Reading the dataset

In [1]:

```
import pandas as pd
text = pd.read_csv('Full-Economic-News-DFE-839861.csv', encoding = 'ISO-8859-1')
text.head()
```

Out[1]:

	_unit_id	_golden	_unit_state	_trusted_judgments	_last_judgment_at	positivity	positivity:confidence	relevance	relevance:confid
0	842613455	False	finalized	3	12/5/15 17:48	3.0	0.6400	yes	
1	842613456	False	finalized	3	12/5/15 16:54	NaN	NaN	no	
2	842613457	False	finalized	3	12/5/15 1:59	NaN	NaN	no	
3	842613458	False	finalized	3	12/5/15 2:19	NaN	0.0000	no	
4	842613459	False	finalized	3	12/5/15 17:48	3.0	0.3257	yes	

In [2]:

```
text.shape
```

Out[2]:

```
(8000, 15)
```

So we have 8000 articles (big corpus)

For us only the 'text' column is useful

Here, we can see that there are some tags like '< /br>', we dont need that so we need to get rid of them.

Also we now need to lowercase and tokenize the corpus.

In [3]:

```
import nltk
nltk.download('punkt')
import math
```

```
[nltk_data] Error loading punkt: <urlopen error [Errno 11001]
[nltk_data]     getaddrinfo failed>
```

In [4]:

```
articles = [x.replace('</br>', ' ') for x in text['text'][:200]]
# using first 200 articles
all_articles = []
for a in articles:
    tokens = nltk.word_tokenize(a)
    all_articles.append([word.lower() for word in tokens if word.isalpha()])

total_tokens = nltk.word_tokenize(' '.join(articles))
all_words = [word.lower() for word in total_tokens if word.isalpha()]

print(all_words[:150])
print(articles[:1])
print('\ntotal number of words:', len(all_words))
```

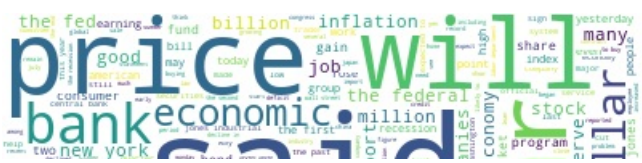
```
['new', 'york', 'yields', 'on', 'most', 'certificates', 'of', 'deposit', 'offered', 'by', 'major',
'banks', 'dropped', 'more', 'than', 'a', 'tenth', 'of', 'a', 'percentage', 'point', 'in', 'the', '
latest', 'week', 'reflecting', 'the', 'overall', 'decline', 'in', 'interest', 'rates', 'on', 'or',
'consumer', 'cds', 'sold', 'directly', 'by', 'banks', 'the', 'average', 'yield', 'on', 'deposits',
'fell', 'to', 'from', 'in', 'the', 'week', 'ended', 'yesterday', 'according', 'to', 'an',
'survey', 'by', 'banxquote', 'money', 'markets', 'a', 'wilmington', 'information', 'service',
'on', 'consumer', 'deposits', 'the', 'average', 'yield', 'sank', 'to', 'from', 'the', 'week', 'bef
ore', 'according', 'to', 'banxquote', 'two', 'banks', 'in', 'the', 'banxquote', 'survey',
'citibank', 'in', 'new', 'york', 'and', 'corestates', 'in', 'pennsylvania', 'are', 'paying',
'less', 'than', 'on', 'threemonth', 'cds', 'declines', 'were', 'somewhat', 'smaller', 'on',
'consumer', 'cds', 'which', 'eased', 'to', 'from', 'banxquote', 'said', 'yields', 'on', 'and', 'tr
easury', 'bills', 'sold', 'at', 'monday', 'auction', 'plummeted', 'more', 'than', 'a', 'fifth', 'o
f', 'a', 'percentage', 'point', 'from', 'the', 'previous', 'week', 'to', 'and', 'respectively', 't
he', 'wall', 'street', 'journal', 'online', 'the', 'morning', 'brief', 'a', 'look', 'at']
['NEW YORK -- Yields on most certificates of deposit offered by major banks dropped more than a
tenth of a percentage point in the latest week, reflecting the overall decline in short-term
interest rates.  On small-denomination, or "consumer," CDs sold directly by banks, the average yie
ld on six-month deposits fell to 5.49% from 5.62% in the week ended yesterday, according to an 18-
bank survey by Banxquote Money Markets, a Wilmington, Del., information service.  On three-month "
consumer" deposits, the average yield sank to 5.29% from 5.42% the week before, according to
Banxquote. Two banks in the Banxquote survey, Citibank in New York and CoreStates in Pennsylvania,
are paying less than 5% on threemonth small-denomination CDs.  Declines were somewhat smaller on f
ive-year consumer CDs, which eased to 7.37% from 7.45%, Banxquote said.  Yields on three-month and
six-month Treasury bills sold at Monday\'s auction plummeted more than a fifth of a percentage poi
nt from the previous week, to 5.46% and 5.63%, respectively.']

total number of words: 43157
```

Time to create our favourite, 'The WordCloud'

In [5]:

```
# I am gonna import the word cloud library
from wordcloud import WordCloud

# for creating a word cloud we need to pull all our text together
# I am gonna do that by joining all the processed_data together in a single variable
para = ','.join(list(all_words))

# Lets create an instance for word cloud
wordcloud = WordCloud(background_color = 'white',
                      max_words = 5000,
                      contour_width = 3,
                      contour_color = 'steelblue')

# generate our word cloud
wordcloud.generate(para)

# Lets visualize it
wordcloud.to_image()
```

Out[5]:



Raw Frequency

In [6]:

```
min_threshold_tau = 5
words_len = len(all_words)
```

In [7]:

```
f = {}
index = {}
for i in range(words_len):
    word = all_words[i]
    if word not in index.keys():
        index[word] = [i]
    else:
        index[word].append(i)
```

In [8]:

```
index_allwords = index.copy()
index_allwords['risen']
```

Out[8]:

```
[12648, 35905, 36095, 37722, 43156]
```

In [9]:

```
while len(index.keys()) > 0:
    index_p = {}
    for j in index.keys():
        if len(index[j]) >= min_threshold_tau:
            f[j] = len(index[j])
            for k in index[j]:
                if k+1 < words_len:
                    new_phrase = j + u' ' + all_words[k + 1]
                    if new_phrase not in index_p.keys():
                        index_p[new_phrase] = [k + 1]
                    else:
                        index_p[new_phrase].append(k + 1)
    index = index_p
    print(list(index.keys())[:10])
```

```
['new york', 'new restrictions', 'new limits', 'new jersey', 'new steps', 'new highs', 'new law',
'new aid', 'new recruits', 'new employees']
['new york yields', 'new york and', 'new york times', 'new york that', 'new york indecision', 'new
york day', 'new york the', 'new york trade', 'new york trading', 'new york in']
['new york the euro', 'new york the wall', 'new york the dollar', 'new york the market', 'new york
the statement', 'new york the federal', 'new york the risk', 'new york a sharp', 'new york a promi
nent', 'new york a plunge']
['new york stock exchange had', 'new york stock exchange more', 'new york stock exchange already',
'new york stock exchange also', 'new york stock exchange in', 'new york stock exchange composite',
'of million or cents a', 'more than a year have', 'more than a year as', 'more than a year ago']
['of million or cents a diluted', 'of million or cents a share', 'the new york stock exchange had'
, 'the new york stock exchange more', 'the new york stock exchange already', 'the new york stock e
xchange also', 'the new york stock exchange in', 'the new york stock exchange composite', 'the dow
jones industrial average fell', 'the dow jones industrial average from']
[]
```

In [10]:

```
print('total phrases above threshold:', len(f.keys()))
print('sample phrases:', list(f.keys())[:10])
```

```
total phrases above threshold: 2272
sample phrases: ['new', 'york', 'yields', 'on', 'most', 'of', 'deposit', 'offered', 'by', 'major']
```

In [11]:

```
f.keys()
```

Out[11]:

```
dict_keys(['new', 'york', 'yields', 'on', 'most', 'of', 'deposit', 'offered', 'by', 'major',
'banks', 'dropped', 'more', 'than', 'a', 'percentage', 'point', 'in', 'the', 'latest', 'week',
'overall', 'decline', 'interest', 'rates', 'or', 'consumer', 'sold', 'directly', 'average',
'yield', 'fell', 'to', 'from', 'ended', 'yesterday', 'according', 'an', 'survey', 'money',
'markets', 'information', 'service', 'sank', 'before', 'two', 'and', 'are', 'less', 'declines', 'w
ere', 'somewhat', 'smaller', 'which', 'eased', 'said', 'treasury', 'bills', 'at', 'monday', 'aucti
on', 'previous', 'wall', 'street', 'journal', 'online', 'morning', 'look', 'day', 'biggest',
'news', 'is', 'every', 'business', 'sign', 'up', 'for', 'here', 'friday', 'with', 'congress', 'out
', 'its', 'summer', 'americans', 'into', 'weekend', 'bush', 'administration', 'states', 'federal',
'government', 'will', 'make', 'it', 'national', 'children', 'insurance', 'program', 'cover',
'families', 'state', 'health', 'was', 'created', 'help', 'whose', 'could', 'but', 'did',
'medicaid', 'officials', 'tell', 'times', 'that', 'changes', 'income', 'focus', 'become',
'private', 'man', 'wrote', 'there', 'would', 'be', 'including', 'plan', 'coverage', 'levels', 'fam
ily', 'three', 'four', 'under', 'limits', 'making', 'have', 'spend', 'one', 'year', 'any', 'wants'
, 'washington', 'least', 'programs', 'as', 'reports', 'no', 'can', 'currently', 'such',
'director', 'policy', 'group', 'since', 'many', 'above', 'ca', 'this', 'commissioner', 'human', 's
ervices', 'cause', 'our', 'already', 'been', 'taking', 'other', 'parents', 'moving', 'their', 'som
e', 'million', 'come', 'end', 'next', 'month', 'if', 'does', 'larger', 'political', 'white',
'house', 'against', 'democrats', 'republicans', 'effort', 'banking', 'reform', 'senate', 'agreed',
'drop', 'efforts', 'allow', 'further', 'securities', 'several', 'committee', 'bill', 'face',
'when', 'starts', 'measure', 'perhaps', 'today', 'also', 'bank', 'last', 'began', 'second',
'attempt', 'pass', 'after', 'part', 'because', 'over', 'how', 'vote', 'fund', 'regulation',
'reserve', 'board', 'ability', 'keep', 'loans', 'until', 'final', 'give', 'leadership', 'time', 's
upport', 'working', 'hard', 'significant', 'john', 'who', 'should', 'include', 'statistics',
'costs', 'drug', 'well', 'known', 'billion', 'per', 'december', 'testimony', 'labor', 'safety', 'm
anagement', 'pressure', 'reduce', 'productivity', 'losses', 'risk', 'personal', 'caused',
'workers', 'sales', 'growing', 'laboratories', 'manufacturers', 'testing', 'common', 'argument', '
following', 'your', 'competitors', 'starting', 'you', 'do', 'work', 'force', 'all', 'companies', '
widespread', 'however', 'they', 'employees', 'civil', 'legal', 'texas', 'his', 'results', 'test',
'cases', 'middle', 'without', 'number', 'important', 'managers', 'about', 'signs', 'initial', 'sta
tus', 'company', 'had', 'may', 'place', 'begin', 'widely', 'often', 'even', 'control', 'study', 'h
igh', 'dollar', 'traders', 'monthly', 'employment', 'report', 'release', 'slightly', 'weaker',
'both', 'euro', 'yen', 'market', 'participants', 'top', 'japanese', 'issue', 'reforms', 'late', 'c
ents', 'stronger', 'level', 'wednesday', 'trading', 'down', 'swiss', 'francs', 'while', 'raise', '
head', 'during', 'trade', 'first', 'roughly', 'weeks', 'advantage', 'stock', 'worries',
'economic', 'recovery', 'ahead', 'stocks', 'declined', 'investors', 'domestic', 'growth',
'continued', 'concerns', 'take', 'economy', 'dow', 'jones', 'industrial', 'points', 'lowest', 'clo
se', 'america', 'technology', 'issued', 'earnings', 'outlook', 'current', 'quarter', 'fiscal',
's', 'p', 'index', 'nasdaq', 'composite', 'lost', 'royal', 'canada', 'posted', 'increased', 'net',
'incomes', 'quarters', 'helped', 'them', 'recent', 'slowdown', 'volatility', 'canadian', 'better',
'has', 'partly', 'credit', 'relatively', 'cautious', 'analyst', 'capital', 'problems', 'later', 'r
ecession', 'added', 'gave', 'extra', 'put', 'he', 'april', 'based', 'dollars', 'us', 'fully', 'sha
re', 'earlier', 'total', 'rose', 'period', 'included', 'full', 'recently', 'businesses', 'firm', '
south', 'corp', 'jump', 'people', 'think', 'inflation', 'driven', 'back', 'early', 'july', 'then',
'secretary', 'reagan', 'seems', 'within', 'community', 'announced', 'spending', 'care', 'decades',
'cost', 'start', 'american', 'social', 'contract', 'nation', 'providers', 'so', 'almost', 'left',
'alone', 'best', 'whatever', 'hospitals', 'usual', 'only', 'very', 'generally', 'ever', 'taken', '
individual', 'these', 'not', 'prices', 'among', 'climbed', 'experts', 'thought', 'short', 'terms',
'deal', 'medical', 'doing', 'good', 'thus', 'itself', 'buyers', 'around', 'products', 'paid', 'thi
rd', 'party', 'either', 'public', 'sector', 'election', 'made', 'came', 'expenditures', 'through',
'rise', 'greater', 'consumption', 'price', 'limited', 'set', 'basis', 'return', 'financial', 'cont
racts', 'provide', 'necessary', 'flat', 'annual', 'acquire', 'valued', 'what', 'analysts',
'expect', 'series', 'closely', 'genetic', 'boom', 'seen', 'san', 'makes', 'systems', 'data',
'order', 'differences', 'might', 'large', 'research', 'june', 'agreement', 'exchange', 'each', 'sh
ares', 'traded', 'suggest', 'things', 'firms', 'foreign', 'currencies', 'although', 'bond',
'advance', 'tuesday', 'rally', 'weakening', 'europe', 'falling', 'lower', 'higher', 'investments',
'attractive', 'closed', 'we', 'off', 'my', 'continue', 'between', 'now', 'assistant', 'vice',
'president', 'discount', 'west', 'german', 'marks', 'thursday', 'strengthened', 'currency', 'gener
al', 'kept', 'negative', 'red', 'surge', 'just', 'auto', 'maker', 'gained', 'considering', 'buying
', 'corporate', 'big', 'finished', 'standard', 'poor', 'pushing', 'rising', 'advanced', 'mci', 'wo
rldcom', 'internet', 'calls', 'expectations', 'revenue', 'grew', 'amid', 'whether', 'bid',
'communications', 'independent', 'remains', 'talks', 'potential', 'acquisition', 'sides', 'fall',
'value', 'compared', 'excluding', 'gain', 'operating', 'expected', 'consensus', 'call', 'author',
```

'i', 'mean', 'depression', 'makers', 'deflation', 'seem', 'like', 'bad', 'welcome', 'getting',
'those', 'where', 'got', 'longer', 'course', 'really', 'never', 'thing', 'steady', 'profits', 'buy
', 'demand', 'cut', 'japan', 'difficult', 'commercial', 'loan', 'volume', 'grow', 'forecast', 'pri
cing', 'projected', 'increase', 'forecasts', 'gross', 'product', 'years', 'estimates', 'richard',
'senior', 'economist', 'going', 'need', 'believe', 'still', 'another', 'junk', 'bonds', 'turn', 'i
ssues', 'highs', 'debt', 'way', 'international', 'germany', 'rate', 'payments', 'unemployment', 'f
ood', 'rather', 'cash', 'consumers', 'single', 'needs', 'machines', 'homes', 'show', 'future', 'pr
ospects', 'change', 'right', 'away', 'long', 'local', 'behind', 'inventories', 'same', 'country',
'instead', 'indian', 'independence', 'central', 'august', 'savings', 'standards', 'reach',
'politicians', 'passed', 'law', 'days', 'reserves', 'used', 'paper', 'power', 'beginning', 'hope',
'yet', 'term', 'specific', 'use', 'agreements', 'bankers', 'desire', 'monetary', 'system', 'core',
'toward', 'beyond', 'anticipated', 'provided', 'powerful', 'followed', 'fed', 'action', 'remain',
'meeting', 'lift', 'funds', 'raised', 'decision', 'despite', 'activity', 'pace', 'clinton', 'aid',
'package', 'mexico', 'quickly', 'intraday', 'lows', 'minutes', 'statement', 'little', 'session', '
solid', 'jobs', 'own', 'home', 'small', 'her', 'older', 'near', 'job', 'situation', 'sometimes', '
sort', 'life', 'took', 'everything', 'much', 'she', 'says', 'opportunity', 'mark', 'investment',
'improving', 'likely', 'weak', 'held', 'consulting', 'reading', 'quarterly', 'improved', 'minus',
'fourth', 'suggests', 'zero', 'slower', 'indicate', 'seasonally', 'adjusted', 'step', 'below', 'ro
bust', 'enough', 'given', 'economists', 'holiday', 'season', 'industry', 'half', 'sharply',
'hiring', 'evidence', 'giant', 'google', 'profit', 'reported', 'record', 'follow', 'strong', 'near
ly', 'tech', 'others', 'meanwhile', 'reached', 'again', 'competition', 'goods', 'months', 'chief',
'executive', 'view', 'started', 'chairman', 'alan', 'joint', 'go', 'panel', 'members', 'nothing',
'plans', 'venture', 'know', 'energy', 'lot', 'takes', 'managing', 'former', 'understand', 'past',
'create', 'seven', 'eight', 'try', 'sense', 'stage', 'liquidity', 'get', 'portfolio', 'though', 'o
il', 'led', 'key', 'say', 'benchmark', 'low', 'extended', 'persuade', 'signal', 'stay', 'city', 'm
ayor', 'cuts', 'due', 'budget', 'deficit', 'spokesman', 'largest', 'office', 'unions', 'pushed', '
sharp', 'inflationary', 'tumbled', 'william', 'inc', 'slump', 'coming', 'problem', 'amount',
'momentum', 'november', 'department', 'producer', 'weakness', 'volatile', 'sectors', 'edged', 'chi
cago', 'gains', 'across', 'equipment', 'showed', 'october', 'production', 'bear', 'gold',
'turning', 'futures', 'commodity', 'tightening', 'fuel', 'investor', 'warburg', 'pincus',
'bought', 'electric', 'pension', 'university', 'slipped', 'performance', 'losing', 'position', 'ju
mped', 'great', 'rapid', 'figures', 'slowing', 'heavy', 'borrowing', 'increases', 'run', 'gas', 'm
organ', 'stanley', 'struggling', 'policies', 'broad', 'commerce', 'maintaining', 'moved',
'household', 'slow', 'confidence', 'hit', 'gauge', 'sentiment', 'move', 'ceiling', 'date',
'authority', 'pay', 'obama', 'trying', 'deficits', 'trillion', 'assets', 'mutual', 'world', 'serio
us', 'six', 'additional', 'crist', 'governor', 'republican', 'old', 'primary', 'democratic',
'reason', 'leading', 'name', 'closer', 'jobless', 'global', 'page', 'disappointing', 'boost', 'med
icare', 'being', 'decade', 'bubble', 'housing', 'called', 'crash', 'once', 'see', 'households', 'f
iling', 'claims', 'benefits', 'five', 'improvement', 'historically', 'few', 'employers', 'grown',
'career', 'development', 'result', 'laid', 'doubt', 'argued', 'asian', 'crisis', 'official', 'janu
ary', 'education', 'numbers', 'bargaining', 'wages', 'seasonal', 'factors', 'released',
'purchasing', 'orders', 'easing', 'court', 'campaign', 'contributions', 'missouri', 'different', '
rights', 'huge', 'needed', 'influence', 'corruption', 'bring', 'critics', 'case', 'limit',
'david', 'prime', 'question', 'cutting', 'finding', 'alternative', 'retail', 'process', 'shot', 'w
ar', 'too', 'aggressive', 'london', 'find', 'line', 'lead', 'manager', 'boston', 'direction',
'fresh', 'continues', 'ims', 'executives', 'approved', 'leader', 'operations', 'mortgage', 'sale',
'gets', 'unit', 'electronic', 'rowe', 'india', 'ago', 'hoping', 'stake', 'uti', 'rivals',
'eventually', 'february', 'increasingly', 'highest', 'saw', 'trader', 'raising', 'stores', 'kind',
'worse', 'citizens', 'class', 'northview', 'concern', 'winners', 'bit', 'announce', 'industrials',
'import', 'ease', 'pressures', 'voted', 'target', 'proposed', 'software', 'hand', 'told', 'meet',
'possible', 'supply', 'suppliers', 'materials', 'opening', 'organization', 'plunge', 'selling',
'estimated', 'far', 'bundesbank', 'intervention', 'dealers', 'especially', 'push', 'estimate',
'surveyed', 'open', 'robert', 'finance', 'adobe', 'continuing', 'frame', 'closing', 'me',
'reflect', 'economies', 'china', 'revised', 'gap', 'imports', 'hold', 'deep', 'measures', 'soon',
'fundamentals', 'rescue', 'range', 'positions', 'deals', 'trust', 'holdings', 'list', 'something',
'procter', 'always', 'professor', 'school', 'march', 'why', 'looking', 'want', 'strategy',
'worst', 'fact', 'reflected', 'commodities', 'fear', 'sell', 'petroleum', 'bureau', 'feel',
'recessions', 'real', 'gdp', 'risen', 'soared', 'airlines', 'represents', 'air', 'association', 'b
en', 'main', 'worth', 'turned', 'romney', 'attacks', 'tight', 'security', 'affect',
'investigation', 'familiar', 'lawmakers', 'payroll', 'tax', 'traditional', 'agency', 'history', 'p
rotection', 'payment', 'bankruptcy', 'commission', 'settlement', 'wo', 'region', 'transportation',
'helping', 'attention', 'taxes', 'shows', 'highly', 'typically', 'anyone', 'broader', 'output', 't
hinking', 'lawrence', 'silver', 'ounce', 'reduced', 'bernanke', 'vietnam', 'leaders', 'worried', '
faster', 'thailand', 'previously', 'rule', 'clear', 'summers', 'economics', 'modest', 'natural', '
european', 'battle', 'exports', 'form', 'tools', 'fix', 'done', 'options', 'records', 'offer', 'st
rategies', 'lending', 'mixed', 'comments', 'save', 'sessions', 'positive', 'asia', 'finally', 'sur
prise', 'wide', 'margin', 'probably', 'strategist', 'includes', 'september', 'strength',
'separate', 'construction', 'mostly', 'conditions', 'mortgages', 'institutions', 'addition',
'taylor', 'regulators', 'expects', 'conference', 'nine', 'nobel', 'prize', 'admit', 'myerson', 'me
chanism', 'design', 'works', 'rules', 'harvard', 'regulatory', 'gates', 'ford', 'offering', 'found
', 'southeast', 'freddie', 'mac', 'loss', 'ending', 'argentina', 'imf', 'emergency', 'trustees', '
retirement', 'odds', 'teachers', 'happen', 'means', 'indexes', 'singapore', 'comes', 'believed', '
women', 'adelphia', 'flow', 'themselves', 'planned', 'invest', 'weekly', 'averaged',
'archipelago', 'atlas', 'fedex', 'legislation', 'bay', 'appeared', 'paul', 'art', 'downturn', 'hap
pens', 'schorsch', 'expanding', 'units', 'coalition', 'seniors', 'pia', 'cftc', 'latin',
'project', 'client', 'ultrashort', 'schwab', 'police', 'yahoo', 'new york', 'york the', 'york a',
'york stock', 'on friday', 'on its', 'on the', 'on a', 'on tuesday', 'on their', 'on thursday', 'm

