

Assignment 1 - Use NLTK to conduct text processing

1. Install Python and NLTK.

Python 3 was installed and NLTK loaded. Then pyCharm was loaded and reset to Python 3.

```
(java:19895): Gtk-WARNING **: Unable to locate theme engine in module_path: "adwaita",  
[chris@Euclid bin]$ python  
Python 3.4.4 (default, Jul 22 2016, 08:05:32)  
[GCC 4.9.1 Compatible Clang 3.8.1 (tags/RELEASE_381/final)] on linux  
Type "help", "copyright", "credits" or "license" for more information.  
>>> import nltk  
>>> nltk.download()  
NLTK Downloader  
-----  
d) Download l) List u) Update c) Config h) Help q) Quit  
Downloader> d  
Download which package (l=list; x=cancel)?  
Identifier> brown  
Downloading package brown to /home/chris/nltk_data...  
Package brown is already up-to-date!  
-----  
d) Download l) List u) Update c) Config h) Help q) Quit  
Downloader> q  
True  
>>> brown.words()  
Traceback (most recent call last):  
  File "<stdin>", line 1, in <module>  
NameError: name 'brown' is not defined  
>>> from nltk.corpus import brown  
>>> brown.words  
<bound method CategorizedTaggedCorpusReader.words of <CategorizedTaggedCorpusReader in '/home/chris/nltk_data/corpora/brown'>>  
>>> brown.words()  
['The', 'Fulton', 'County', 'Grand', 'Jury', 'said', ...]  
>>>
```

Illustration 1: Command line screen shot of brown.words()

2. Tokenize the documents into words, remove stop word, and conduct stemming.

Using code in Appendix A. The documents in the path containing the .txt extension were converted to lower case and tokenized. Stopwords and stemming were removed in the tokenizing process. A separate tokenizer was created to pass to the tfidf vectorizer. The tokenized, stemmed words are listed in Appendix B.

3. Calculate tf-idf for each word in the document and generate document-word matrix.

The tf-idf for each word was calculated and is attached in Appendix C.

4. Calculate pairwise cosine similarity for the documents.

The file list and cosine similarity matrix is as follows:

Document cosine similarity matrix:

```
['100554newsML.txt', '100593newsML.txt', '100618newsML.txt', '130040newsML.txt',  
'137871newsML.txt']  
[[ 1.          0.80887809  0.82951572  0.44085006  0.40472577]]  
[[ 0.80887809  1.          0.97518013  0.36302538  0.32228205]]  
[[ 0.82951572  0.97518013  1.          0.36699489  0.3327344  ]]  
[[ 0.44085006  0.36302538  0.36699489  1.          0.30557307]]  
[[ 0.40472577  0.32228205  0.3327344  0.30557307  1.          ]]
```

5. To validate the code used to create the values, a document was duplicated and the resulting matrix is as follows:

Document cosine similarity matrix:

```
['100554newsML.txt', '100593newsML.txt', '100618newsML.txt', '130040newsML.txt',  
'137871newsML.txt', 'acopyof100554.txt']  
[[ 1.          0.81481268  0.83636607  0.42807788  0.39302942  1.          ]]  
[[ 0.81481268  1.          0.97397535  0.35161264  0.31194988  0.81481268]]  
[[ 0.83636607  0.97397535  1.          0.3562864  0.32259638  0.83636607]]  
[[ 0.42807788  0.35161264  0.3562864  1.          0.28193159  0.42807788]]  
[[ 0.39302942  0.31194988  0.32259638  0.28193159  1.          0.39302942]]  
[[ 1.          0.81481268  0.83636607  0.42807788  0.39302942  1.          ]]
```

Appendix A – Project code (Python)

```
import os
import nltk
import string

from nltk.stem.porter import PorterStemmer
from nltk.corpus import stopwords

from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity

x = []
file_list = []
file_content = []
file_tokens = []
ps = PorterStemmer()
file_stems = []
stop_words = set(stopwords.words("english"))

"""
    The tokenize function cnverts the text to tokens, then the stop words are
    removed and then the
    tokens are stemmed. The stem_tokens requires stop words to be set since
    tokenize is called from
    the tfidf function.
"""
def tokenize(text):
    tokens = nltk.word_tokenize(text)
    stems = stem_tokens(tokens, ps)
    return stems

""" stem tokens requires stop_words to be set """
def stem_tokens(tokens, stemmer):
    stems = []
    for token in tokens:
        if token not in stop_words:
            stems.append(stemmer.stem(token))
    return stems

#{ "100554newsML.txt", "100593newsML.txt", "100618newsML.txt", "130040newsML.txt",
"137871newsML.txt"}
""" This program assumes all .txt files in its interpreted tree are to be processed
    the file list is created, the files tokenized, then stop words are removed and
    stems located"""
path="/home/chris/Dropbox/MSCS/CAP6776 - Information and Data
Retrieval/Project01/hw 1 datasets/"
for files in os.listdir(path):
    if files.endswith('.txt'):
        file_list.append(files)

file_list = sorted(file_list)
print("Read from file list, convert to lower case then remove punctuation:")
for filename in file_list:
    content =
open(path+filename).read().lower().translate(dict.fromkeys(string.punctuation))
# convert to lowercase here for ease of processing
x.append(content)
```

```

print("Tokenize: remove stop words and conduct stemming:")
for stuff in x:
    print(tokenize(stuff))

print("Vectorize, removing stop words and tokenize:")
vectorizer = TfidfVectorizer(tokenizer=tokenize, stop_words='english',
                             lowercase=False)

#    print(f)
tfidf = vectorizer.fit_transform(x)

#get word array (feature_names)

words = vectorizer.get_feature_names()

""" Print unsorted document word matrix.
    only use words found in documents ignore sparse array
"""
for c in tfidf.nonzero()[1]:
    print (words[c], ' - ', tfidf[0,c])

print("Document cosine similarity matrix:")

cos_matrix=[]
for r in tfidf:
    cos_matrix.append(cosine_similarity(r, tfidf))

print(file_list)
for n in cos_matrix:
    print(n)

