', 'sorry', 'say
ing', 'me']
good
['though', 'him', 'thanks', 'treated', 'father', 'taking', 'take', 'moveme
nt', 'nor', 'men']
slowly
['amounts', 'gives', 'parties', 'rather', 'lower', 'temperature', 'particu
lar', 'exposure', 'aimed', 'considered']
100
['tour', 'hotel', 'hotels', 'height', '1000', '1', 'white', '90', '=', 'mm
']
```

In [92]:

```
run(windowSize=5, noComp=50, leftOnly=False, rightOnly=True)
```

```
Window : 5 ; No components : 50 ; Left only : False ; Right only : True
-----------------------------------------------------------------------
Time taken is ----  50.06431269645691
boy
['sea', 'him', 'dear', 'green', 'dream', 'wood', 'father', 'daughter', 'mo
ther', 'sought']
sunday
['bible', 'evening', 'john', '11th', '4th', 'acts', 'morning', 'saturday',
'june', '1st']
eat
["'re", 'stay', "'ll", 'just', 'did', 'lot', 'know', 'looks', 'saying', 'b
it']
good
['yet', 'true', 'around', 'even', 'again', 'life', 'saw', 'really', 'right
', 'way']
slowly
['properly', 'impossible', 'red', 'shown', 'majority', 'suggests', 'facts'
, 'takes', 'brand', 'move']
100
['capacity', 'length', 'column', '1000', 'document', 'cm', '1', 'mm', '=',
'10']
```

In [93]:
```
run(windowSize=5, noComp=100, leftOnly=False, rightOnly=True)
```

```
Window : 5 ; No components : 100 ; Left only : False ; Right only : True
----------------------------------------------------------------------
Time taken is ----  53.526697635650635
boy
['him', 'dear', 'arms', 'green', 'dream', 'wood', 'sea', 'daughter', 'moth
er', 'sought']
sunday
['bible', 'evening', 'saturday', '11th', '4th', 'acts', 'morning', 'john',
'june', '1st']
eat
['just', 'got', "'re", 'think', 'bit', 'did', 'feel', 'looks', 'lot', 'thi
ng']
good
['yet', 'saw', "'ll", 'well', 'though', 'got', 'make', 'way', 'said', 'aga
in']
slowly
['properly', 'however', 'fact', 'surrounding', 'therapy', 'necessary', 'fa
cts', 'understood', 'pressure', 'climate']
100
['capacity', 'length', 'score', '1000', 'document', 'cm', '1', 'mm', '=',
'10']
```

In [94]:
```
run(windowSize=5, noComp=200, leftOnly=False, rightOnly=True)
```

```
Window : 5 ; No components : 200 ; Left only : False ; Right only : True
----------------------------------------------------------------------
Time taken is ----  55.80970644950867
boy
['true', 'him', 'dear', 'green', 'dream', 'wood', 'sea', 'daughter', 'moth
er', 'sought']
sunday
['bible', 'evening', '2nd', '11th', '4th', 'acts', 'morning', 'saturday',
'june', '1st']
eat
['bit', 'feel', "'ll", 'think', 'just', 'did', 'got', 'saying', 'seemed',
"'d"]
good
['yet', 'saw', "'ll", 'well', 'though', 'got', 'dream', 'said', 'true', 'a
gain']
slowly
['eligible', 'qualification', 'social', 'climate', 'take', 'facts', 'regis
tered', 'pressure', 'worker', 'fact']
100
['capacity', 'volume', 'length', '1000', 'document', '1', 'mm', 'cm', '=',
'10']
```

In [95]:
```
run(windowSize=10, noComp=10, leftOnly=False, rightOnly=True)
```

```
Window : 10 ; No components : 10 ; Left only : False ; Right only : True
------------------------------------------------------------------------
Time taken is ----  45.79260516166687
boy
['adobe', 'attend', 'final', 'green', 'location', 'forced', 'birds', 'loca
tions', 'speaker', 'supplied']
sunday
['literature', '17th', 'seminar', 'saturday', 'pub', 'morning', '1st', 'di
rector', 'dates', 'routes']
eat
['happen', 'something', 'heavy', "'ve", 'stop', "'re", 'am', 'me', 'saying
', 'sorry']
good
['though', 'him', 'thanks', 'treated', 'father', 'taking', 'take', 'moveme
nt', 'nor', 'men']
slowly
['amounts', 'gives', 'parties', 'rather', 'lower', 'temperature', 'particu
lar', 'exposure', 'aimed', 'considered']
100
['tour', 'hotel', 'hotels', 'height', '1000', '1', 'white', '90', '=', 'mm
']
```

In [96]:
```
run(windowSize=10, noComp=50, leftOnly=False, rightOnly=True)
```

```
Window : 10 ; No components : 50 ; Left only : False ; Right only : True
------------------------------------------------------------------------
Time taken is ----  54.28583335876465
boy
['sea', 'him', 'dear', 'green', 'dream', 'wood', 'father', 'daughter', 'mo
ther', 'sought']
sunday
['bible', 'evening', 'john', '11th', '4th', 'acts', 'morning', 'saturday',
'june', '1st']
eat
["'re", 'lot', "'ll", 'stay', 'looks', 'did', 'just', 'know', 'saying', 'b
it']
good
['yet', 'true', 'around', 'even', 'again', 'life', 'saw', 'really', 'right
', 'way']
slowly
['properly', 'impossible', 'red', 'shown', 'majority', 'suggests', 'facts'
, 'takes', 'brand', 'move']
100
['capacity', 'length', 'column', '1000', 'document', 'cm', '1', 'mm', '=',
'10']
```

In [97]:
```
run(windowSize=10, noComp=100, leftOnly=False, rightOnly=True)
```

```
Window : 10 ; No components : 100 ; Left only : False ; Right only : True
----------------------------------------------------------------------
Time taken is ----  51.38403511047363
boy
['him', 'dear', 'arms', 'green', 'dream', 'wood', 'sea', 'daughter', 'moth
er', 'sought']
sunday
['bible', 'evening', 'saturday', '11th', '4th', 'acts', 'morning', 'john',
'june', '1st']
eat
['just', 'got', "'re", 'think', 'bit', 'did', 'feel', 'looks', 'lot', 'thi
ng']
good
['yet', 'saw', "'ll", 'well', 'though', 'got', 'make', 'way', 'said', 'aga
in']
slowly
['properly', 'however', 'fact', 'surrounding', 'therapy', 'necessary', 'fa
cts', 'understood', 'pressure', 'climate']
100
['capacity', 'length', 'score', '1000', 'document', 'cm', '1', 'mm', '=',
'10']
```

In [98]:
```
run(windowSize=10, noComp=200, leftOnly=False, rightOnly=True)
```

```
Window : 10 ; No components : 200 ; Left only : False ; Right only : True
----------------------------------------------------------------------
Time taken is ----  56.408291816711426
boy
['true', 'him', 'dear', 'green', 'dream', 'wood', 'sea', 'daughter', 'moth
er', 'sought']
sunday
['bible', 'evening', '2nd', '11th', '4th', 'acts', 'morning', 'saturday',
'june', '1st']
eat
['bit', 'feel', "'ll", 'think', 'just', 'did', 'got', 'saying', 'seemed',
"'d"]
good
['yet', 'saw', "'ll", 'well', 'though', 'got', 'dream', 'said', 'true', 'a
gain']
slowly
['eligible', 'qualification', 'social', 'climate', 'take', 'facts', 'regis
tered', 'pressure', 'worker', 'fact']
100
['capacity', 'volume', 'length', '1000', 'document', '1', 'mm', 'cm', '=',
'10']
```

KMeans Clustering

In [108]:

```
# K-means Clustering
from sklearn.cluster import KMeans

cmatrix = populate_cmatrix(file_name, iunigrams, idimensions, window = 5, le
svd = TruncatedSVD(n_components = 100, random_state=42)
svd.fit(cmatrix)
cmatrix = svd.transform(cmatrix)

kmeans = KMeans(n_clusters=100, random_state=0).fit(cmatrix)
i=0
wordTolabel={}
for word in dimensions:
    wordTolabel[word] = kmeans.labels_[i]
    i=i+1
sortedByLabel = sorted(wordTolabel.items(), key = lambda x : x[1], reverse =
print(sortedByLabel)
```

```
[('graduate', 0), ('constant', 0), ('ball', 0), ('principal', 0), ('update
', 0), ('contains', 0), ('error', 0), ('posts', 0), ('national', 0), ('rec
ords', 0), ('perhaps', 0), ('included', 0), ('claimed', 0), ('intended', 0
), ('credit', 0), ('flexible', 0), ('finally', 0), ('budget', 0), ('missin
g', 0), ('follow', 0), ('manager', 0), ('external', 0), ('medieval', 0), (
'administration', 0), ('pensions', 0), ('engineers', 0), ('end', 0), ('sta
tion', 0), ('80', 0), ('copyright', 0), ('act', 0), ('nearly', 0), ('form'
, 0), ('5.', 0), ('sell', 0), ('30', 0), ('sir', 0), ('easily', 0), ('vote
', 0), ('q', 0), ('reducing', 0), ('below', 0), ('exercise', 0), ('rail',
0), ('quality', 0), ('consider', 0), ('0', 0), ('exhibition', 0), ('signs'
, 0), ('pair', 0), ('sponsorship', 0), ('strategies', 0), ('huge', 0), ('r
esearch', 0), ('sky', 0), ('employers', 0), ('map', 0), ('ii', 0), ('energ
y', 0), ('leading', 0), ('properly', 0), ('turns', 0), ('looks', 0), ('cas
es', 0), ('train', 0), ('items', 0), ('running', 0), ('exciting', 0), ('ce
ntury', 0), ('organisation', 0), ('errors', 0), ('qualified', 0), ('miss',
0), ('fashion', 0), ('activities', 0), ('ordered', 0), ('behalf', 0), ('dv
d', 0), ('approach', 0), ('8.', 0), ('roles', 0), ('incident', 0), ('wider
', 0), ('museum', 0), ('wild', 0), ('maintain', 0), ('base', 0), ('trials'
, 0), ('developing', 0), ('big', 0), ('experts', 0), ('independent', 0), (
'views', 0), ('odds', 0), ('club', 0), ('overseas', 0), ('corporation', 0)
```

OBSERVATIONS:

The words didnt group into clusters of different POS tags. Majority of words are gouped into cluster 0. When number of clusters are increased, we find a different clustering scheme where words are distributed among other clusters also. The above is the experiment with n_clusters = 100

In [ ]: