nlp-ex3

February 8, 2024

[1]:
```
import pandas as pd
import numpy as np
```

[2]:
```
sen= """Millions go missing at China bank Two senior officials at one of␣
 ↪China's top commercial banks have reportedly disappeared after funds
worth up to $120m (£64m) went missing.
The pair both worked at Bank of China in the northern city of Harbin, the South␣
 ↪China Morning Post
reported. The latest scandal at Bank of China will do nothing to reassure␣
 ↪foreign investors that China's
big four banks are ready for international listings. Government policy sees the␣
 ↪bank listings as vital
economic reforms. Bank of China is one of two frontrunners in the race to list␣
 ↪overseas. The other is
China Construction Bank. Both are expected to list abroad during 2005.
They shared a $45bn state bailout in 2003, to help clean up their balance␣
 ↪sheets in preparation for a
foreign stock market debut.
However, a report in the China-published Economic Observer said on Monday that␣
 ↪the two banks may
have scrapped plans to list in New York because of the cost of meeting␣
 ↪regulatory requirements
imposed since the Enron scandal. Bank of China is the country's biggest foreign␣
 ↪exchange dealer, while
China Construction Bank is the largest deposit holder. China's banking sector␣
 ↪is burdened with at least
$190bn of bad debt according to official data, though most observers believe␣
 ↪the true figure is far
higher. Officially, one in five loans is not being repaid. Attempts to␣
 ↪strengthen internal controls and
tighten lending policies have uncovered a succession of scandals involving␣
 ↪embezzlement by bank
officials and loans-for-favours. The most high-profile case involved the␣
 ↪ex-president of Bank of China,
Wang Xuebing, jailed for 12 years in 2003. Although, he committed the offences␣
 ↪whilst running Bank
```

```
of China in New York, Mr.Wang was head of China Construction Bank when the␣
 ↪scandal broke. Earlier
this month, a China Construction Bank branch manager was jailed for life in a␣
 ↪separate case.
China's banks used to act as cash offices for state enterprises and did not␣
 ↪require checks on credit
worthiness. The introduction of market reforms has been accompanied by attempts␣
 ↪to modernize the
banking sector, but links between banks and local government remain strong.␣
 ↪Last year, China's
premier, Wen Jiabao, targeted bank lending practices in a series of speeches,␣
 ↪and regulators ordered
all big loans to be scrutinized, in an attempt to cool down irresponsible␣
 ↪lending. China's leaders see
reforming the top four banks as vital to distribute capital to profitable␣
 ↪companies and protect the health
of China's economic boom. But two problems persist. First, inefficient state␣
 ↪enterprises continue to
receive protection from bankruptcy because they employ large numbers of people.␣
 ↪Second, many
questionable loans come not from the big four, but from smaller banks. Another␣
 ↪high-profile financial
firm, China Life, is facing shareholder lawsuits and a probe by the US␣
 ↪Securities and Exchange
Commission following its 2004 New York listing over its failure to disclose␣
 ↪accounting irregularities
at its parent company."""
```

[3]:
```
import nltk
```

[4]:
```
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import SnowballStemmer, WordNetLemmatizer
from nltk.tag import pos_tag
from nltk.chunk import ne_chunk
import string
```

[5]:
```
nltk.download('stopwords')
nltk.download('punkt')
nltk.download('wordnet')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\User\AppData\Roaming\nltk_data…
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to
```

```
[nltk_data]     C:\Users\User\AppData\Roaming\nltk_data…
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\User\AppData\Roaming\nltk_data…
[nltk_data]   Package wordnet is already up-to-date!
```

[5]:
```
True
```

[6]:
```
sentences= sen.split(sep='.')
df= pd.DataFrame({'text': sentences})
df.head()
```

[6]:
```
                                                text
0  Millions go missing at China bank Two senior o…
1  \nThe pair both worked at Bank of China in the…
2   The latest scandal at Bank of China will do n…
3   Government policy sees the bank listings as v…
4   Bank of China is one of two frontrunners in t…
```