ost of', 'of a', 'of the', 'of this', 'of companies', 'of their', 'of some', 'of america', 'of can
ada', 'of all', 'of million', 'of it', 'of about', 'of its', 'of economic', 'of business', 'of dol
lars', 'of these', 'of in', 'of government', 'of and', 'of an', 'of people', 'of new', 'of money',
'of bonds', 'of them', 'by the', 'by a', 'banks to', 'more than', 'than a', 'than of', 'than the',
'than in', 'than billion', 'a larger', 'a year', 'a share', 'a drop', 'a basis', 'a series', 'a bi
g', 'a gain', 'a rate', 'a little', 'a increase', 'a seasonally', 'a annual', 'a strong', 'a
decline', 'a level', 'a lot', 'a billion', 'a sharp', 'a sign', 'a rise', 'a record', 'a month', '
a result', 'a more', 'a number', 'a large', 'a few', 'a bit', 'a good', 'a new', 'a high', 'a
professor', 'a recession', 'a million', 'a major', 'percentage point', 'point of', 'in the', 'in i
nterest', 'in new', 'in one', 'in an', 'in part', 'in some', 'in a', 'in its', 'in what', 'in late
', 'in their', 'in any', 'in this', 'in december', 'in may', 'in recent', 'in november', 'in octob
er', 'in when', 'in july', 'in june', 'in august', 'in january', 'in and', 'in annual', 'in april'
, 'in washington', 'in more', 'in which', 'in fact', 'in many', 'in september', 'the latest', 'the
overall', 'the average', 'the week', 'the previous', 'the wall', 'the day', 'the bush', 'the feder
al', 'the state', 'the new', 'the program', 'the end', 'the white', 'the securities', 'the
senate', 'the house', 'the bank', 'the bill', 'the risk', 'the results', 'the second', 'the study'
, 'the dollar', 'the euro', 'the yen', 'the issue', 'the first', 'the economic', 'the employment',
'the economy', 'the dow', 'the current', 'the next', 'the s', 'the nasdaq', 'the recent', 'the can
adian', 'the last', 'the quarter', 'the period', 'the reagan', 'the and', 'the american', 'the nat
ion', 'the terms', 'the deal', 'the late', 'the public', 'the rise', 'the firm', 'the two', 'the d
ecline', 'the bond', 'the japanese', 'the year', 'the stock', 'the standard', 'the fall', 'the tal
ks', 'the company', 'the is', 'the markets', 'the only', 'the country', 'the central', 'the law',
'the fed', 'the rate', 'the session', 'the time', 'the outlook', 'the group', 'the fourth', 'the t
hird', 'the recession', 'the pace', 'the same', 'the past', 'the capital', 'the city', 'the larges
t', 'the most', 'the treasury', 'the labor', 'the board', 'the price', 'the commerce', 'the
gains', 'the university', 'the government', 'the world', 'the jobless', 'the housing', 'the
crash', 'the biggest', 'the number', 'the recovery', 'the job', 'the unemployment', 'the numbers',
'the department', 'the amount', 'the market', 'the way', 'the highest', 'the big', 'the trade', 't
he currency', 'the report', 'the budget', 'the deficit', 'the mark', 'the start', 'the president',
'the potential', 'the other', 'the national', 'the financial', 'the consensus', 'the money', 'the
final', 'the index', 'the region', 'the global', 'the mortgage', 'the imf', 'the news', 'the cftc'
, 'the project', 'the art', 'decline in', 'interest rates', 'interest rate', 'rates and', 'rates i
n', 'rates that', 'rates to', 'or to', 'or at', 'or the', 'or cents', 'or more', 'or that', 'or a'
, 'consumer credit', 'consumer prices', 'yield on', 'fell to', 'fell points', 'to from', 'to an',
'to and', 'to the', 'to help', 'to its', 'to have', 'to allow', 'to keep', 'to give', 'to reduce',
'to make', 'to raise', 'to million', 'to a', 'to be', 'to do', 'to buy', 'to billion', 'to think',
'to take', 'to start', 'to continue', 'to go', 'to get', 'to say', 'to pay', 'to their', 'to find'
, 'to investors', 'to push', 'to in', 'to see', 'to increase', 'to more', 'to people', 'from in',
'from the', 'from a', 'from last', 'from an', 'yesterday as', 'according to', 'an index', 'an
average', 'an annual', 'an ounce', 'markets and', 'before the', 'two years', 'and the', 'and new',
'and for', 'and some', 'and other', 'and a', 'and investors', 'and interest', 'and with', 'and has
', 'and energy', 'and they', 'and that', 'and more', 'and many', 'and inflation', 'and in', 'and s
o', 'and analysts', 'and it', 'are the', 'are expected', 'are likely', 'less than', 'declines in',
'which has', 'which is', 'said it', 'said the', 'said that', 'said a', 'said in', 'treasury
secretary', 'at the', 'at least', 'at a', 'at its', 'at to', 'at billion', 'wall street', 'street
journal', 'look at', 'is to', 'is the', 'is one', 'is a', 'is it', 'is expected', 'is that', 'is a
t', 'is still', 'is likely', 'is about', 'is up', 'is no', 'is not', 'business investment', 'sign
that', 'sign of', 'up for', 'up from', 'up their', 'up in', 'up to', 'up the', 'for the', 'for a',
'for their', 'for each', 'for more', 'friday the', 'with the', 'with an', 'with a', 'with its',
'with billion', 'out of', 'its own', 'into a', 'into the', 'bush administration', 'federal
reserve', 'will make', 'will be', 'it is', 'it will', 'it would', 'it was', 'it could', 'it also',
'it had', 'it a', 'it has', 'insurance companies', 'program to', 'health care', 'was at', 'was a',
'was the', 'was up', 'could be', 'but the', 'but there', 'but it', 'but they', 'but not', 'but tha
t', 'that the', 'that is', 'that would', 'that have', 'that was', 'that are', 'that has', 'that we'
, 'that it', 'that in', 'that they', 'that many', 'that a', 'that he', 'that its', 'that some',
'that means', 'become a', 'there is', 'there a', 'there are', 'there were', 'would be', 'would hav
e', 'be the', 'be a', 'plan to', 'three months', 'three years', 'four years', 'under the', 'have t
o', 'have also', 'have been', 'have the', 'have a', 'one of', 'year according', 'year the', 'year
to', 'year earlier', 'year but', 'year and', 'year ago', 'year end', 'washington the', 'as the', 'a
s investors', 'as of', 'as one', 'as a', 'as they', 'as much', 'as many', 'as well', 'can be', 'su
ch as', 'such a', 'director of', 'policy makers', 'since april', 'since the', 'many economists', '
many investors', 'this is', 'this year', 'this was', 'this week', 'this past', 'been a', 'their bi
lls', 'their own', 'some of', 'million or', 'million in', 'end of', 'next month', 'next year', 'if
you', 'if they', 'if the', 'if it', 'if we', 'white house', 'against the', 'effort to', 'agreed to
', 'drop in', 'efforts to', 'when the', 'when he', 'when it', 'when a', 'bank of', 'last week',
'last year', 'last month', 'second quarter', 'attempt to', 'after the', 'after a', 'part of', 'bec
ause of', 'because it', 'because the', 'over the', 'how the', 'how much', 'reserve board',
'ability to', 'keep the', 'give the', 'time to', 'time in', 'support for', 'who are', 'should be',
'should have', 'known as', 'billion in', 'labor department', 'labor market', 'pressure on', 'reduc
e the', 'risk of', 'sales and', 'following the', 'starting to', 'all the', 'all of', 'all that',
'companies that', 'companies to', 'companies and', 'however the', 'they have', 'they were', 'they
had', 'they do', 'they are', 'number of', 'about the', 'about a', 'about million', 'about to', 'ab
out of', 'signs of', 'had been', 'may be', 'may have', 'even as', 'even if', 'even more', 'dollar
was', 'dollar and', 'traders said', 'employment report', 'market that', 'market the', 'market and'
, 'market is', 'issue of', 'late wednesday', 'cents or', 'cents a', 'down from', 'while the', 'hea
d of', 'during the', 'trade deficit', 'first time', 'first quarter', 'stock prices', 'stock market
', 'stock exchange', 'economic recovery', 'economic growth', 'economic crisis', 'ahead of', 'stock

```
s rose', 'declined to', 'investors are', 'growth in', 'continued to', 'concerns about', 'economy t
he', 'economy has', 'economy and', 'economy is', 'dow jones', 'jones industrial', 'industrial aver
age', 'points or', 'points to', 'close to', 'outlook for', 'quarter and', 'quarter of', 's p', 'in
dex fell', 'index of', 'index rose', 'nasdaq composite', 'nasdaq stock', 'composite index', 'royal
bank', 'increased to', 'them to', 'recent years', 'better than', 'has been', 'has to', 'has
become', 'has a', 'credit markets', 'capital markets', 'capital gains', 'recession in', 'he said',
'he says', 'share of', 'rose to', 'rose points', 'rose in', 'people are', 'think of', 'inflation i
s', 'inflation and', 'back in', 'social security', 'only a', 'prices of', 'prices and', 'prices
on', 'prices are', 'terms of', 'around the', 'third quarter', 'came in', 'through the', 'rise in',
'price index', 'price of', 'price increases', 'financial markets', 'financial crisis', 'annual rat
e', 'valued at', 'analysts said', 'series of', 'data the', 'might be', 'traded at', 'although
the', 'bond prices', 'bond market', 'lower interest', 'higher interest', 'closed at', 'we had', 'w
e have', 'we are', 'continue to', 'between the', 'now is', 'vice president', 'president of',
'surge in', 'standard poor', 'poor index', 'mci worldcom', 'whether the', 'value of', 'compared wi
th', 'compared to', 'expected to', 'like it', 'like the', 'demand for', 'cut in', 'japan and',
'increase in', 'increase the', 'increase of', 'years the', 'years ago', 'years old', 'economist at
', 'going to', 'believe the', 'bonds and', 'rate of', 'rate the', 'rate in', 'rate for', 'rate
cut', 'unemployment rate', 'food and', 'rather than', 'show that', 'instead of', 'central bank', '
central bankers', 'central banks', 'days of', 'used to', 'monetary policy', 'fed has', 'fed
chairman', 'fed officials', 'funds and', 'pace of', 'much as', 'much of', 'she says', 'opportunity
to', 'investment in', 'likely to', 'fourth quarter', 'seasonally adjusted', 'below the', 'enough t
o', 'half of', 'meanwhile the', 'goods and', 'months of', 'months ago', 'chief economist',
'chairman of', 'chairman ben', 'plans to', 'energy prices', 'lot of', 'managing director', 'past w
eek', 'past three', 'past two', 'try to', 'oil prices', 'led by', 'due to', 'budget deficit',
'amount of', 'department said', 'producer price', 'weakness in', 'gains have', 'across the', 'warb
urg pincus', 'university of', 'morgan stanley', 'commerce department', 'confidence in', 'trying to
', 'mutual funds', 'few weeks', 'purchasing managers', 'retail sales', 'shot up', 'continues to',
'mortgage rates', 'ago the', 'raising the', 'kind of', 'concern that', 'surveyed by', 'imports of'
, 'professor of', 'professor at', 'want to', 'ben bernanke', 'familiar with', 'natural gas',
'mechanism design', 'freddie mac', 'new york the', 'new york a', 'new york stock', 'york stock exc
hange', 'on the nasdaq', 'most of the', 'of the fed', 'of the most', 'of the economy', 'of the yea
r', 'of the new', 'of the world', 'of million or', 'by the federal', 'more than a', 'more than of'
, 'more than billion', 'than a year', 'a year earlier', 'a year ago', 'a series of', 'a seasonally
adjusted', 'a annual rate', 'a decline in', 'a lot of', 'a number of', 'in the latest', 'in the fo
urth', 'in the first', 'in the dollar', 'in the stock', 'in the economy', 'in the second', 'in the
past', 'in the market', 'in interest rates', 'in new york', 'in recent years', 'in more than', 'th
e wall street', 'the bush administration', 'the federal reserve', 'the new york', 'the end of', 't
he white house', 'the second quarter', 'the dollar was', 'the first time', 'the first quarter', 't
he economic recovery', 'the economy the', 'the economy has', 'the economy and', 'the economy is',
'the dow jones', 'the s p', 'the nasdaq composite', 'the nasdaq stock', 'the bond market', 'the st
ock market', 'the standard poor', 'the central bank', 'the fed has', 'the fourth quarter', 'the pa
st three', 'the past two', 'the labor department', 'the commerce department', 'the university of',
'the number of', 'the unemployment rate', 'the market is', 'the trade deficit', 'the financial cri
sis', 'the financial markets', 'decline in the', 'interest rates and', 'or to and', 'or to the', '
or cents a', 'to and the', 'to keep the', 'to be the', 'to billion in', 'according to a',
'according to the', 'and interest rates', 'are expected to', 'are likely to', 'said in a', 'at the
end', 'at a annual', 'wall street journal', 'is one of', 'is expected to', 'is that the', 'is like
ly to', 'for the first', 'for the year', 'for the second', 'for the week', 'federal reserve board'
, 'it was a', 'but it is', 'that the economy', 'that the market', 'that the federal', 'there is no
', 'one of the', 'year according to', 'as the economy', 'as a result', 'as much as', 'this past we
ek', 'some of the', 'million or cents', 'end of the', 'part of a', 'over the past', 'over the
next', 'billion in the', 'number of companies', 'cents or to', 'cents a share', 'down from last',
'dow jones industrial', 'jones industrial average', 'points or to', 'nasdaq composite index',
'nasdaq stock market', 'annual rate of', 'lower interest rates', 'vice president of', 'standard po
or index', 'compared with a', 'expected to be', 'cut in the', 'unemployment rate for', 'food and e
nergy', 'likely to be', 'chief economist at', 'chairman of the', 'chairman ben bernanke',
'familiar with the', 'new york stock exchange', 'of million or cents', 'more than a year', 'in the
fourth quarter', 'in the first quarter', 'in the stock market', 'in new york the', 'the wall stree
t journal', 'the federal reserve board', 'the new york stock', 'the end of the', 'the dow jones in
dustrial', 'the nasdaq composite index', 'the nasdaq stock market', 'the standard poor index', 'to
billion in the', 'at the end of', 'at a annual rate', 'is one of the', 'is expected to be', 'for t
he first time', 'that the economy is', 'that the federal reserve', 'million or cents a', 'dow jone
s industrial average', 'of million or cents a', 'the new york stock exchange', 'the dow jones
industrial average'])
```

Quality score

In [12]:

```
phrases = []
for w in f.keys():
    if len(w.split(u' ')) > 1:
        phrases.append(w)
```

In [13]:

```
print('total 2 words and above phrases:', len(phrases))
print('sample:', phrases[:10])
```

```
total 2 words and above phrases: 972
sample: ['new york', 'york the', 'york a', 'york stock', 'on friday', 'on its', 'on the', 'on a',
'on tuesday', 'on their']
```

Concordance

In [14]:

```
PMI = {}
PLK = {}
```

In [15]:

```
def prob_u(phrase, f):
    words = phrase.split(u' ')
    denom = 0
    for j in words:
        denom += f[word]
    return f[phrase]*1. / denom
```

In [16]:

```
def cal_PMI_PLK(phrase, f, pmi, plk):
    words = phrase.split(u' ')
    prob_whole = prob_u(phrase, f)
    if len(words) > 2:
        min_mutual_info = float('inf')
        best_u_left = None
        best_u_right = None
        for i in range(1, len(words)):
            u_left = u' '.join(words[:i])
            u_right = u' '.join(words[i:])
            info = math.log(prob_whole * 1. / (prob_u(u_left, f) * prob_u(u_right, f)))
            if info < min_mutual_info:
                min_mutual_info = info
                best_u_left = u_left
                best_u_right = u_right
    else:
        best_u_left = words[0]
        best_u_right = words[1]

    pmi[phrase] = math.log(prob_whole * 1. / (prob_u(best_u_left, f) * prob_u(best_u_right, f)))
    plk[phrase] = prob_whole * math.log(prob_whole * 1. / (prob_u(best_u_left, f) *
prob_u(best_u_right, f)))
```

In [17]:

```
for p in phrases:
    cal_PMI_PLK(p, f, PMI, PLK)
```

In [18]:

```
print(len(PMI.keys()))
print(list(PMI.keys())[:10])
print(PMI[list(PMI.keys())[0]])
print(PLK[list(PLK.keys())[0]])
```

```
972
['new york', 'york the', 'york a', 'york stock', 'on friday', 'on its', 'on the', 'on a', 'on tues
day', 'on their']
-4.165113633110308
-31.65486361163834
```

Informativeness

In [19]:

```
IDF_phrase = {}
IDF_word = {}
```

In [20]:

```
def cal_word_IDF(docs, idf_word):
    total_length = len(docs)
    for w in all_words:
        count = 0
        for d in docs:
            if w in d:
                count += 1
        if count > 0:
            idf_word[w] = math.log(total_length * 1. / count)
        else:
            idf_word[w] = 0
```

In [21]:

```
cal_word_IDF(all_articles, IDF_word)
```

In [22]:

```
print(list(IDF_word.keys())[:5])
index = list(IDF_word.keys())[:5]
for i in index:
    print(IDF_word[i])
```

```
['new', 'york', 'yields', 'on', 'most']
0.7339691750802005
1.3470736479666094
2.900422093749666
0.22314355131420976
1.40649706843741
```

In [23]:

```
def cal_phrase_IDF(phrases, word_idf, phrase_idf):
    for p in phrases:
        idf = 0
        words = p.split(u' ')
        for w in words:
            idf += word_idf[w]
        phrase_idf[p] = idf * 1. / len(words)
```

In [24]:

```
cal_phrase_IDF(phrases, IDF_word, IDF_phrase)
```

In [25]:

```
print(list(IDF_phrase.keys())[:10])
for p in list(IDF_phrase.keys())[:10]:
    print(IDF_phrase[p])
```

```
['new york', 'york the', 'york a', 'york stock', 'on friday', 'on its', 'on the', 'on a', 'on tues
day', 'on their']
1.040521411523405
0.6760430948950769
0.6965587932340082
1.3976217064022938
1.0601317681000455
0.48379201313085274
0.11407804656887698
0.134593744907808
```

0.13459374490786853
1.4412017941234938
0.5635058815949038

Training a Classifier

In [26]:

```
data = {}
for p in phrases:
    data[p] = []
    data[p].append(float(PMI[p]))
    data[p].append(float(PLK[p]))
    data[p].append(float(IDF_phrase[p]))
```

In [27]:

```
import numpy as np
import sklearn
```

In [28]:

```
print(data.keys())
```

dict_keys(['new york', 'york the', 'york a', 'york stock', 'on friday', 'on its', 'on the', 'on a', 'on tuesday', 'on their', 'on thursday', 'most of', 'of a', 'of the', 'of this', 'of companies', 'of their', 'of some', 'of america', 'of canada', 'of all', 'of million', 'of it', 'of about', 'of its', 'of economic', 'of business', 'of dollars', 'of these', 'of in', 'of government', 'of and', 'of an', 'of people', 'of new', 'of money', 'of bonds', 'of them', 'by the', 'by a', 'banks to', 'more than', 'than a', 'than of', 'than the', 'than in', 'than billion', 'a larger', 'a year', 'a share', 'a drop', 'a basis', 'a series', 'a big', 'a gain', 'a rate', 'a little', 'a increase', 'a seasonally', 'a annual', 'a strong', 'a decline', 'a level', 'a lot', 'a billion', 'a sharp', 'a sign', 'a rise', 'a record', 'a month', 'a result', 'a more', 'a number', 'a large', 'a few', 'a bit', 'a good', 'a new', 'a high', 'a professor', 'a recession', 'a million', 'a major', 'percentage point', 'point of', 'in the', 'in interest', 'in new', 'in one', 'in an', 'in part', 'in some', 'in a', 'in its', 'in what', 'in late', 'in their', 'in any', 'in this', 'in december', 'in may', 'in recent', 'in november', 'in october', 'in when', 'in july', 'in june', 'in august', 'in january', 'in and', 'in annual', 'in april', 'in washington', 'in more', 'in which', 'in fact', 'in many', 'in september', 'the latest', 'the overall', 'the average', 'the week', 'the previous', 'the wall', 'the day', 'the bush', 'the federal', 'the state', 'the new', 'the program', 'the end', 'the white', 'the securities', 'the senate', 'the house', 'the bank', 'the bill', 'the risk', 'the results', 'the second', 'the study', 'the dollar', 'the euro', 'the yen', 'the issue', 'the first', 'the economic', 'the employment', 'the economy', 'the dow', 'the current', 'the next', 'the s', 'the nasdaq', 'the recent', 'the canadian', 'the last', 'the quarter', 'the period', 'the reagan', 'the and', 'the american', 'the nation', 'the terms', 'the deal', 'the late', 'the public', 'the rise', 'the firm', 'the two', 'the decline', 'the bond', 'the japanese', 'the year', 'the stock', 'the standard', 'the fall', 'the talks', 'the company', 'the is', 'the markets', 'the only', 'the country', 'the central', 'the law', 'the fed', 'the rate', 'the session', 'the time', 'the outlook', 'the group', 'the fourth', 'the third', 'the recession', 'the pace', 'the same', 'the past', 'the capital', 'the city', 'the largest', 'the most', 'the treasury', 'the labor', 'the board', 'the price', 'the commerce', 'the gains', 'the university', 'the government', 'the world', 'the jobless', 'the housing', 'the crash', 'the biggest', 'the number', 'the recovery', 'the job', 'the unemployment', 'the numbers', 'the department', 'the amount', 'the market', 'the way', 'the highest', 'the big', 'the trade', 'the currency', 'the report', 'the budget', 'the deficit', 'the mark', 'the start', 'the president', 'the potential', 'the other', 'the national', 'the financial', 'the consensus', 'the money', 'the final', 'the index', 'the region', 'the global', 'the mortgage', 'the imf', 'the news', 'the cftc', 'the project', 'the art', 'decline in', 'interest rates', 'interest rate', 'rates and', 'rates in', 'rates that', 'rates to', 'or to', 'or at', 'or the', 'or cents', 'or more', 'or that', 'or a', 'consumer credit', 'consumer prices', 'yield on', 'fell to', 'fell points', 'to from', 'to an', 'to and', 'to the', 'to help', 'to its', 'to have', 'to allow', 'to keep', 'to give', 'to reduce', 'to make', 'to raise', 'to million', 'to a', 'to be', 'to do', 'to buy', 'to billion', 'to think', 'to take', 'to start', 'to continue', 'to go', 'to get', 'to say', 'to pay', 'to their', 'to find', 'to investors', 'to push', 'to in', 'to see', 'to increase', 'to more', 'to people', 'from in', 'from the', 'from a', 'from last', 'from an', 'yesterday as', 'according to', 'an index', 'an average', 'an annual', 'an ounce', 'markets and', 'before the', 'two years', 'and the', 'and new', 'and for', 'and some', 'and other', 'and a', 'and investors', 'and interest', 'and with', 'and has', 'and energy', 'and they', 'and that', 'and more', 'and many', 'and inflation', 'and in', 'and so', 'and analysts', 'and it', 'are the', 'are expected', 'are likely', 'less than', 'declines in', 'which has', 'which is', 'said it', 'said the', 'said that', 'said a', 'said in', 'treasury secretary', 'at the', 'at least', 'at a', 'at its', 'at to', 'at billion', 'wall street', 'street journal', 'look at', 'is to', 'is the', 'is one', 'is a', 'is it', 'is expected', 'is that', 'is at', 'is still', 'is likely', 'is about', 'is up', 'is no

', 'is not', 'business investment', 'sign that', 'sign of', 'up for', 'up from', 'up their', 'up i
n', 'up to', 'up the', 'for the', 'for a', 'for their', 'for each', 'for more', 'friday the', 'wit
h the', 'with an', 'with a', 'with its', 'with billion', 'out of', 'its own', 'into a', 'into the',
'bush administration', 'federal reserve', 'will make', 'will be', 'it is', 'it will', 'it would',
'it was', 'it could', 'it also', 'it had', 'it a', 'it has', 'insurance companies', 'program to',
'health care', 'was at', 'was a', 'was the', 'was up', 'could be', 'but the', 'but there', 'but it
', 'but they', 'but not', 'but that', 'that the', 'that is', 'that would', 'that have', 'that
was', 'that are', 'that has', 'that we', 'that it', 'that in', 'that they', 'that many', 'that a',
'that he', 'that its', 'that some', 'that means', 'become a', 'there is', 'there a', 'there are',
'there were', 'would be', 'would have', 'be the', 'be a', 'plan to', 'three months', 'three years'
, 'four years', 'under the', 'have to', 'have also', 'have been', 'have the', 'have a', 'one of',
'year according', 'year the', 'year to', 'year earlier', 'year but', 'year and', 'year ago', 'year
end', 'washington the', 'as the', 'as investors', 'as of', 'as one', 'as a', 'as they', 'as much',
'as many', 'as well', 'can be', 'such as', 'such a', 'director of', 'policy makers', 'since april'
, 'since the', 'many economists', 'many investors', 'this is', 'this year', 'this was', 'this week'
, 'this past', 'been a', 'their bills', 'their own', 'some of', 'million or', 'million in', 'end o
f', 'next month', 'next year', 'if you', 'if they', 'if the', 'if it', 'if we', 'white house', 'ag
ainst the', 'effort to', 'agreed to', 'drop in', 'efforts to', 'when the', 'when he', 'when it',
'when a', 'bank of', 'last week', 'last year', 'last month', 'second quarter', 'attempt to',
'after the', 'after a', 'part of', 'because of', 'because it', 'because the', 'over the', 'how the
', 'how much', 'reserve board', 'ability to', 'keep the', 'give the', 'time to', 'time in',
'support for', 'who are', 'should be', 'should have', 'known as', 'billion in', 'labor
department', 'labor market', 'pressure on', 'reduce the', 'risk of', 'sales and', 'following the',
'starting to', 'all the', 'all of', 'all that', 'companies that', 'companies to', 'companies and',
'however the', 'they have', 'they were', 'they had', 'they do', 'they are', 'number of', 'about
the', 'about a', 'about million', 'about to', 'about of', 'signs of', 'had been', 'may be', 'may h
ave', 'even as', 'even if', 'even more', 'dollar was', 'dollar and', 'traders said', 'employment r
eport', 'market that', 'market the', 'market and', 'market is', 'issue of', 'late wednesday',
'cents or', 'cents a', 'down from', 'while the', 'head of', 'during the', 'trade deficit', 'first
time', 'first quarter', 'stock prices', 'stock market', 'stock exchange', 'economic recovery', 'ec
onomic growth', 'economic crisis', 'ahead of', 'stocks rose', 'declined to', 'investors are', 'gro
wth in', 'continued to', 'concerns about', 'economy the', 'economy has', 'economy and', 'economy i
s', 'dow jones', 'jones industrial', 'industrial average', 'points or', 'points to', 'close to', '
outlook for', 'quarter and', 'quarter of', 's p', 'index fell', 'index of', 'index rose', 'nasdaq
composite', 'nasdaq stock', 'composite index', 'royal bank', 'increased to', 'them to', 'recent
years', 'better than', 'has been', 'has to', 'has become', 'has a', 'credit markets', 'capital mar
kets', 'capital gains', 'recession in', 'he said', 'he says', 'share of', 'rose to', 'rose
points', 'rose in', 'people are', 'think of', 'inflation is', 'inflation and', 'back in', 'social
security', 'only a', 'prices of', 'prices and', 'prices on', 'prices are', 'terms of', 'around
the', 'third quarter', 'came in', 'through the', 'rise in', 'price index', 'price of', 'price incr
eases', 'financial markets', 'financial crisis', 'annual rate', 'valued at', 'analysts said', 'ser
ies of', 'data the', 'might be', 'traded at', 'although the', 'bond prices', 'bond market', 'lower
interest', 'higher interest', 'closed at', 'we had', 'we have', 'we are', 'continue to', 'between
the', 'now is', 'vice president', 'president of', 'surge in', 'standard poor', 'poor index', 'mci
worldcom', 'whether the', 'value of', 'compared with', 'compared to', 'expected to', 'like it', 'l
ike the', 'demand for', 'cut in', 'japan and', 'increase in', 'increase the', 'increase of',
'years the', 'years ago', 'years old', 'economist at', 'going to', 'believe the', 'bonds and',
'rate of', 'rate the', 'rate in', 'rate for', 'rate cut', 'unemployment rate', 'food and', 'rather
than', 'show that', 'instead of', 'central bank', 'central bankers', 'central banks', 'days of', '
used to', 'monetary policy', 'fed has', 'fed chairman', 'fed officials', 'funds and', 'pace of', '
much as', 'much of', 'she says', 'opportunity to', 'investment in', 'likely to', 'fourth quarter',
'seasonally adjusted', 'below the', 'enough to', 'half of', 'meanwhile the', 'goods and', 'months
of', 'months ago', 'chief economist', 'chairman of', 'chairman ben', 'plans to', 'energy prices',
'lot of', 'managing director', 'past week', 'past three', 'past two', 'try to', 'oil prices', 'led
by', 'due to', 'budget deficit', 'amount of', 'department said', 'producer price', 'weakness in',
'gains have', 'across the', 'warburg pincus', 'university of', 'morgan stanley', 'commerce
department', 'confidence in', 'trying to', 'mutual funds', 'few weeks', 'purchasing managers', 're
tail sales', 'shot up', 'continues to', 'mortgage rates', 'ago the', 'raising the', 'kind of',
'concern that', 'surveyed by', 'imports of', 'professor of', 'professor at', 'want to', 'ben
bernanke', 'familiar with', 'natural gas', 'mechanism design', 'freddie mac', 'new york the', 'new
york a', 'new york stock', 'york stock exchange', 'on the nasdaq', 'most of the', 'of the fed', 'o
f the most', 'of the economy', 'of the year', 'of the new', 'of the world', 'of million or', 'by t
he federal', 'more than a', 'more than of', 'more than billion', 'than a year', 'a year earlier',
'a year ago', 'a series of', 'a seasonally adjusted', 'a annual rate', 'a decline in', 'a lot of',
'a number of', 'in the latest', 'in the fourth', 'in the first', 'in the dollar', 'in the stock',
'in the economy', 'in the second', 'in the past', 'in the market', 'in interest rates', 'in new yo
rk', 'in recent years', 'in more than', 'the wall street', 'the bush administration', 'the federal
reserve', 'the new york', 'the end of', 'the white house', 'the second quarter', 'the dollar was',
'the first time', 'the first quarter', 'the economic recovery', 'the economy the', 'the economy ha
s', 'the economy and', 'the economy is', 'the dow jones', 'the s p', 'the nasdaq composite', 'the
nasdaq stock', 'the bond market', 'the stock market', 'the standard poor', 'the central bank', 'th
e fed has', 'the fourth quarter', 'the past three', 'the past two', 'the labor department', 'the c
ommerce department', 'the university of', 'the number of', 'the unemployment rate', 'the market is
', 'the trade deficit', 'the financial crisis', 'the financial markets', 'decline in the',
'interest rates and', 'or to and', 'or to the', 'or cents a', 'to and the', 'to keep the', 'to be
the', 'to billion in', 'according to a', 'according to the', 'and interest rates', 'are expected t
o', 'are likely to', 'said in a', 'at the end', 'at a annual', 'wall street journal', 'is one of',

```
'is expected to', 'is that the', 'is likely to', 'for the first', 'for the year', 'for the second'
, 'for the week', 'federal reserve board', 'it was a', 'but it is', 'that the economy', 'that the
market', 'that the federal', 'there is no', 'one of the', 'year according to', 'as the economy', '
as a result', 'as much as', 'this past week', 'some of the', 'million or cents', 'end of the', 'pa
rt of a', 'over the past', 'over the next', 'billion in the', 'number of companies', 'cents or to'
, 'cents a share', 'down from last', 'dow jones industrial', 'jones industrial average', 'points o
r to', 'nasdaq composite index', 'nasdaq stock market', 'annual rate of', 'lower interest rates',
'vice president of', 'standard poor index', 'compared with a', 'expected to be', 'cut in the', 'un
employment rate for', 'food and energy', 'likely to be', 'chief economist at', 'chairman of the',
'chairman ben bernanke', 'familiar with the', 'new york stock exchange', 'of million or cents', 'm
ore than a year', 'in the fourth quarter', 'in the first quarter', 'in the stock market', 'in new
york the', 'the wall street journal', 'the federal reserve board', 'the new york stock', 'the end
of the', 'the dow jones industrial', 'the nasdaq composite index', 'the nasdaq stock market', 'the
standard poor index', 'to billion in the', 'at the end of', 'at a annual rate', 'is one of the', '
is expected to be', 'for the first time', 'that the economy is', 'that the federal reserve', 'mill
ion or cents a', 'dow jones industrial average', 'of million or cents a', 'the new york stock
exchange', 'the dow jones industrial average'])
```

In [29]:

```
qual_phrase = [u'dow jones industrial average',
               u'chairman ben bernanke',
               u'economic recovery',
               u'business investment',
               u'nasdaq composite index',
               u'commerce department',
               u'interest rates',
              u'seasonally adjusted',
               u'labor department',
               u'central banks']
rand_phrase = [u'to billion',
               u'and new',
               u'but it is',
               u'starting to',
               u'if we',
               u'year and',
               u'said in',
               u'rise in',
               u'or to the',
               u'professor at']
```

In [30]:

```
train_data = []
for j in qual_phrase:
        train_data.append([1] + data[j])

for j in rand_phrase:
        train_data.append([0] + data[j])
train_data = np.array(train_data)
np.random.shuffle(train_data)
```

In [31]:

```
print(train_data.shape)
X = train_data[:, 1:]
y = train_data[:, 0]
print(X.shape)
print(y.shape)
print(train_data[:10])
```

```
(20, 4)
(20, 3)
(20,)
[[ 0.         -8.46005125 -5.64003417  0.26194802]
 [ 0.         -6.07335171 -3.64401102  1.40795692]
 [ 0.         -7.07898996 -9.20268695  0.97420664]
 [ 1.         -1.38629436 -0.97040605  3.22675   ]
 [ 0.         -8.7514998  -8.7514998   0.43661562]
 [ 1.         -3.47506723 -2.78005378  2.0993299 ]
 [ 0.         -6.39612934 -3.19806467  1.85199655]
 [ 1.         -4.81741006 -4.33566906  1.68886251]
 [ 1.         -2.98365969 -2.98365969  2.57927771]
 [ 0.         -6.0310698  -2.0103566   0.4424538 ]]
```

In [32]:

```
test_data = []
test_phrases = []
for p in phrases:
    if p not in qual_phrase and p not in rand_phrase:
        test_data.append(data[p])
        test_phrases.append(p)
test_data = np.array(test_data)
print(test_data)
```

```
[[ -4.16511363 -31.65486361   1.04052141]
 [ -9.01286457  -9.91415102   0.67604309]
 [ -8.77854772  -4.38927386   0.69655879]
 ...
 [ -5.67246379  -1.13449276   0.98308594]
 [ -6.61002287  -1.58640549   1.16736204]
 [ -6.64177157  -8.23579675   1.26848199]]
```

In [33]:

```
from sklearn.ensemble import RandomForestClassifier
clf = RandomForestClassifier(max_depth=3, n_estimators=10,random_state=0)
clf.fit(X, y)
```

Out[33]:

```
RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                       criterion='gini', max_depth=3, max_features='auto',
                       max_leaf_nodes=None, max_samples=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=10,
                       n_jobs=None, oob_score=False, random_state=0, verbose=0,
                       warm_start=False)
```

In [34]:

```
predictions = []
pred_data = clf.predict_proba(test_data)
for k in range(len(test_data)):
    predictions.append([test_phrases[k], pred_data[k][1]])
print(predictions[:10])
```

```
[['new york', 0.6], ['york the', 0.0], ['york a', 0.02222222222222222], ['york stock',
0.04722222222222222], ['on friday', 0.2], ['on its', 0.02222222222222222], ['on the', 0.0], ['on a
', 0.0], ['on tuesday', 0.14722222222222223], ['on their', 0.0]]
```

In [35]:

```
best_predictions = []
worst_predictions = []
for p in predictions:
    if p[1] >= 0.8:
        best_predictions.append(p)
    elif p[1] < 0.2:
        worst_predictions.append(p)
print('best predictions:', best_predictions[:10])
print('worst predictions:', worst_predictions[:10])
```

```
best predictions: [['percentage point', 0.975], ['consumer credit', 0.975], ['an ounce', 0.975], [
'two years', 0.825], ['treasury secretary', 0.975], ['wall street', 0.8222222222222222], ['street
journal', 0.975], ['bush administration', 0.975], ['insurance companies', 0.875], ['health care',
0.975]]
worst predictions: [['york the', 0.0], ['york a', 0.02222222222222222], ['york stock',
0.04722222222222222], ['on its', 0.02222222222222222], ['on the', 0.0], ['on a', 0.0], ['on
tuesday', 0.14722222222222223], ['on their', 0.0], ['most of', 0.0], ['of a', 0.0]]
```