```

Appendix B – Tokenized words

Tokenize: remove stop words and conduct stemming:

```

['channel', 'tunnel', 'oper', 'eurotunnel', 'monday', 'announc', 'detail', 'deal', 'give',
'bank'
, 'creditor', '45.5', 'percent', 'compani', 'return', 'wipe', '1.0', 'billion', 'pound',
'(', '$'
, '1.6', 'billion', ')', 'massiv', 'debt', '.', 'long-await', 'highli', 'complex',
'restructur',
'nearli', 'nearli', 'nine', 'billion', 'pound', 'debt', 'unpaid', 'interest', 'throw',
'compani',
'lifelin', 'could', 'secur', 'still', 'like', 'difficult', 'futur', '.', 'deal', ',',
'announc',
'simultan', 'pari', 'london', ',', 'bring', 'compani', 'back', 'brink', 'bankruptci',
'leav', 'c
urrent', 'sharehold', ',', 'alreadi', 'seen', 'invest', 'dwindl', ',', 'own', '54.5',
'percent',
'compani', '.', '`', 'fix', 'cap', 'interest', 'payment', 'arrang', 'pay', 'avail', 'cash',
',',
"'''", 'eurotunnel', 'co-chairman', 'alastair', 'morton', 'told', 'report', 'news', 'confer',
',',
, '`', 'avoid', 'name', 'game', '.', "'''", 'morton', 'said', 'plan', 'provid', 'anglo-

```

french', '
compani', 'medium', 'term', 'financi', 'stabil', 'consolid', 'commerci', 'posit', 'develop',
'ope
r', ',', 'ad', 'firm', 'make', 'profit', 'interest', '.', 'although', 'sharehold', 'see',
'hold',
'dilut', ',', 'offer', 'prospect', 'brighter', 'futur', 'urg', 'patient', 'month',
'uncertainti'
, 'eurotunnel', 'wrestl', 'reduc', 'crippl', 'interest', 'payment', 'negoti', 'tunnel',
"s", 'co
nstruct', '.', 'eurotunnel', ',', 'taken', 'around', 'half', 'market', 'busiest', 'cross-
channel'
, 'rout', 'european', 'ferri', 'compani', ',', 'said', 'strong', 'oper', 'perform', 'could',
'all
ow', 'pay', 'first', 'dividend', 'within', 'next', '10', 'year', '.', 'french', 'co-
chairman', 'p
atrack', 'ponsol', 'told', 'report', 'pari', 'news', 'confer', 'dividend', 'could', 'come',
'earl
i', '2004', 'compani', 'perform', '``', 'well', '"', '.', 'eurotunnel', 'bank', 'come',
'ingeni'
, 'formula', 'help', 'compani', 'get', 'earli', 'year', 'deal', ',', 'despit', 'swap',
'debt', 'e
quiti', 'bond', ',', 'still', 'abl', 'afford', 'annual', 'interest', 'bill', '400',
'million', 'p
ound', '.', 'revenu', ',', 'cost', 'depreci', ',', 'less', '400', 'million', 'pound', ',',
'compa
ni', 'issu', '``', 'stabilis', 'note', '"', 'maximum', '1.85', 'billion', 'pound', 'bank',
'.'
'eurotunnel', 'would', 'pay', 'interest', 'note', '(', 'would', 'constitut', 'debt', 'issu',
)',
'ten', 'year', '.', 'analyst', 'said', 'deal', ',', 'eurotunnel', "s", 'abil', 'financ',
'debt'
, 'would', 'becom', 'sustain', ',', 'least', 'year', '.', '``', 'look', 'current', 'cash',
'flow'
, '150', '200', 'million', 'pound', 'year', ',', 'ca', "n't", 'find', '(', 'meet', 'bill',
)', '
roll', 'forward', 'stabilis', 'note', ',', 'keep', 'go', 'seven', ',', 'eight', ',', 'nine',
'yea
r', ',', '"', 'said', 'analyst', 'one', 'major', 'invest', 'bank', '.', '``', 'time', ',',
"',
'ad', '.', 'compani', 'said', 'statement', 'still', 'consider', 'work', 'done', 'finalis',
'agre
, 'detail', 'plan', 'submit', 'sharehold', 'bank', 'group', 'approv', ',', 'probabl',
'earli', '
spring', '1997.', 'eurotunnel', 'said', 'debt-for-equ', 'swap', 'would', '130', 'penc', ',',
'10.
40', 'franc', ',', 'per', 'share', '--', 'consider', 'level', '160', 'penc', 'wide',
'report', 'r
un', 'deal', 'compani', 'said', '3.7', 'billion', 'pound', 'debt', 'would', 'convert',
'new', 'fi
nanci', 'instrument', 'exist', 'sharehold', 'would', 'abl', 'particip', 'issu', '.',
'choos', 'ta
ke', 'free', 'warrant', 'entitl', 'subscrib', ',', 'eurotunnel', 'said', 'sharehold', '"',
'inter
est', 'may', 'reduc', '39', 'percent', 'compani', 'end', 'decemb', '2003.', 'eurotunnel',
"s", '

share', ',', 'suspend', 'last', 'week', '113.5', 'penc', 'ahead', 'monday', '"s', 'announc',
,',',
'resum', 'trade', 'tuesday', '.', 'sharehold', '225', 'creditor', 'bank', 'agre', 'deal',
,',',
'', '"m", 'hope', '"m", 'take', '(', 'approv', ')', 'grant', ',', '"', 'morton', 'admit',
,',',
'', 'sharehold', 'pretti', 'angri', 'franc', '.', '"', 'ask', 'would', 'happen', 'bank',
'reje
ct', 'deal', ',', 'morton', 'said', ',', '``', 'nobodi', 'want', 'collaps', ',', 'nobodi',
'want'
, 'doomsday', 'scenario', '.', '"', '(', '\$', '1=6393', 'pound', ')']
['anglo-french', 'channel', 'tunnel', 'oper', 'eurotunnel', 'monday', 'announc', 'deal',
'give',
'creditor', 'bank', '45.5', 'percent', 'compani', 'return', 'wipe', 'one', 'billion',
'pound', '(
, '\$', '1.56', 'billion', ')', 'debt', '.', 'long-await', 'restructur', 'bring', 'end',
'month',
'wrangl', 'eurotunnel', '225', 'bank', 'owe', 'nearli', 'nine', 'billion', 'pound', '(',
'\$', '1
4.1', 'billion', ')', '.', 'deal', ',', 'announc', 'simultan', 'pari', 'london', ',',
'bring', 'c
ompani', 'back', 'brink', 'insolv', 'leav', 'sharehold', 'own', '54.5', 'percent',
'compani', '.'
, '``', 'restructur', 'plan', 'provid', 'eurotunnel', 'medium-term', 'financi', 'stabil',
'allow'
, 'consolid', 'substanti', 'commerci', 'achiev', 'date', 'develop', 'oper', ',', '"',
'eurotunne
l', 'co-chairman', 'alastair', 'morton', 'said', '.', 'firm', 'make', 'profit', 'interest',
,',',
'ad', '.', 'although', 'sharehold', 'see', 'interest', 'dilut', ',', 'offer', 'prospect',
'bright
er', 'futur', 'month', 'uncertainiti', 'eurotunnel', 'wrestl', 'reduc', 'crippl', 'interest',
'pay
ment', 'negoti', 'tunnel', '"s', 'construct', '.', 'eurotunnel', ',', 'taken', 'around',
'half',
'cross-channel', 'market', 'european', 'ferri', 'compani', ',', 'said', 'strong', 'oper',
'perfor
m', 'could', 'allow', 'pay', 'first', 'dividend', 'within', 'next', '10', 'year', '.',
'french',
'co-chairman', 'patrick', 'ponsol', 'said', 'sharehold', 'would', 'patient', 'could',
'reap', 'be
nefit', 'compani', '"s', 'success', '.', 'call', 'debt', 'restructur', 'plan', '``',
'accept', 'c
ompromis', '"', 'holder', 'eurotunnel', 'share', '.', 'compani', 'said', 'still',
'consider', 'w
ork', 'done', 'finalis', 'agre', 'detail', 'plan', 'submit', 'sharehold', 'full', '225',
'bank',
'syndic', 'approv', ',', 'probabl', 'earli', '1997.', 'monday', '"s', 'announc', 'follow',
'two',
'week', 'highli', 'secret', 'negoti', 'eurotunnel', 'six', 'lead', 'bank', '.', 'extend',
'24',
'``', 'instruct', 'bank', '"', 'meet', 'late', 'last', 'week', 'london', '.', 'eurotunnel',
'sai
d', 'debt-for-equ', 'swap', 'would', '130', 'penc', ',', '10.40', 'franc', ',', 'per',
'share', '

., 'consider', 'level', 'around', '160', 'penc', 'wide', 'report', 'announc', 'deal', ',',
'redu
c', 'outstand', 'debt', '8.7', 'billion', 'pound', '(', '\$', '13.6', 'billion', ')', '1.0',
'bill
ion', '(', '\$', '1.56', 'billion', ')', '.', 'compani', 'said', '3.7', 'billion', 'pound',
'(', '
\$', '5.8', 'billion', ')', 'debt', 'would', 'convert', 'new', 'financi', 'instrument',
'exist', '
sharehold', 'would', 'abl', 'particip', 'issu', '.', 'choos', 'take', 'free', 'warrant',
'entitl'
, 'subscrib', ',', 'eurotunnel', 'said', 'sharehold', '"', 'interest', 'may', 'reduc', '39',
'per
cent', 'compani', 'end', 'decemb', '2003.', 'eurotunnel', '"s', 'share', ',', 'suspend',
'last',
'week', '113.5', 'penc', 'ahead', 'monday', '"s', 'announc', ',', 'resum', 'trade',
'tuesday', '
, 'compani', 'said', '.']
['anglo-french', 'channel', 'tunnel', 'oper', 'eurotunnel', 'monday', 'announc', 'deal',
'give',
'creditor', 'bank', '45.5', 'percent', 'compani', 'return', 'wipe', 'one', 'billion',
'pound', '('
, '\$', '1.56', 'billion', ')', 'debt', 'mountain', '.', 'long-await', 'restructur',
'bring', 'en
d', 'month', 'wrangl', 'eurotunnel', '225', 'bank', 'owe', 'nearli', 'nine', 'billion',
'pound',
, 'deal', ',', 'announc', 'simultan', 'pari', 'london', ',', 'bring', 'compani', 'back',
'brin
k', 'insolv', 'leav', 'sharehold', 'own', '54.5', 'percent', 'compani', '.', '``',
'restructur',
'plan', 'provid', 'eurotunnel', 'medium', 'term', 'financi', 'stabil', 'allow', 'consolid',
'subs
tanti', 'commerci', 'achiev', 'date', 'develop', 'oper', ',', '"', 'eurotunnel', 'co-
chairman',
'alastair', 'morton', 'said', '.', 'firm', 'make', 'profit', 'interest', ',', 'ad', '.',
'althoug
h', 'sharehold', 'see', 'interest', 'dilut', ',', 'offer', 'prospect', 'brighter', 'futur',
'mont
h', 'uncertain', 'eurotunnel', 'wrestl', 'reduc', 'crippl', 'interest', 'payment',
'negoti', 't
unnel', '"s', 'construct', '.', 'eurotunnel', ',', 'taken', 'around', 'half', 'cross-
channel', 'm
arket', 'european', 'ferri', 'compani', ',', 'said', 'strong', 'oper', 'perform', 'could',
'allow
, 'pay', 'first', 'dividend', 'within', 'next', '10', 'year', '.', 'french', 'co-chairman',
'pat
rick', 'ponsol', 'said', 'sharehold', 'would', 'patient', 'could', 'reap', 'benefit',
'compani',
"s", 'success', '.', 'call', 'debt', 'restructur', 'plan', '``', 'accept', 'compromis',
"''", 'h
older', 'eurotunnel', 'share', '.', 'compani', 'said', 'statement', 'still', 'consider',
'work',
'done', 'finalis', 'agre', 'detail', 'plan', 'submit', 'sharehold', 'full', '225', 'bank',
'syndi
c', 'approv', ',', 'probabl', 'earli', '1997.', 'monday', '"s', 'announc', 'follow', 'two',
'week

, 'highli', 'secret', 'negoti', 'eurotunnel', 'six', 'lead', 'bank', '.', 'extend', '24',
,
'instruct', 'bank', '"', 'meet', 'late', 'last', 'week', 'london', '.', 'eurotunnel',
'said', 'd
ebt-for-equ', 'swap', 'would', '130', 'penc', ',', '10.40', 'franc', ',', 'per', 'share',
,', 'c
onsider', 'level', 'around', '160', 'penc', 'wide', 'report', 'run', 'deal', ',', 'reduc',
'outst
and', 'debt', '8.7', 'billion', 'pound', '1.0', 'billion', '.', 'compani', 'said', '3.7',
'billio
n', 'pound', 'debt', 'would', 'convert', 'new', 'financi', 'instrument', 'exist',
'sharehold', 'w
ould', 'abl', 'particip', 'issu', '.', 'choos', 'take', 'free', 'warrant', 'entitl',
'subscrib',
,', 'eurotunnel', 'said', 'sharehold', '"', 'interest', 'may', 'reduc', '39', 'percent',
'compan
i', 'end', 'decemb', '2003.', 'eurotunnel', "'s", 'share', ',', 'suspend', 'last', 'week',
'113.5
, 'penc', 'ahead', 'monday', "'s", 'announc', ',', 'resum', 'trade', 'tuesday', ',',
'compani',
'said', '.', '(', '\$', '1=.6393', 'pound', ')']
['britain', "'s", 'barclay', 'plc', 'said', 'monday', 'talk', 'u.s.', 'group', 'morgan',
'stanley
, 'global', 'custodi', 'busi', 'bank', 'sourc', 'confirm', 'unit', 'sale', '.', 'barclay',
'morg
an', 'stanley', 'coy', 'respond', 'press', 'specul', ',', 'say', '`', 'discuss', 'explor',
'pote
nti', 'futur', 'co-oper', '"', '.', 'move', 'would', 'anoth', 'step', 'consolid', 'global',
'cus
todi', 'busi', 'becom', 'domin', 'bank', 'will', 'make', 'larg', 'front-end', 'invest',
'system',
'necessari', 'later', 'rake', 'fee', '.', 'barclay', '"', 'custodi', 'busi', '350', '400',
'staf
f', 'worldwid', ',', 'work', 'far', 'east', ',', 'time', 'treat', 'bank', "'s", 'core',
'busi', '
, 'secur', 'worth', '150', 'billion', 'pound', '(', '\$', '239.2', 'billion', ')', 'place',
'uni
t', '.', 'interim', 'result', ',', 'publish', 'august', ',', 'barclay', 'includ', 'custodi',
,
, 'busi', 'transit', '"', 'categori', 'includ', 'restructur', 'busi', 'franc', 'unit',
'state',
, 'categori', 'includ', 'custodi', 'made', 'profit', '23', 'million', 'pound', 'half-
year', 'j
une', '30.', 'report', 'said', ',', '`', 'busi', 'transit', 'primarili', 'compris', 'lend',
'ass
et', 'unlik', 'long-term', 'interest', 'group', 'requir', 'signific', 'restructur', '.',
, 'c
lear', 'deal', 'finalis', 'analyst', 'said', 'would', 'mark', 'step', 'consolid', 'global',
'cust
odi', 'market', '.', '`', 'consolid', 'natur', 'look', 'econom', 'busi', ',', '"', 'said',
'joh
n', 'leonard', ',', 'bank', 'analyst', 'salomon', 'brother', '.', '`', "'s", 'scale',
'economi',
'buisi', '.', '"', 'leonard', 'said', 'bank', 'taken', 'decis', 'invest', 'larg', 'sum',
'need'

, 'get', 'comput', 'system', 'scratch', 'competit', '.', 'custodi', 'busi', 'rang',
'tradit', 'sa
fekeep', 'secur', ',', 'britain', 'domin', 'lloyd', 'tsb', 'royal', 'bank', 'scotland', ',',
'per
form', 'measur', 'stock', 'lend', '.', ',', 'ceas', 'larg', 'risk-fre', 'busi', 'although',
'risk
, 'level', 'still', 'rel', 'low', '.', '`', 'guarante', 'stead', 'incom', 'stream', "'s",
'tot
al', 'insensit', 'market', 'volum', ',', '"', 'leonard', 'ad', '.', 'analyst', 'said',
'barclay'
, 'presum', 'decid', 'without', 'invest', ',', 'busi', 'small', 'compet', 'global', 'stage',
'..',
'expect', 'morgan', 'stanley', 'pay', 'well', '100', 'million', 'pound', 'busi', '.', '`',
'cou
ld', 'well', 'less', 'fifti', 'million', '(', 'pound', ')', ',', '"', 'said', 'one', '.',
'finan
c', 'worker', '"', 'union', 'bifu', 'attack', 'leak', 'discuss', ',', 'say', 'staff',
'face', 'un
certain', 'futur', 'busi', 'sold', '.', 'call', 'bank', '`', 'come', 'clean', '"',
'intent', '.'
, '(', '\$', '1=6270', 'pound', ')']
['italian', 'busi', 'tycoon', 'silvio', 'berlusconi', 'wednesday', 'lost', 'appeal',
'london', 'h
igh', 'court', 'transfer', 'document', 'italian', 'author', 'investig', 'alleg', 'fraud',
'fals',
'account', '.', 'britain', "'s", 'seriou', 'fraud', 'offic', 'seiz', '15', 'bag',
'document', 'a
pril', 'offic', 'london', 'lawyer', 'david', 'mill', 'repres', 'interest', 'former',
'italian', '
prime', 'minist', 'berlusconi', 'britain', '.', 'judg', ',', 'lord', 'justic', 'simon',
'brown',
'mr', 'justic', 'gage', ',', 'reject', 'ground', 'appeal', ',', 'brown', 'agre', 'hear',
'applic'
, 'appel', 'friday', 'morn', ',', 'could', 'lead', 'appeal', 'hous', 'lord', ',',
'although', 'le
gal', 'sourc', 'said', 'unlik', 'succeed', '.', 'hear', 'stem', 'appeal', 'berlusconi', ',',
'ita
lian', 'prime', 'minist', 'seven', 'month', '1994', 'trial', 'milan', 'corrupt', 'charg',
, ',', 'c
ompani', 'fininvest', 'compani', 'presid', 'fedel', 'confalonieri', 'court', 'rule',
'earlier', '
year', 'paper', 'sent', 'itali', '.', 'brown', 'said', 'paper', 'link', 'fraud', 'fals',
'account
, 'alleg', 'amount', '51', 'million', 'pound', '(', '\$', '82', 'million', ')', '`',
'surreptiti
, 'remov', 'fininvest', 'use', 'crimin', 'purpos', '.', '"', 'note', 'prosecut',
'alreadi', 'un
derway', 'itali', 'berlusconi', 'bribe', 'revenu', 'inspector', 'make', 'illicit', 'donat',
'form
er', 'prime', 'minist', 'bettino', 'craxi', '.', 'italian', 'investig', 'suspect',
'berlusconi',
, ',', 'among', 'hold', ',', 'own', 'three', 'tv', 'channel', 'footbal', 'club', 'ac',
'milan', ',',
, 'may', 'use', 'offshor', 'compani', 'channel', 'payment', 'politician', '.', 'end',
'judgement'

, 'brown', 'point', ',', '``', 'imper', 'recognis', ',', 'howev', ',', 'none', 'applic',
'yet', '
convict', 'anyth', 'noth', 'said', 'thought', 'rais', 'least', 'presumpt', 'guilt', '.',
'', 'r
eject', 'ground', 'appeal', 'basi', 'italian', 'request', 'wide', ',', 'justifi', 'offenc',
'ques
tion', '``', 'polit', '','', '.', 'one', 'ground', 'appeal', 'sfo', 'gone', 'berlusconi',
"s", 'l
awyer', "s", 'offic', '``', 'fish', 'expedit', '','', '--', 'unspecif', 'search', '--',
'illeg',
'english', 'law', '.', 'brown', 'describ', 'case', '``', 'wide-rang', ',', 'multi-facet',
',', 'i
ntern', 'fraud', 'involv', 'far-reach', 'alleg', 'larg', 'number', 'individu', 'connect',
'even',
'larger', 'number', 'compani', '.', '','', 'sfo', 'said', '``', 'happi', '','', 'rule', '.',
'``',
'sfo', "s", 'fight', 'fraud', 'intern', 'level', 'strengthen', 'welcom', 'judgement', ',',
''''
, 'said', 'statement', '.', 'sfo', 'said', 'judg', 'reject', 'main', 'argument', 'put',
'forward'
, 'appel', '.', '``', 'look', 'forward', 'speedi', 'transfer', 'document', 'milan',
'investig', '
get', 'job', ',', '','', 'said', '.', 'hear', 'friday', 'decid', 'whether', 'berlusconi',
'appeal'
, 'today', "s", 'rule', 'british', 'hous', 'lord', '.', 'sfo', 'senior', 'assist',
'director', '
chri', 'dickson', 'said', 'judg', 'current', 'hear', 'awar', ',', 'hous', 'lord', 'hear',
'take',
'place', 'next', 'year', ',', 'would', 'mean', 'paper', 'would', 'arriv', 'late', 'use',
'italia
n', 'proceed', '.', 'said', 'appel', 'given', 'leav', 'appeal', 'friday', 'expect', 'paper',
'tra
nsfer', 'immedi', ',', 'although', 'decis', 'home', 'offic', '.', '(', '\$', '1=.6260',
'pound', '
)']

Appendix C – Tokenized Words with tf-idf Vectors

channel - 0.0195579654464	45.5 - 0.0232492064976) - 0.0827100095316
tunnel - 0.0464984129952	percent - 0.0697476194928	massiv - 0.0347152484798
oper - 0.0697476194928	compani - 0.234695585357	debt - 0.139495238986
eurotunnel -	return - 0.0232492064976	. - 0.330840038127
0.232492064976	wipe - 0.0232492064976	long-await -
monday - 0.0391159308928	1.0 - 0.0232492064976	0.0232492064976
announc - 0.0697476194928	billion - 0.0977898272319	highli - 0.0232492064976
deal - 0.136905758125	pound - 0.132336015251	complex - 0.0347152484798
bank - 0.136905758125	(- 0.0827100095316	restructur -
creditor -	\$ - 0.0330840038127	0.0195579654464
0.0464984129952	1.6 - 0.0347152484798	nearli - 0.0464984129952

unpaid - 0.0347152484798	posit - 0.0347152484798	million - 0.0697476194928
throw - 0.0347152484798	develop - 0.0232492064976	revenu - 0.0280080438885
lifelin - 0.0347152484798	ad - 0.0391159308928	cost - 0.0347152484798
secur - 0.0280080438885	firm - 0.0232492064976	depreci - 0.0347152484798
like - 0.0347152484798	make - 0.0165420019063	issu - 0.0697476194928
difficult -	profit - 0.0195579654464	stabilis -
0.0347152484798	hold - 0.0280080438885	0.0694304969596
futur - 0.0391159308928	dilut - 0.0232492064976	note - 0.0840241316656
, - 0.545886062909	offer - 0.0232492064976	maximum - 0.0347152484798
simultan -	prospect -	1.85 - 0.0347152484798
0.0232492064976	0.0232492064976	constitut -
pari - 0.0464984129952	brighter -	0.0347152484798
london - 0.0195579654464	0.0232492064976	analyst - 0.056016087777
bring - 0.0232492064976	urg - 0.0347152484798	abil - 0.0347152484798
brink - 0.0232492064976	patient - 0.0232492064976	financ - 0.0280080438885
bankruptci -	month - 0.0195579654464	becom - 0.0280080438885
0.0347152484798	uncertainti -	sustain - 0.0347152484798
leav - 0.0195579654464	0.0232492064976	look - 0.0232492064976
current - 0.056016087777	wrestl - 0.0232492064976	flow - 0.0347152484798
sharehold -	reduc - 0.0464984129952	150 - 0.0280080438885
0.162744445483	crippl - 0.0232492064976	200 - 0.0347152484798
alreadi - 0.0280080438885	negoti - 0.0232492064976	ca - 0.0347152484798
seen - 0.0347152484798	's - 0.0661680076253	n't - 0.0347152484798
invest - 0.056016087777	construct -	meet - 0.0232492064976
dwindl - 0.0347152484798	0.0232492064976	roll - 0.0347152484798
54.5 - 0.0232492064976	taken - 0.0195579654464	forward - 0.0280080438885
` - 0.148878017157	half - 0.0232492064976	seven - 0.0280080438885
fix - 0.0347152484798	market - 0.0195579654464	major - 0.0347152484798
cap - 0.0347152484798	busiest - 0.0347152484798	time - 0.0280080438885
payment - 0.0391159308928	cross-channel -	statement -
arrang - 0.0347152484798	0.0232492064976	0.0232492064976
pay - 0.0586738963391	rout - 0.0347152484798	consider -
avail - 0.0347152484798	european -	0.0464984129952
cash - 0.0694304969596	0.0232492064976	work - 0.0195579654464
' - 0.148878017157	ferri - 0.0232492064976	finalis - 0.0195579654464
co-chairman -	strong - 0.0232492064976	agre - 0.0391159308928
0.0464984129952	perform - 0.0391159308928	submit - 0.0232492064976
alastair -	allow - 0.0232492064976	group - 0.0280080438885
0.0232492064976	dividend -	approv - 0.0464984129952
morton - 0.0929968259904	0.0464984129952	probabl - 0.0232492064976
told - 0.0694304969596	10 - 0.0232492064976	spring - 0.0347152484798
report - 0.0586738963391	year - 0.117347792678	1997. - 0.0232492064976
news - 0.0694304969596	french - 0.0232492064976	debt-for-equ -
confer - 0.0694304969596	patrick - 0.0232492064976	0.0232492064976
avoid - 0.0347152484798	ponsol - 0.0232492064976	130 - 0.0232492064976
game - 0.0347152484798	come - 0.056016087777	penc - 0.0697476194928
said - 0.148878017157	earli - 0.0697476194928	10.40 - 0.0232492064976
plan - 0.0464984129952	2004 - 0.0347152484798	franc - 0.0391159308928
provid - 0.0232492064976	ingeni - 0.0347152484798	share - 0.0464984129952
anglo-french -	formula - 0.0347152484798	-- - 0.0280080438885
0.0232492064976	help - 0.0347152484798	level - 0.0165420019063
medium - 0.0280080438885	despit - 0.0347152484798	160 - 0.0232492064976
term - 0.0280080438885	swap - 0.0464984129952	wide - 0.0195579654464
financi - 0.0464984129952	equiti - 0.0347152484798	run - 0.0280080438885
stabil - 0.0232492064976	bond - 0.0347152484798	3.7 - 0.0232492064976
consolid -	abl - 0.0464984129952	convert - 0.0232492064976
0.0195579654464	afford - 0.0347152484798	new - 0.0232492064976
commerci -	annual - 0.0347152484798	instrument -
0.0232492064976	400 - 0.056016087777	0.0232492064976

exist - 0.0232492064976	\$ - 0.0330840038127	reduc - 0.0464984129952
particip - 0.0232492064976) - 0.0827100095316	crippl - 0.0232492064976
choos - 0.0232492064976	debt - 0.139495238986	negoti - 0.0232492064976
free - 0.0232492064976	. - 0.330840038127	's - 0.0661680076253
warrant - 0.0232492064976	long-await - 0.0232492064976	construct - 0.0232492064976
entitl - 0.0232492064976	highli - 0.0232492064976	taken - 0.0195579654464
subscrib - 0.0232492064976	restructur - 0.0195579654464	half - 0.0232492064976
' - 0.0195579654464	nearli - 0.0464984129952	market - 0.0195579654464
39 - 0.0232492064976	futur - 0.0391159308928	cross-channel - 0.0232492064976
end - 0.0195579654464	, - 0.545886062909	european - 0.0232492064976
decemb - 0.0232492064976	simultan - 0.0232492064976	ferri - 0.0232492064976
2003. - 0.0232492064976	pari - 0.0464984129952	strong - 0.0232492064976
suspend - 0.0232492064976	london - 0.0195579654464	perform - 0.0391159308928
week - 0.0232492064976	bring - 0.0232492064976	allow - 0.0232492064976
113.5 - 0.0232492064976	brink - 0.0232492064976	dividend - 0.0464984129952
ahead - 0.0232492064976	leav - 0.0195579654464	10 - 0.0232492064976
resum - 0.0232492064976	sharehold - 0.162744445483	year - 0.117347792678
trade - 0.0232492064976	54.5 - 0.0232492064976	french - 0.0232492064976
tuesday - 0.0232492064976	` - 0.148878017157	patrick - 0.0232492064976
225 - 0.0232492064976	payment - 0.0391159308928	ponsol - 0.0232492064976
'm - 0.0694304969596	pay - 0.0586738963391	earli - 0.0697476194928
hope - 0.0347152484798	' - 0.148878017157	swap - 0.0464984129952
grant - 0.0347152484798	co-chairman - 0.0464984129952	abl - 0.0464984129952
admit - 0.0347152484798	alastair - 0.0232492064976	issu - 0.0697476194928
pretti - 0.0347152484798	morton - 0.0929968259904	meet - 0.0232492064976
angri - 0.0347152484798	report - 0.0586738963391	consider - 0.0464984129952
ask - 0.0347152484798	said - 0.148878017157	work - 0.0195579654464
happen - 0.0347152484798	plan - 0.0464984129952	finalis - 0.0195579654464
reject - 0.0280080438885	provid - 0.0232492064976	agre - 0.0391159308928
nobodi - 0.0694304969596	anglo-french - 0.0232492064976	submit - 0.0232492064976
want - 0.0694304969596	financi - 0.0464984129952	approv - 0.0464984129952
collaps - 0.0347152484798	stabil - 0.0232492064976	probabl - 0.0232492064976
doomsday - 0.0347152484798	consolid - 0.0195579654464	1997. - 0.0232492064976
scenario - 0.0347152484798	commerci - 0.0232492064976	debt-for-equ - 0.0232492064976
1=.6393 - 0.0280080438885	develop - 0.0232492064976	130 - 0.0232492064976
channel - 0.0195579654464	ad - 0.0391159308928	penc - 0.0697476194928
tunnel - 0.0464984129952	firm - 0.0232492064976	10.40 - 0.0232492064976
oper - 0.0697476194928	make - 0.0165420019063	franc - 0.0391159308928
eurotunnel - 0.232492064976	profit - 0.0195579654464	share - 0.0464984129952
monday - 0.0391159308928	dilut - 0.0232492064976	level - 0.0165420019063
announc - 0.0697476194928	offer - 0.0232492064976	160 - 0.0232492064976
deal - 0.136905758125	prospect - 0.0232492064976	wide - 0.0195579654464
bank - 0.136905758125	brighter - 0.0232492064976	3.7 - 0.0232492064976
creditor - 0.0464984129952	patient - 0.0232492064976	convert - 0.0232492064976
45.5 - 0.0232492064976	month - 0.0195579654464	new - 0.0232492064976
percent - 0.0697476194928	uncertainti - 0.0232492064976	instrument - 0.0232492064976
compani - 0.234695585357	wrestl - 0.0232492064976	exist - 0.0232492064976
return - 0.0232492064976		particip - 0.0232492064976
wipe - 0.0232492064976		choos - 0.0232492064976
1.0 - 0.0232492064976		free - 0.0232492064976
billion - 0.0977898272319		warrant - 0.0232492064976
pound - 0.132336015251		
(- 0.0827100095316		

entitl - 0.0232492064976	1.0 - 0.0232492064976	brighter -
subscrib -	billion - 0.0977898272319	0.0232492064976
0.0232492064976	pound - 0.132336015251	patient - 0.0232492064976
' - 0.0195579654464	(- 0.0827100095316	month - 0.0195579654464
39 - 0.0232492064976	\$ - 0.0330840038127	uncertainiti -
end - 0.0195579654464) - 0.0827100095316	0.0232492064976
decemb - 0.0232492064976	debt - 0.139495238986	wrestl - 0.0232492064976
2003. - 0.0232492064976	. - 0.330840038127	reduc - 0.0464984129952
suspend - 0.0232492064976	long-await -	crippl - 0.0232492064976
week - 0.0232492064976	0.0232492064976	negoti - 0.0232492064976
113.5 - 0.0232492064976	highli - 0.0232492064976	's - 0.0661680076253
ahead - 0.0232492064976	restructur -	construct -
resum - 0.0232492064976	0.0195579654464	0.0232492064976
trade - 0.0232492064976	nearli - 0.0464984129952	taken - 0.0195579654464
tuesday - 0.0232492064976	futur - 0.0391159308928	half - 0.0232492064976
225 - 0.0232492064976	, - 0.545886062909	market - 0.0195579654464
1.56 - 0.0	simultan -	cross-channel -
wrangl - 0.0	0.0232492064976	0.0232492064976
owe - 0.0	pari - 0.0464984129952	european -
14.1 - 0.0	london - 0.0195579654464	0.0232492064976
insolv - 0.0	bring - 0.0232492064976	ferri - 0.0232492064976
medium-term - 0.0	brink - 0.0232492064976	strong - 0.0232492064976
substanti - 0.0	leav - 0.0195579654464	perform - 0.0391159308928
achiev - 0.0	sharehold -	allow - 0.0232492064976
date - 0.0	0.162744445483	dividend -
reap - 0.0	54.5 - 0.0232492064976	0.0464984129952
benefit - 0.0	` - 0.148878017157	10 - 0.0232492064976
success - 0.0	payment - 0.0391159308928	year - 0.117347792678
accept - 0.0	pay - 0.0586738963391	french - 0.0232492064976
compromis - 0.0	' - 0.148878017157	patrick - 0.0232492064976
holder - 0.0	co-chairman -	ponsol - 0.0232492064976
syndic - 0.0	0.0464984129952	earli - 0.0697476194928
follow - 0.0	alastair -	swap - 0.0464984129952
secret - 0.0	0.0232492064976	abl - 0.0464984129952
lead - 0.0	morton - 0.0929968259904	issu - 0.0697476194928
extend - 0.0	report - 0.0586738963391	meet - 0.0232492064976
24 - 0.0	said - 0.148878017157	statement -
instruct - 0.0	plan - 0.0464984129952	0.0232492064976
late - 0.0	provid - 0.0232492064976	consider -
outstand - 0.0	anglo-french -	0.0464984129952
8.7 - 0.0	0.0232492064976	work - 0.0195579654464
13.6 - 0.0	medium - 0.0280080438885	finalis - 0.0195579654464
5.8 - 0.0	term - 0.0280080438885	agre - 0.0391159308928
channel - 0.0195579654464	financi - 0.0464984129952	submit - 0.0232492064976
tunnel - 0.0464984129952	stabil - 0.0232492064976	approv - 0.0464984129952
oper - 0.0697476194928	consolid -	probabl - 0.0232492064976
eurotunnel -	0.0195579654464	1997. - 0.0232492064976
0.232492064976	commerci -	debt-for-equ -
monday - 0.0391159308928	0.0232492064976	0.0232492064976
announc - 0.0697476194928	develop - 0.0232492064976	130 - 0.0232492064976
deal - 0.136905758125	ad - 0.0391159308928	penc - 0.0697476194928
bank - 0.136905758125	firm - 0.0232492064976	10.40 - 0.0232492064976
creditor -	make - 0.0165420019063	franc - 0.0391159308928
0.0464984129952	profit - 0.0195579654464	share - 0.0464984129952
45.5 - 0.0232492064976	dilut - 0.0232492064976	level - 0.0165420019063
percent - 0.0697476194928	offer - 0.0232492064976	160 - 0.0232492064976
compani - 0.234695585357	prospect -	wide - 0.0195579654464
return - 0.0232492064976	0.0232492064976	run - 0.0280080438885
wipe - 0.0232492064976		3.7 - 0.0232492064976

convert - 0.0232492064976	. - 0.330840038127	anoth - 0.0
new - 0.0232492064976	restructur -	step - 0.0
instrument -	0.0195579654464	domin - 0.0
0.0232492064976	secur - 0.0280080438885	larg - 0.0
exist - 0.0232492064976	futur - 0.0391159308928	front-end - 0.0
particip -	, - 0.545886062909	necessari - 0.0
0.0232492064976	invest - 0.056016087777	later - 0.0
choos - 0.0232492064976	` - 0.148878017157	rake - 0.0
free - 0.0232492064976	pay - 0.0586738963391	fee - 0.0
warrant - 0.0232492064976	' - 0.148878017157	350 - 0.0
entitl - 0.0232492064976	report - 0.0586738963391	staff - 0.0
subscrib -	said - 0.148878017157	worldwid - 0.0
0.0232492064976	consolid -	far - 0.0
' - 0.0195579654464	0.0195579654464	east - 0.0
39 - 0.0232492064976	ad - 0.0391159308928	treat - 0.0
end - 0.0195579654464	make - 0.0165420019063	core - 0.0
decemb - 0.0232492064976	profit - 0.0195579654464	worth - 0.0
2003. - 0.0232492064976	's - 0.0661680076253	239.2 - 0.0
suspend - 0.0232492064976	taken - 0.0195579654464	place - 0.0
week - 0.0232492064976	market - 0.0195579654464	interim - 0.0
113.5 - 0.0232492064976	perform - 0.0391159308928	result - 0.0
ahead - 0.0232492064976	come - 0.056016087777	publish - 0.0
resum - 0.0232492064976	400 - 0.056016087777	august - 0.0
trade - 0.0232492064976	million - 0.0697476194928	includ - 0.0
tuesday - 0.0232492064976	analyst - 0.056016087777	transit - 0.0
225 - 0.0232492064976	financ - 0.0280080438885	categori - 0.0
1=.6393 - 0.0280080438885	becom - 0.0280080438885	state - 0.0
1.56 - 0.0	look - 0.0232492064976	23 - 0.0
wrangl - 0.0	150 - 0.0280080438885	half-year - 0.0
owe - 0.0	time - 0.0280080438885	june - 0.0
insolv - 0.0	work - 0.0195579654464	30. - 0.0
substanti - 0.0	finalis - 0.0195579654464	primarili - 0.0
achiev - 0.0	group - 0.0280080438885	compris - 0.0
date - 0.0	franc - 0.0391159308928	lend - 0.0
reap - 0.0	level - 0.0165420019063	asset - 0.0
benefit - 0.0	' - 0.0195579654464	unlik - 0.0
success - 0.0	britain - 0.0	long-term - 0.0
accept - 0.0	barclay - 0.0	requir - 0.0
compromis - 0.0	plc - 0.0	signific - 0.0
holder - 0.0	talk - 0.0	clear - 0.0
syndic - 0.0	u.s. - 0.0	mark - 0.0
follow - 0.0	morgan - 0.0	natur - 0.0
secret - 0.0	stanley - 0.0	econom - 0.0
lead - 0.0	global - 0.0	john - 0.0
extend - 0.0	custodi - 0.0	leonard - 0.0
24 - 0.0	busi - 0.0	salomon - 0.0
instruct - 0.0	sourc - 0.0	brother - 0.0
late - 0.0	confirm - 0.0	scale - 0.0
outstand - 0.0	unit - 0.0	economi - 0.0
8.7 - 0.0	sale - 0.0	buisi - 0.0
mountain - 0.0	coy - 0.0	decis - 0.0
monday - 0.0391159308928	respond - 0.0	sum - 0.0
deal - 0.136905758125	press - 0.0	need - 0.0
bank - 0.136905758125	specul - 0.0	comput - 0.0
billion - 0.0977898272319	say - 0.0	scratch - 0.0
pound - 0.132336015251	discuss - 0.0	competit - 0.0
(- 0.0827100095316	explor - 0.0	rang - 0.0
\$ - 0.0330840038127	potenti - 0.0	tradit - 0.0
) - 0.0827100095316	co-oper - 0.0	safekeep - 0.0

lloyd	-	0.0	revenu	-	0.0280080438885	ground	-	0.0
tsb	-	0.0	note	-	0.0840241316656	hear	-	0.0
royal	-	0.0	look	-	0.0232492064976	applic	-	0.0
scotland	-	0.0	forward	-	0.0280080438885	appel	-	0.0
measur	-	0.0	seven	-	0.0280080438885	friday	-	0.0
stock	-	0.0	statement	-		morn	-	0.0
ceas	-	0.0	0.0232492064976			hous	-	0.0
risk-fre	-	0.0	agre	-	0.0391159308928	legal	-	0.0
risk	-	0.0	--	-	0.0280080438885	succeed	-	0.0
rel	-	0.0	level	-	0.0165420019063	stem	-	0.0
low	-	0.0	wide	-	0.0195579654464	1994	-	0.0
guarante	-	0.0	end	-	0.0195579654464	trial	-	0.0
stead	-	0.0	reject	-	0.0280080438885	milan	-	0.0
incom	-	0.0	lead	-	0.0	corrupt	-	0.0
stream	-	0.0	late	-	0.0	charg	-	0.0
total	-	0.0	britain	-	0.0	fininvest	-	0.0
insensit	-	0.0	busi	-	0.0	presid	-	0.0
volum	-	0.0	sourc	-	0.0	fedel	-	0.0
presum	-	0.0	larg	-	0.0	confalonieri	-	0.0
decid	-	0.0	place	-	0.0	rule	-	0.0
small	-	0.0	unlik	-	0.0	earlier	-	0.0
compet	-	0.0	decis	-	0.0	paper	-	0.0
stage	-	0.0	decid	-	0.0	sent	-	0.0
expect	-	0.0	expect	-	0.0	itali	-	0.0
100	-	0.0	italian	-	0.0	link	-	0.0
fifti	-	0.0	tycoon	-	0.0	51	-	0.0
worker	-	0.0	silvio	-	0.0	82	-	0.0
union	-	0.0	berlusconi	-	0.0	surreptiti	-	0.0
bifu	-	0.0	wednesday	-	0.0	remov	-	0.0
attack	-	0.0	lost	-	0.0	use	-	0.0
leak	-	0.0	appeal	-	0.0	crimin	-	0.0
face	-	0.0	high	-	0.0	purpos	-	0.0
uncertain	-	0.0	court	-	0.0	prosecut	-	0.0
sold	-	0.0	transfer	-	0.0	underway	-	0.0
clean	-	0.0	document	-	0.0	bribe	-	0.0
intent	-	0.0	author	-	0.0	inspector	-	0.0
1=.6270	-	0.0	investig	-	0.0	illicit	-	0.0
channel	-	0.0195579654464	alleg	-	0.0	donat	-	0.0
compani	-	0.234695585357	fraud	-	0.0	bettino	-	0.0
pound	-	0.132336015251	fals	-	0.0	craxi	-	0.0
(-	0.0827100095316	account	-	0.0	suspect	-	0.0
\$	-	0.0330840038127	seriou	-	0.0	tv	-	0.0
)	-	0.0827100095316	offic	-	0.0	football	-	0.0
.	-	0.330840038127	seiz	-	0.0	club	-	0.0
,	-	0.545886062909	15	-	0.0	ac	-	0.0
london	-	0.0195579654464	bag	-	0.0	offshor	-	0.0
leav	-	0.0195579654464	april	-	0.0	politician	-	0.0
current	-	0.056016087777	lawyer	-	0.0	judgement	-	0.0
alreadi	-	0.0280080438885	david	-	0.0	point	-	0.0
`	-	0.148878017157	repres	-	0.0	imper	-	0.0
payment	-	0.0391159308928	prime	-	0.0	recognis	-	0.0
"	-	0.148878017157	minist	-	0.0	howev	-	0.0
said	-	0.148878017157	judg	-	0.0	convict	-	0.0
make	-	0.0165420019063	lord	-	0.0	anyth	-	0.0
hold	-	0.0280080438885	justic	-	0.0	noth	-	0.0
month	-	0.0195579654464	simon	-	0.0	thought	-	0.0
's	-	0.0661680076253	brown	-	0.0	rais	-	0.0
year	-	0.117347792678	mr	-	0.0	presumpt	-	0.0
million	-	0.0697476194928	gag	-	0.0	guilt	-	0.0

basi - 0.0
request - 0.0
justifi - 0.0
offenc - 0.0
question - 0.0
polit - 0.0
sfo - 0.0
gone - 0.0
fish - 0.0
expedit - 0.0
unspecif - 0.0
search - 0.0
illeg - 0.0
english - 0.0
law - 0.0
describ - 0.0
case - 0.0

wide-rang - 0.0
multi-facet - 0.0
intern - 0.0
involv - 0.0
far-reach - 0.0
number - 0.0
individu - 0.0
connect - 0.0
larger - 0.0
happi - 0.0
fight - 0.0
strengthen - 0.0
welcom - 0.0
main - 0.0
argument - 0.0
speedi - 0.0
job - 0.0

today - 0.0
british - 0.0
senior - 0.0
assist - 0.0
director - 0.0
chri - 0.0
dickson - 0.0
awar - 0.0
mean - 0.0
arriv - 0.0
proceed - 0.0
given - 0.0
immedi - 0.0
home - 0.0
1=.6260 - 0.0