[7]:
```
def concat_text(tokens):
    return " ".join([token for token in tokens])
```

[8]:
```
df['text'] = df['text'].apply(word_tokenize)
df['text'] = df['text'].apply(concat_text)
df.head()
```

[8]:
```
                                                text
0  Millions go missing at China bank Two senior o…
1  The pair both worked at Bank of China in the n…
2  The latest scandal at Bank of China will do no…
3  Government policy sees the bank listings as vi…
4  Bank of China is one of two frontrunners in th…
```

[9]:
```
def remove_punc(text):
    removed_text = ""
    for char in str(text.lower()):
        if char not in string.punctuation:
            removed_text+=char
    return removed_text
```

[10]:
```
df['Punc'] = df['text'].apply(remove_punc)
df.head()
```

[10]:
```
                                                text  \
0  Millions go missing at China bank Two senior o…
1  The pair both worked at Bank of China in the n…
2  The latest scandal at Bank of China will do no…
```

```
3  Government policy sees the bank listings as vi…
4  Bank of China is one of two frontrunners in th…

                                                Punc
0  millions go missing at china bank two senior o…
1  the pair both worked at bank of china in the n…
2  the latest scandal at bank of china will do no…
3  government policy sees the bank listings as vi…
4  bank of china is one of two frontrunners in th…
```

[11]:
```
df= df.iloc[:25]
```

[12]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25 entries, 0 to 24
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   text    25 non-null     object
 1   Punc    25 non-null     object
dtypes: object(2)
memory usage: 532.0+ bytes
```

[13]:
```
df['tokenized'] = df['Punc'].apply(word_tokenize)
df.head()
```

[13]:
```
                                                text  \
0  Millions go missing at China bank Two senior o…
1  The pair both worked at Bank of China in the n…
2  The latest scandal at Bank of China will do no…
3  Government policy sees the bank listings as vi…
4  Bank of China is one of two frontrunners in th…

                                                Punc  \
0  millions go missing at china bank two senior o…
1  the pair both worked at bank of china in the n…
2  the latest scandal at bank of china will do no…
3  government policy sees the bank listings as vi…
4  bank of china is one of two frontrunners in th…

                                           tokenized
0  [millions, go, missing, at, china, bank, two, …
1  [the, pair, both, worked, at, bank, of, china,…
2  [the, latest, scandal, at, bank, of, china, wi…
3  [government, policy, sees, the, bank, listings…
4  [bank, of, china, is, one, of, two, frontrunne…
```

[14]:
```
def rem_stop(tokens):
    stop_words = set(stopwords.words('english'))
    filtered_tokens = [token for token in tokens if token.lower() not in␣
 ↪stop_words]
    return filtered_tokens
```

[15]:
```
df['stopwords']= df['tokenized'].apply(rem_stop)
df.head()
```

[15]:

```
                                                text  \
0  Millions go missing at China bank Two senior o…
1  The pair both worked at Bank of China in the n…
2  The latest scandal at Bank of China will do no…
3  Government policy sees the bank listings as vi…
4  Bank of China is one of two frontrunners in th…

                                                Punc  \
0  millions go missing at china bank two senior o…
1  the pair both worked at bank of china in the n…
2  the latest scandal at bank of china will do no…
3  government policy sees the bank listings as vi…
4  bank of china is one of two frontrunners in th…

                                           tokenized  \
0  [millions, go, missing, at, china, bank, two, …
1  [the, pair, both, worked, at, bank, of, china,…
2  [the, latest, scandal, at, bank, of, china, wi…
3  [government, policy, sees, the, bank, listings…
4  [bank, of, china, is, one, of, two, frontrunne…

                                           stopwords
0  [millions, go, missing, china, bank, two, seni…
1  [pair, worked, bank, china, northern, city, ha…
2  [latest, scandal, bank, china, nothing, reassu…
3  [government, policy, sees, bank, listings, vit…
4  [bank, china, one, two, frontrunners, race, li…
```

[16]:
```
def lemma_tokens(tokens):
    lemmatizer = WordNetLemmatizer()
    tokens = [lemmatizer.lemmatize(token) for token in tokens]
    return tokens
```

[17]:
```
df['lemma'] = df['stopwords'].apply(lemma_tokens)
df.head()
```

[17]:

```
                                                text  \
0  Millions go missing at China bank Two senior o…
```

```
1  The pair both worked at Bank of China in the n…
2  The latest scandal at Bank of China will do no…
3  Government policy sees the bank listings as vi…
4  Bank of China is one of two frontrunners in th…

                                                Punc  \
0  millions go missing at china bank two senior o…
1  the pair both worked at bank of china in the n…
2  the latest scandal at bank of china will do no…
3  government policy sees the bank listings as vi…
4  bank of china is one of two frontrunners in th…

                                            tokenized  \
0  [millions, go, missing, at, china, bank, two, …
1  [the, pair, both, worked, at, bank, of, china,…
2  [the, latest, scandal, at, bank, of, china, wi…
3  [government, policy, sees, the, bank, listings…
4  [bank, of, china, is, one, of, two, frontrunne…

                                            stopwords  \
0  [millions, go, missing, china, bank, two, seni…
1  [pair, worked, bank, china, northern, city, ha…
2  [latest, scandal, bank, china, nothing, reassu…
3  [government, policy, sees, bank, listings, vit…
4  [bank, china, one, two, frontrunners, race, li…

                                               lemma
0  [million, go, missing, china, bank, two, senio…
1  [pair, worked, bank, china, northern, city, ha…
2  [latest, scandal, bank, china, nothing, reassu…
3  [government, policy, see, bank, listing, vital…
4  [bank, china, one, two, frontrunners, race, li…
```

[18]:
```
df['preprocessed_text'] = df['lemma'].apply(concat_text)
df.head()
```

[18]:
```
                                                text  \
0  Millions go missing at China bank Two senior o…
1  The pair both worked at Bank of China in the n…
2  The latest scandal at Bank of China will do no…
3  Government policy sees the bank listings as vi…
4  Bank of China is one of two frontrunners in th…

                                                Punc  \
0  millions go missing at china bank two senior o…
1  the pair both worked at bank of china in the n…
2  the latest scandal at bank of china will do no…
```

```
3  government policy sees the bank listings as vi…
4  bank of china is one of two frontrunners in th…

                                              tokenized  \
0  [millions, go, missing, at, china, bank, two, …
1  [the, pair, both, worked, at, bank, of, china,…
2  [the, latest, scandal, at, bank, of, china, wi…
3  [government, policy, sees, the, bank, listings…
4  [bank, of, china, is, one, of, two, frontrunne…

                                              stopwords  \
0  [millions, go, missing, china, bank, two, seni…
1  [pair, worked, bank, china, northern, city, ha…
2  [latest, scandal, bank, china, nothing, reassu…
3  [government, policy, sees, bank, listings, vit…
4  [bank, china, one, two, frontrunners, race, li…

                                                lemma  \
0  [million, go, missing, china, bank, two, senio…
1  [pair, worked, bank, china, northern, city, ha…
2  [latest, scandal, bank, china, nothing, reassu…
3  [government, policy, see, bank, listing, vital…
4  [bank, china, one, two, frontrunners, race, li…

                                    preprocessed_text
0  million go missing china bank two senior offic…
1  pair worked bank china northern city harbin so…
2  latest scandal bank china nothing reassure for…
3  government policy see bank listing vital econo…
4  bank china one two frontrunners race list over…
```

[19]:
```
df.head()
```

[19]:
```
                                                text  \
0  Millions go missing at China bank Two senior o…
1  The pair both worked at Bank of China in the n…
2  The latest scandal at Bank of China will do no…
3  Government policy sees the bank listings as vi…
4  Bank of China is one of two frontrunners in th…

                                                Punc  \
0  millions go missing at china bank two senior o…
1  the pair both worked at bank of china in the n…
2  the latest scandal at bank of china will do no…
3  government policy sees the bank listings as vi…
4  bank of china is one of two frontrunners in th…
```

```
                                                tokenized  \
0  [millions, go, missing, at, china, bank, two, …
1  [the, pair, both, worked, at, bank, of, china,…
2  [the, latest, scandal, at, bank, of, china, wi…
3  [government, policy, sees, the, bank, listings…
4  [bank, of, china, is, one, of, two, frontrunne…

                                                stopwords  \
0  [millions, go, missing, china, bank, two, seni…
1  [pair, worked, bank, china, northern, city, ha…
2  [latest, scandal, bank, china, nothing, reassu…
3  [government, policy, sees, bank, listings, vit…
4  [bank, china, one, two, frontrunners, race, li…

                                                  lemma  \
0  [million, go, missing, china, bank, two, senio…
1  [pair, worked, bank, china, northern, city, ha…
2  [latest, scandal, bank, china, nothing, reassu…
3  [government, policy, see, bank, listing, vital…
4  [bank, china, one, two, frontrunners, race, li…

                                   preprocessed_text
0  million go missing china bank two senior offic…
1  pair worked bank china northern city harbin so…
2  latest scandal bank china nothing reassure for…
3  government policy see bank listing vital econo…
4  bank china one two frontrunners race list over…
```

1 Summary for BOW

[20]:
```
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer()
count_matrix = cv.fit_transform(df['preprocessed_text'].values.tolist())
count_matrix
```

[20]:
```
<25x214 sparse matrix of type '<class 'numpy.int64'>'
        with 301 stored elements in Compressed Sparse Row format>
```

[21]:
```
from sklearn.metrics.pairwise import cosine_similarity
cosine_sim_bow = cosine_similarity(count_matrix, count_matrix)
```

[22]:
```
def get_summary(doc_index, similarity_matrix, documents, threshold=0.2):
    summary = " "
    similar_indices = (similarity_matrix[doc_index] > threshold).nonzero()[0]
    similar_documents = documents[similar_indices].tolist()
    for i in similar_documents:
```

```
        summary = summary+ i + ". "
    return summary
```

[23]:
```
document_index = 17
summary_bow = get_summary(document_index, cosine_sim_bow, df['text'])

print(summary_bow)
```

```
 Millions go missing at China bank Two senior officials at one of China 's top
commercial banks have reportedly disappeared after funds worth up to $ 120m (
£64m ) went missing. The pair both worked at Bank of China in the northern city
of Harbin , the South China Morning Post reported. The latest scandal at Bank of
China will do nothing to reassure foreign investors that China's big four banks
are ready for international listings. Bank of China is one of two frontrunners
in the race to list overseas. The other is China Construction Bank. Bank of
China is the country 's biggest foreign exchange dealer , while China
Construction Bank is the largest deposit holder. Wang was head of China
Construction Bank when the scandal broke. China 's banks used to act as cash
offices for state enterprises and did not require checks on credit worthiness.
```

2 Summary for TFIDF

[24]:
```
from sklearn.feature_extraction.text import TfidfVectorizer
tfidf= TfidfVectorizer()
tfidf_matrix = tfidf.fit_transform(df['preprocessed_text'].values.tolist())
tfidf_matrix
```

[24]:
```
<25x214 sparse matrix of type '<class 'numpy.float64'>'
        with 301 stored elements in Compressed Sparse Row format>
```

[25]:
```
cosine_sim_tfidf = cosine_similarity(tfidf_matrix, tfidf_matrix)
document_index = 17
summary_tfidf = get_summary(document_index, cosine_sim_tfidf, df['text'],␣
 ↪threshold=0.03)

print(summary_tfidf)
```

```
 Millions go missing at China bank Two senior officials at one of China 's top
commercial banks have reportedly disappeared after funds worth up to $ 120m (
£64m ) went missing. The pair both worked at Bank of China in the northern city
of Harbin , the South China Morning Post reported. The latest scandal at Bank of
China will do nothing to reassure foreign investors that China's big four banks
are ready for international listings. Bank of China is one of two frontrunners
in the race to list overseas. The other is China Construction Bank. They shared
a $ 45bn state bailout in 2003 , to help clean up their balance sheets in
preparation for a foreign stock market debut. Bank of China is the country 's
biggest foreign exchange dealer , while China Construction Bank is the largest
```

deposit holder. The most high-profile case involved the ex-president of Bank of China , Wang Xuebing , jailed for 12 years in 2003. Although , he committed the offences whilst running Bank of China in New York , Mr. Wang was head of China Construction Bank when the scandal broke. Earlier this month , a China Construction Bank branch manager was jailed for life in a separate case. China 's banks used to act as cash offices for state enterprises and did not require checks on credit worthiness. China 's leaders see reforming the top four banks as vital to distribute capital to profitable companies and protect the health of China 's economic boom. First , inefficient state enterprises continue to receive protection from bankruptcy because they employ large numbers of people.

3 Summary for CBOW

[26]:
```
from gensim.models.word2vec import Word2Vec

cbow = Word2Vec(df['preprocessed_text'], vector_size=150, window=5,␣
 ↪min_count=2, sg=0)
vocab = cbow.wv.index_to_key

def get_mean_vector(model, sentence):
    words = [word for word in sentence if word in vocab]
    if len(words) >= 1:
        return np.mean(model.wv[words], axis=0)
    return np.zeros((150,))

cbow_vector = [get_mean_vector(cbow, sentence) for sentence in␣
 ↪df['preprocessed_text']]
```

```
C:\Users\User\miniconda3\Lib\site-packages\paramiko\transport.py:219:
CryptographyDeprecationWarning: Blowfish has been deprecated
  "class": algorithms.Blowfish,
```

[27]:
```
cosine_sim_cbow = cosine_similarity(cbow_vector, cbow_vector)
document_index = 17
summary_cbow = get_summary(document_index, cosine_sim_cbow, df['text'],␣
 ↪threshold=0.99999)

print(summary_cbow)
```

```
 Bank of China is one of two frontrunners in the race to list overseas. They
shared a $ 45bn state bailout in 2003 , to help clean up their balance sheets in
preparation for a foreign stock market debut. However , a report in the China-
published Economic Observer said on Monday that the two banks may have scrapped
plans to list in New York because of the cost of meeting regulatory requirements
imposed since the Enron scandal. Bank of China is the country 's biggest foreign
exchange dealer , while China Construction Bank is the largest deposit holder.
China 's banking sector is burdened with at least $ 190bn of bad debt according
```

to official data , though most observers believe the true figure is far higher. Earlier this month , a China Construction Bank branch manager was jailed for life in a separate case. China 's banks used to act as cash offices for state enterprises and did not require checks on credit worthiness. Last year , China's premier , Wen Jiabao , targeted bank lending practices in a series of speeches , and regulators ordered all big loans to be scrutinized , in an attempt to cool down irresponsible lending. First , inefficient state enterprises continue to receive protection from bankruptcy because they employ large numbers of people.

4 Summary for Skipgram

[28]:
```
sg = Word2Vec(df['preprocessed_text'].values.tolist(), vector_size=150,␣
 ↪window=5, min_count=2, sg=1)
vocab = sg.wv.index_to_key

def get_mean_vector(model, sentence):
    words = [word for word in sentence if word in vocab]
    if len(words) >= 1:
        return np.mean(model.wv[words], axis=0)
    return np.zeros((150,))

sg_vector = []
for sentence in df['preprocessed_text'].values.tolist():
    sg_vector.append(get_mean_vector(sg, sentence))

sg_vector = np.array(sg_vector)
```

[29]:
```
cosine_sim_sg = cosine_similarity(sg_vector, sg_vector)
document_index = 17
summary_sg = get_summary(document_index, cosine_sim_sg, df['text'], threshold=0.
 ↪999987)

print(summary_sg)
```

Bank of China is one of two frontrunners in the race to list overseas. They shared a $ 45bn state bailout in 2003 , to help clean up their balance sheets in preparation for a foreign stock market debut. However , a report in the China-published Economic Observer said on Monday that the two banks may have scrapped plans to list in New York because of the cost of meeting regulatory requirements imposed since the Enron scandal. Bank of China is the country 's biggest foreign exchange dealer , while China Construction Bank is the largest deposit holder. China 's banking sector is burdened with at least $ 190bn of bad debt according to official data , though most observers believe the true figure is far higher. Earlier this month , a China Construction Bank branch manager was jailed for life in a separate case. China 's banks used to act as cash offices for state enterprises and did not require checks on credit worthiness. Last year , China's premier , Wen Jiabao , targeted bank lending practices in a series of speeches ,

and regulators ordered all big loans to be scrutinized , in an attempt to cool down irresponsible lending. First , inefficient state enterprises continue to receive protection from bankruptcy because they employ large numbers of people.

5 Summary for Word2Vec

[30]:
```
w2v_vector = (cbow_vector+sg_vector)/2
```

[31]:
```
cosine_sim_w2v = cosine_similarity(w2v_vector, w2v_vector)
document_index = 17
summary_w2v = get_summary(document_index, cosine_sim_w2v, df['text'],␣
 ↪threshold= 0.03)

print(summary_sg)
```

Bank of China is one of two frontrunners in the race to list overseas. They shared a $ 45bn state bailout in 2003 , to help clean up their balance sheets in preparation for a foreign stock market debut. However , a report in the China-published Economic Observer said on Monday that the two banks may have scrapped plans to list in New York because of the cost of meeting regulatory requirements imposed since the Enron scandal. Bank of China is the country 's biggest foreign exchange dealer , while China Construction Bank is the largest deposit holder. China 's banking sector is burdened with at least $ 190bn of bad debt according to official data , though most observers believe the true figure is far higher. Earlier this month , a China Construction Bank branch manager was jailed for life in a separate case. China 's banks used to act as cash offices for state enterprises and did not require checks on credit worthiness. Last year , China's premier , Wen Jiabao , targeted bank lending practices in a series of speeches , and regulators ordered all big loans to be scrutinized , in an attempt to cool down irresponsible lending. First , inefficient state enterprises continue to receive protection from bankruptcy because they employ large numbers of people.

6 Summary for GloVe

[32]:
```
from gensim.models import KeyedVectors
from gensim.scripts.glove2word2vec import glove2word2vec

glove_vectors_file = r'E:\NLP Lab\glove\glove.6B.100d.txt'

temp_word2vec_file = "temp_word2vec_file.txt"
glove2word2vec(glove_input_file=glove_vectors_file,␣
 ↪word2vec_output_file=temp_word2vec_file)

glove_model = KeyedVectors.load_word2vec_format(temp_word2vec_file,␣
 ↪binary=False)
```

```
def get_doc_vector(doc, model):
    vectors = []
    for word in doc:
        if word in model:
            vectors.append(model[word])
    if vectors:
        return np.mean(vectors, axis=0)
    else:
        return np.zeros(model.vector_size)

doc_vectors = np.array([get_doc_vector(doc, glove_model) for doc in␣
 ↪df['preprocessed_text']])
```

```
C:\Users\User\AppData\Local\Temp\ipykernel_17160\4169749988.py:7:
DeprecationWarning: Call to deprecated `glove2word2vec`
(KeyedVectors.load_word2vec_format(.., binary=False, no_header=True) loads GLoVE
text vectors.).
  glove2word2vec(glove_input_file=glove_vectors_file,
word2vec_output_file=temp_word2vec_file)
```

[33]:
```
cosine_sim_glove = cosine_similarity(doc_vectors, doc_vectors)
document_index = 17
summary_glove = get_summary(document_index, cosine_sim_glove, df['text'],␣
 ↪threshold=0.987)

print(summary_glove)
```

```
 The pair both worked at Bank of China in the northern city of Harbin , the
South China Morning Post reported. The latest scandal at Bank of China will do
nothing to reassure foreign investors that China's big four banks are ready for
international listings. Government policy sees the bank listings as vital
economic reforms. Bank of China is one of two frontrunners in the race to list
overseas. They shared a $ 45bn state bailout in 2003 , to help clean up their
balance sheets in preparation for a foreign stock market debut. However , a
report in the China-published Economic Observer said on Monday that the two
banks may have scrapped plans to list in New York because of the cost of meeting
regulatory requirements imposed since the Enron scandal. Bank of China is the
country 's biggest foreign exchange dealer , while China Construction Bank is
the largest deposit holder. China 's banking sector is burdened with at least $
190bn of bad debt according to official data , though most observers believe the
true figure is far higher. Attempts to strengthen internal controls and tighten
lending policies have uncovered a succession of scandals involving embezzlement
by bank officials and loans-for-favours. Earlier this month , a China
Construction Bank branch manager was jailed for life in a separate case. China
's banks used to act as cash offices for state enterprises and did not require
checks on credit worthiness. The introduction of market reforms has been
```

accompanied by attempts to modernize the banking sector , but links between banks and local government remain strong. Last year , China's premier , Wen Jiabao , targeted bank lending practices in a series of speeches , and regulators ordered all big loans to be scrutinized , in an attempt to cool down irresponsible lending. China 's leaders see reforming the top four banks as vital to distribute capital to profitable companies and protect the health of China 's economic boom. First , inefficient state enterprises continue to receive protection from bankruptcy because they employ large numbers of people. Another high-profile financial firm , China Life , is facing shareholder lawsuits and a probe by the US Securities and Exchange Commission following its 2004 New York listing over its failure to disclose accounting irregularities at its parent company.

7 Summary for FastText

[34]:
```
from gensim.models import FastText
ft=FastText(df['preprocessed_text'].values.
 ↪tolist(),vector_size=100,window=5,min_count=1,workers=4)
similar=ft.wv.most_similar('bank')

def get_mean_vector(model, sentence):
    words = [word for word in sentence if word in vocab]
    if len(words) >= 1:
        return np.mean(model.wv[words], axis=0)
    return np.zeros((100,))

ft_vector = []
for sentence in df['preprocessed_text'].values.tolist():
    ft_vector.append(get_mean_vector(ft, sentence))

ft_vector = np.array(ft_vector)
```

[35]:
```
cosine_sim_ft = cosine_similarity(ft_vector, ft_vector)
document_index = 17
summary_ft = get_summary(document_index, cosine_sim_ft, df['text'], threshold=0.
 ↪99999)

print(summary_ft)
```

The latest scandal at Bank of China will do nothing to reassure foreign investors that China's big four banks are ready for international listings. Bank of China is one of two frontrunners in the race to list overseas. They shared a $ 45bn state bailout in 2003 , to help clean up their balance sheets in preparation for a foreign stock market debut. However , a report in the China-published Economic Observer said on Monday that the two banks may have scrapped plans to list in New York because of the cost of meeting regulatory requirements imposed since the Enron scandal. Bank of China is the country 's biggest foreign

exchange dealer , while China Construction Bank is the largest deposit holder. China 's banking sector is burdened with at least $ 190bn of bad debt according to official data , though most observers believe the true figure is far higher. Earlier this month , a China Construction Bank branch manager was jailed for life in a separate case. China 's banks used to act as cash offices for state enterprises and did not require checks on credit worthiness. Last year , China's premier , Wen Jiabao , targeted bank lending practices in a series of speeches , and regulators ordered all big loans to be scrutinized , in an attempt to cool down irresponsible lending. China 's leaders see reforming the top four banks as vital to distribute capital to profitable companies and protect the health of China 's economic boom. First , inefficient state enterprises continue to receive protection from bankruptcy because they employ large numbers of people. Another high-profile financial firm , China Life , is facing shareholder lawsuits and a probe by the US Securities and Exchange Commission following its 2004 New York listing over its failure to disclose accounting irregularities at its parent company.

[ ]: