WebInfo-exp1 实验报告

组员：

廖佳怡 PB19151776

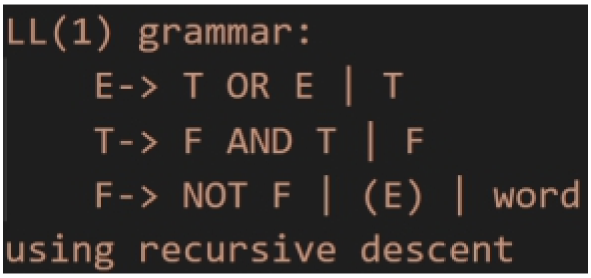
谢新格 PB19081644

基本要求

1. 将每篇文档进行了分词，词根化，去除停用词处理，能处理所有的30w篇文档。根据倒排索引算法建立倒排索引表并存储生成的文件。

- 在`Doc_Sparse.py`的`Doc_Sparse`类中完成预处理，使用了`nltk`库中的函数如`nltk.sent_tokenize`,`nltk.word_tokenize`进行分词，`nltk.stem.PorterStemmer.stem`进行词根化和`nltk.corpus.stopwords`进行停用词处理。使用`nltk.FreqDist`统计词频。
- 在`PL_construct.py`中建立倒排表索引表`PostingList`，生成三个文件`DocMap.json`,`PostingList.json`,`WordMap.json`存储在`output`文件夹中

2. Bool查询，使用了LL(1)文法：

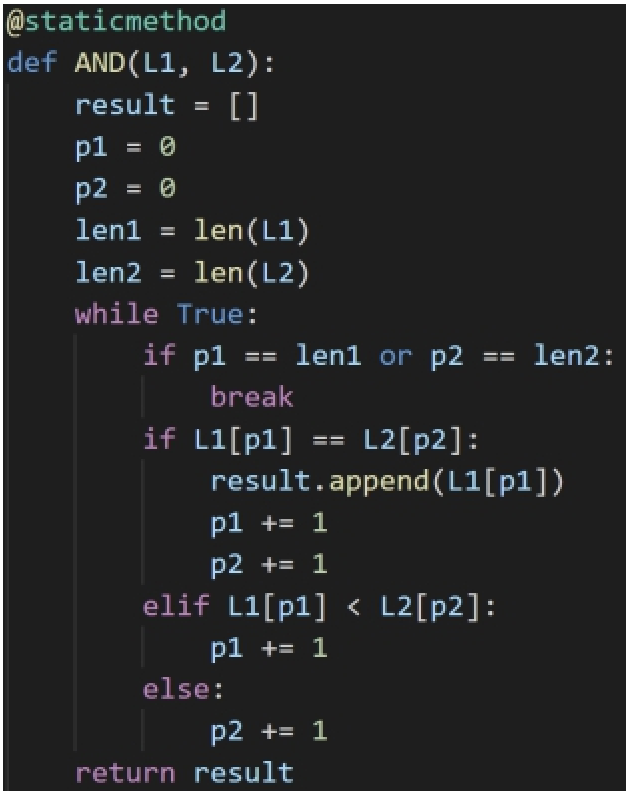


```
LL(1) grammar:
    E-> T OR E | T
    T-> F AND T | F
    F-> NOT F | (E) | word
using recursive descent
```

支持多括号，满足AND，OR，NOT的优先级，手动构建了递归下降的写法。

对于AND, OR, NOT 在倒排表上的算法，完成了3个算法，AND 和 OR 的算法时间复杂度是O（m+n）m, n分别是两个倒排表的长度。另外，由于文档进行了词根化，对于查询词也会先进行词根化

比方关于AND的算法，代码如下：



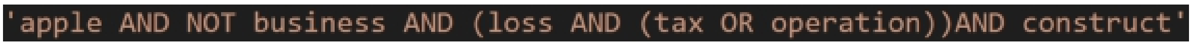
```python
@staticmethod
def AND(L1, L2):
    result = []
    p1 = 0
    p2 = 0
    len1 = len(L1)
    len2 = len(L2)
    while True:
        if p1 == len1 or p2 == len2:
            break
        if L1[p1] == L2[p2]:
            result.append(L1[p1])
            p1 += 1
            p2 += 1
        elif L1[p1] < L2[p2]:
            p1 += 1
        else:
            p2 += 1
    return result
```

对于词项的列表仅扫描一次，时间复杂度为O(m+n)

展示结果

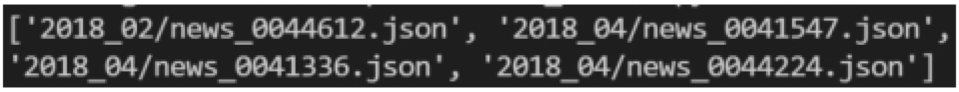
由于bool查询容易出现盛宴或饥荒的情况，所以只打印了文章位置与名字，而不直接打印文章内容。

查询：



```
'apple AND NOT business AND (loss AND (tax OR operation))AND construct'
```

结果：



```
['2018_02/news_0044612.json', '2018_04/news_0041547.json',
'2018_04/news_0041336.json', '2018_04/news_0044224.json']
```

打开其中一篇文档，其内容为：

```
Asian stock indexes closed higher on Wednesday after a session of choppy trade in
Japanese markets. More convincing gains were seen in Taiwan and Hong Kong markets
.
Japan's benchmark Nikkei 225 index closed up 0.21 percent, or 45.71 points, at
21,970.81 after a session of choppy trade. Manufacturing stocks finished the day
mostly higher, with Fanuc Manufacturing advancing 0.4 percent by the end of the
session.
Automakers and technology names were mixed, while bank stocks came under
pressure. Among blue chip names, Toyota gained 0.12 percent, Honda rose 1.64
percent and SoftBank Group slipped 0.55 percent on the day.
Symbol Name Price Change %Change NIKKEI --- HSI --- ASX 200 --- SHANGHAI ---
KOSPI ---
CNBC 100 --- Meanwhile, the Kospi edged up by 0.6 percent to close at 2,429.65.

Steelmakers traded lower after the South Korean government submitted a World
Trade Organization complaint over U.S. duties, although they finished the day off
their session lows. Posco closed lower by 0.14 percent and Hyundai Steel was down
0.57 percent on the
```

day.

Of note, the governor of the country's central bank on Wednesday said it was "prepared to respond" should the Federal Reserve raise interest rates more aggressively than markets were expecting, Reuters reported.

Down Under, the S&P/ASX 200 closed higher by 0.05 percent at 5,943.7 after hovering around the flat line for most of the session as investors focused on earnings releases. Gains in the consumer staples and discretionary sectors were offset by losses in the materials sector.

Major miners finished the session in the red: Shares of Fortescue Metals Group fell 4.66 percent after the company earlier reported that net profit after tax fell 44 percent to $681 million in the six months ending December. BHP closed down 4.76 percent after the mining major reported first-half earnings on Tuesday.

Australian conglomerate Wesfarmers reported on Tuesday that first-half net profit after tax came in at 212 million Australian dollars ($167 million) — an 86.6 percent decline compared to one year ago. Excluding a A$1.3 billion writedown, net profit stood at A$1.54 billion ($1.21 billion), a 2.7 percent decline from a year ago. Wesfarmers shares were up 3.02 percent by the end of the day.

Meanwhile, shares of a2 Milk surged 26.48 percent after the company announced record half-year profit. Investors also cheered a manufacturing and distribution agreement that the company signed with New Zealand-listed Fonterra.

"Global risk sentiments may remain rangebound post-holidays," OCBC Treasury Research said in a morning note.

Gains in Hong Kong were more decisive, with the Hang Seng Index rising 1.66 percent by 3:02 p.m. HK/SIN as financials led gains on the index. China Construction Bank rose 2.88 percent to contribute 75 points to the index's overall 492.84-point gain, while HSBC tacked on 1.05 percent an hour before the market close.

The property sector turned positive as the session wore on, with large cap property stocks carving out gains. Country Garden gained 4.27 percent and CK Asset traded flat. Meanwhile, oil-related and technology stocks also climbed, with CNOOC rising 2.47 percent and index heavyweight Tencent advancing 1.93 percent.

Elsewhere, Taiwan's benchmark Taiex closed higher by 2.81 percent as markets re-opened for trade after the Lunar New Year holiday. Apple suppliers put in a strong showing, with Largan Precision and Pegatron rising 7.34 percent and 5.12 percent, respectively.

Markets in China remained closed for the Lunar New Year holiday.

The Dow Jones industrial average lost 1.01 percent, or 254.63 points, snapping a six-day winning streak. The steep losses came on the back of Walmart stock tumbling 10.2 percent after the retailer reported lower-than-expected earnings . Other U.S. stock indexes saw smaller declines.

Higher U.S. bond yields also pressured stocks stateside in the last session. The 2-year Treasury note yield stood at 2.26 percent on Wednesday after touching its highest levels in almost a decade overnight. The yield on the benchmark 10-year U.S. Treasury note last stood at 2.895 percent.

Ahead, investors awaited the release of minutes from the Federal Reserve due during U.S. hours. The release of the minutes come after a recent rout in stock markets on concerns over rising interest rates. Markets are awaiting the minutes for clues on the central bank's future policy.

The dollar firmed on Wednesday, extending gains made overnight following the overnight

```
U.S. Treasury auction . The dollar index, which tracks the U.S. currency against
a basket of six rivals, stood at 89.834 at 2:46 p.m. HK/SIN. Against the yen, the
dollar edged up to trade at 107.71, above the 106 handle it started the week at.
On the commodities front, U.S. crude futures declined 0.99 percent to trade at
$61.18 per barrel. Brent crude futures edged down by 0.7 percent to trade at
$64.79.
```

可以看出，这篇文章是符合bool查询要求的。

3. 将TF-IDF矩阵用稀疏矩阵的方式存储，存为npz文件，只需300MB多空间

在`tf_idf_construct.py`的`TF_IDF_construct`类中计算TF-IDF矩阵，由于其稀疏性质，为节省存储空间，利用`scipy`和`numpy`库将其以稀疏矩阵方式存储为`TF_IDF.npz`。

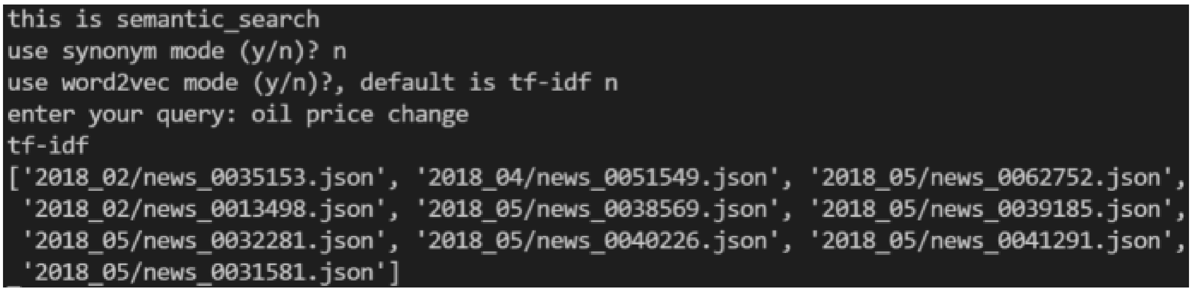
4. 通过余弦相似度排序得到前10相关的文档

在`semantic_search.py`中对query进行分词预处理和计算TF-IDF向量，并根据与TF-IDF矩阵中各文档向量的余弦相似度，排序得到前十相关的文档返回给用户。

结果展示

接下来展示TF-IDF查询情况，查询：`oil price change`

结果为：



```
this is semantic_search
use synonym mode (y/n)? n
use word2vec mode (y/n)?, default is tf-idf n
enter your query: oil price change
tf-idf
['2018_02/news_0035153.json', '2018_04/news_0051549.json', '2018_05/news_0062752.json',
 '2018_02/news_0013498.json', '2018_05/news_0038569.json', '2018_05/news_0039185.json',
 '2018_05/news_0032281.json', '2018_05/news_0040226.json', '2018_05/news_0041291.json',
 '2018_05/news_0031581.json']
```

打开前两篇文档：

第一篇

```
Here's where oil prices are headed next: Oil analyst 1 Hour Ago Mike Kelly,
Seaport Global Securities senior analyst, discusses the volatility in the oil
market and where prices could go from here..
```

第二篇

```
136 COMMENTS Oil prices are headed toward $70 a barrel, a weight on the U.S.
economy that is bearable for now but could pose trouble if prices keep climbing.
The last time U.S. oil prices were at $70, in 2014, they were in the middle of a
steep
collapse. Many investors believed then that prices would soon stabilize, or even
recover. Instead, they continued to plunge, eventually hitting a bottom in 2016
at $26. That
tumble caused acute pain for oil producers, whose troubles rippled out into
stocks, bonds and the broader economy.
```

This year’s rally is a sign of how much has changed in a few years. Global growth has picked up, while U.S. unemployment has fallen. A gambit by the world’s largest oil producers to cut production has been succeeding in eliminating a massive glut, with help from soaring demand.

Oil prices have climbed more than 60% since last summer’s lows, and U.S. producers are exporting more crude than ever.

For now, some investors say oil prices are lodged in a range that could benefit the U.S. economy by bolstering the recovering energy industry without curtailing demand.

Yet even with the economy chugging along, rising oil prices dredge up fresh concerns. If crude continues to move higher, it could begin to stifle economic growth. Higher consumer prices for gasoline and other energy products act like a tax, while pushing inflation higher and increasing pressure on the Federal Reserve to raise interest rates more aggressively.

That, in turn, could slow growth and weigh on the stock market, which has already been knocked around by trade tensions, rising bond yields and recent bouts of volatility. Inflation concerns pushed the yield on the 10-year Treasury note to the highest since 2014 on Friday, while major U.S. stock indexes closed lower, wiping out much of the recent gains after a string of upbeat earnings.

“Nothing can suck cash flow out of the economy faster than rising oil prices,” said Joseph LaVorgna, chief economist for the Americas at Natixis .

When oil prices fell below $40 a barrel, financial distress from the energy sector started to spread, said Jason Thomas, director of research at the Carlyle Group .

But if oil prices continue rising, they could boost inflation expectations, which would raise bond yields and the cost of financing.

“We’re starting to move out of that Goldilocks zone,” Mr. Thomas said. “Certainly $10 to $15 a barrel more there starts to be this drag.”

President Donald Trump tweeted Friday that oil prices are “artificially Very High!”—a sentiment that would have been unthinkable even a few months ago. Oil prices tumbled after his comment, but recovered to settle at $68.38 a barrel Friday.

A major force behind rising oil prices has been a policy reversal from the Organization of the Petroleum Exporting Countries. In 2014, the group opted to continue pumping oil at high rates in an effort to protect its market share against encroaching U.S. shale producers. Two years later, OPEC reversed course, enlisting other major producers such as Russia in a coordinated production cut that has helped to nearly eliminate a supply overhang.

“The conversation is changing,” said Antoine Halff, senior research scholar at Columbia University’s Center on Global Energy Policy. “A year ago the conversation was ‘lower for longer’ and the ‘age of abundance’” for oil, he said. Now, “the idea of cheap oil forever is being challenged.”

A booming global economy has also been key, keeping demand high as excess oil and fuel gets soaked up by consumers around the world. The first quarter was likely the strongest for global oil demand growth, year over year, since the fourth quarter of 2010, Goldman Sachs said.

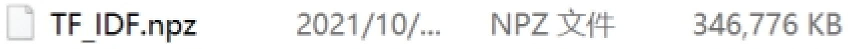
But higher prices could threaten that. When drivers take to the road this summer, they will likely be paying the highest prices for gasoline since 2014. That will likely negate any financial benefits from tax cuts this year for low-income households, according

```
to Deutsche Bank, and could further eat into disposable income.
“The higher prices get, over all, the consumer side of the economy will be
affected. It’s like a tax increase on consumers as gasoline prices go up,” said
Ann-Louise Hittle,
vice president of oil markets at research firm Wood Mackenzie.
Some analysts believe that concerns that higher prices will cut into demand are
overblown.
For one thing, oil’s gains have been gradual. Gasoline prices are still far from
highs
in 2008, when the national average topped $4 a gallon at times.
Demand has remained strong even as oil and fuel prices have been rising, and many
analysts believe that prices still aren’t high enough to prompt big changes in
behavior.
And with the U.S. now on track to overtake Russia as the world’s largest oil
producer,
a large swath of the U.S. economy stands to benefit from higher prices. Oil
prices are
even higher abroad, which has made it lucrative for U.S. producers to ship more
crude overseas. Brent, the global benchmark, climbed to $74.06 a barrel Friday.
“In the past, any time oil prices have gone up it was as a result of supply
constraints and the U.S. was at the mercy of foreign oil,” said Joseph Tanious,
senior investment
strategist at Bessemer Trust. “But U.S. oil production has picked up in a
meaningful way—there could be also some benefits to having modestly rising oil
prices.”
Oil’s rise has started to lift energy companies’ share prices, which had been
slow to react to higher prices. Oil-and-gas companies have taken over as the U.S.
stock market’s priciest segment, according to Credit Suisse analysts. Energy
shares have gained 1.5%
so far this year after a nearly 10% gain over the past month.
But even some producers worry about what will happen if higher oil prices stick
around
too long.
“We’re going to lose demand. It’s going to move more toward alternative energy,”
Scott
Sheffield, chairman of Pioneer Natural Resources Co., said at an energy
conference last week. “I don’t think it does anybody any good to see $70, $80
crude.”
Write to Stephanie Yang at stephanie.yang@wsj.com and Alison Sider at
alison.sider@wsj.com
```

可见语义情况符合良好。

选做内容：

1. 对TF-IDF矩阵进行了空间优化，用稀疏矩阵的方式，若稠密矩阵存储，预计需要上百GB空间，而稀疏矩阵只需要300MB多的空间



TF_IDF.npz 2021/10/... NPZ 文件 346,776 KB

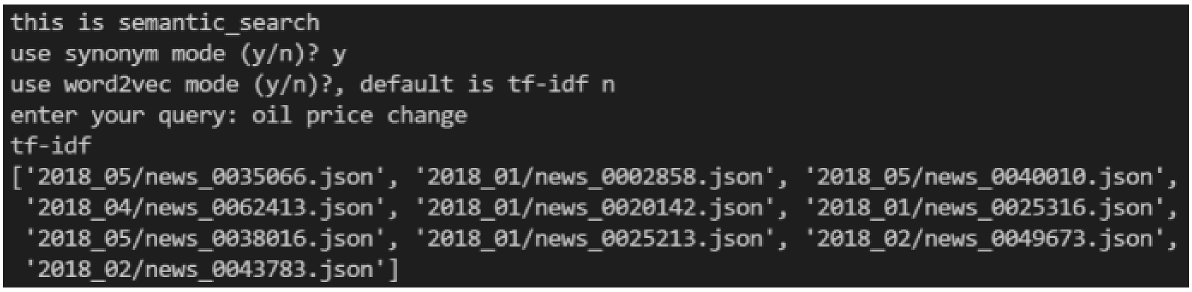
2. 采用外部知识库（同义词表）优化索引效果

在`Semantic_Search`中新增参数`synonym_tag`来开启同义词表优化功能。

优化结果展示（与原来版本对比）

和之前一样搜索 `oil price change`

结果如下:



```
this is semantic_search
use synonym mode (y/n)? y
use word2vec mode (y/n)?, default is tf-idf n
enter your query: oil price change
tf-idf
['2018_05/news_0035066.json', '2018_01/news_0002858.json', '2018_05/news_0040010.json',
 '2018_04/news_0062413.json', '2018_01/news_0020142.json', '2018_01/news_0025316.json',
 '2018_05/news_0038016.json', '2018_01/news_0025213.json', '2018_02/news_0049673.json',
 '2018_02/news_0043783.json']
```

第一篇文档内容:

```
* U.S. crude stocks rise by 4.9 mln bbl to 435.6 mln bbl -API
* Physical spot cargoes trade at discount to financial crude
SINGAPORE, May 16 (Reuters) - Oil prices fell on Wednesday, weighed down by ample
supplies despite ongoing output cuts by producer cartel OPEC and looming U.S.
sanctions against major crude exporter Iran.
Brent crude futures, the international benchmark for oil prices, were at $78.07
per barrel at 0024 GMT, down 36 cents, or 0.5 percent, from their last close.
U.S. West Texas Intermediate (WTI) crude futures were at $71.02 a barrel, down 28
cents, or 0.4 percent, from their last settlement.
Despite the dips, both financial oil benchmarks remained close to their November
2014 highs of $79.47 and $71.92 a barrel respectively, reached the previous day.
But there are signs in physical crude markets that may give pause to financial
investors.
Spot crude oil cargo prices are at their steepest discounts to futures prices in
years
as sellers are struggling to find buyers for West African, Russian and Kazakh
cargoes,
while pipeline bottlenecks trap supply in west Texas and Canada.
The bottleneck in North America likely contributed to a 4.9 million barrel rise
in U.S. crude oil inventories, to 435.6 million barrels, that the American
Petroleum Institute reported on Tuesday.
"The API inventory data in the U.S. fits with ... a topping pattern or at least a
decent pause for oil prices at the moment," said Greg McKenna, chief market
strategist at futures brokerage AxiTrader.
Despite Wednesday's dips and some indicators implying the financial oil has
overshot physical oil, overall crude market conditions have tightened since 2017
when the Organization of the Petroleum Exporting Countries (OPEC), led by Saudi
Arabia, started to withhold supplies to push up oil prices.
With renewed U.S. sanctions looming against OPEC-member Iran and oil demand
strong, analysts said crude markets will likely remain relatively tight for much
of the year.
Stronger oil prices are also spilling into other markets.
"A rising oil price brings upside price risk to all commodities," Morgan Stanley
said in a note to clients this week.
The U.S. bank said rising diesel prices contributed 10-20 percent to cash costs
in the
metals and dry-bulk sectors, while the price of oil also significantly
contributed to power generation.
```

```
"Finally, transport costs (5-20 percent of cash costs) will also rise in
response, with the heaviest impact on bulk commodity producers," Morgan Stanley
said. (Reporting by Henning Gloystein; Editing by Joseph Radford)
```

可以看出，引入同义词表后，搜索内容更加优质！

3. 采用word2vec语义表征方式表征查询和文档

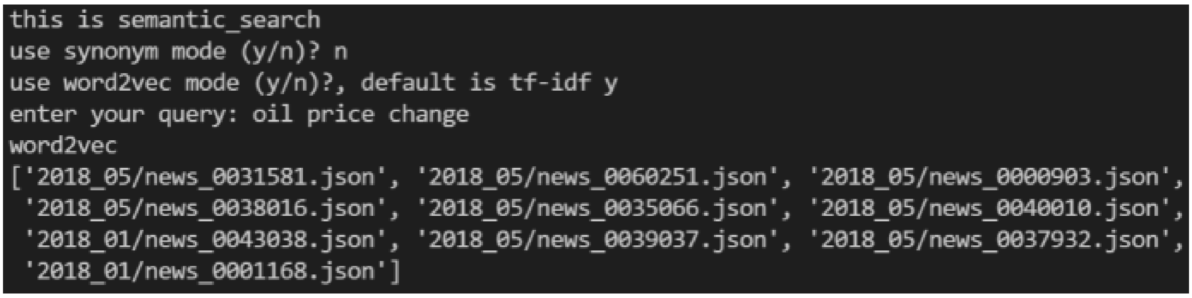
在`Semantic_Search`中参数`embedding_type`来决定语义表征方式。

`word2vec_train.py`中借助`gensim.models.word2vec`使用所有文档的语料训练word2vec模型并存储在`word2vec.model`文件中。

在`semantic_search.py`中加载该模型，并计算各文档和query的相似度和排序。

结果展示（与原来版本对比）

查询`oil price change`结果如下：



```
this is semantic_search
use synonym mode (y/n)? n
use word2vec mode (y/n)?, default is tf-idf y
enter your query: oil price change
word2vec
['2018_05/news_0031581.json', '2018_05/news_0060251.json', '2018_05/news_0000903.json',
 '2018_05/news_0038016.json', '2018_05/news_0035066.json', '2018_05/news_0040010.json',
 '2018_01/news_0043038.json', '2018_05/news_0039037.json', '2018_05/news_0037932.json',
 '2018_01/news_0001168.json']
```

其中一篇文档内容：

```
* U.S. crude stocks rise by 4.9 mln bbl to 435.6 mln bbl -API
* Physical spot cargoes trade at discount to financial crude
* Production by oil majors rising - S&P Global Ratings (Adds S&P Global quote,
updates
prices)
SINGAPORE, May 16 (Reuters) - Oil prices fell on Wednesday, weighed down by ample
supplies despite ongoing output cuts by producer cartel OPEC and looming U.S.
sanctions against major crude exporter Iran.
Brent crude futures were at $78.17 per barrel at 0210 GMT, down 26 cents, or 0.3
percent, from their last close.
U.S. West Texas Intermediate (WTI) crude futures were at $71.02 a barrel, down 29
cents, or 0.4 percent, from their last settlement.
Despite the dips, both financial oil benchmarks remained close to their November
2014 highs of $79.47 and $71.92 a barrel respectively, reached the previous day.
But there are signs in physical crude markets that may give pause to financial
investors.
There are also signs that oil production will rise, especially at majors like
ExxonMobil, Royal Dutch Shell , Chevron, BP and Total.
"Aggregate production - both actual and projected - is growing for the majors,"
S&P Global Ratings said in a report published on Tuesday.
Spot crude oil cargo prices are at their steepest discounts to futures prices in
years
as sellers are struggling to find buyers for West African, Russian and Kazakh
cargoes,
while pipeline bottlenecks trap supply in west Texas and Canada.
```

```
The bottleneck in North America likely contributed to a 4.9 million barrel rise
in U.S. crude oil inventories, to 435.6 million barrels, that the American
Petroleum Institute reported on Tuesday.
"The API inventory data in the U.S. fits with ... a topping pattern or at least a
decent pause for oil prices at the moment," said Greg McKenna, chief market
strategist at futures brokerage AxiTrader.
Despite Wednesday's dips and some indicators implying the financial oil has
overshot physical oil, overall crude market conditions have tightened since 2017
when the Organization of the Petroleum Exporting Countries (OPEC), led by Saudi
Arabia, started to withhold supplies to push up oil prices.
With renewed U.S. sanctions looming against OPEC-member Iran and oil demand
strong, analysts said crude markets will likely remain relatively tight for much
of the year.
Stronger oil prices are also spilling into other markets.
"A rising oil price brings upside price risk to all commodities," Morgan Stanley
said in a note to clients this week.
```

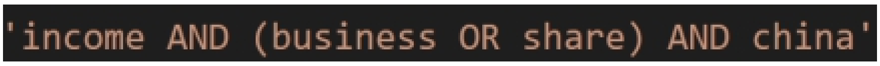
可见，使用word2vec语义表征方式，搜索时间约是普通语义查询的一半，即性能有了提高，且效果有了显著提升！

4. 图片查询

从 json 文件中 main_image字段获取，支持bool查询和semantic 查询，有些文档main_image字段为空，则会跳过这些文档，另外，如果链接是https://s4.reutersmedia.net/resources_v2/images/rcom-default.png，则认为这是无效链接，并跳过。最多返回10篇文档的图片链接，并打开图片，注意，很多URL需要VPN才能访问，所以在打开图片时需要配置VPN，所以在代码中默认不直接打开（可以设置直接打开）

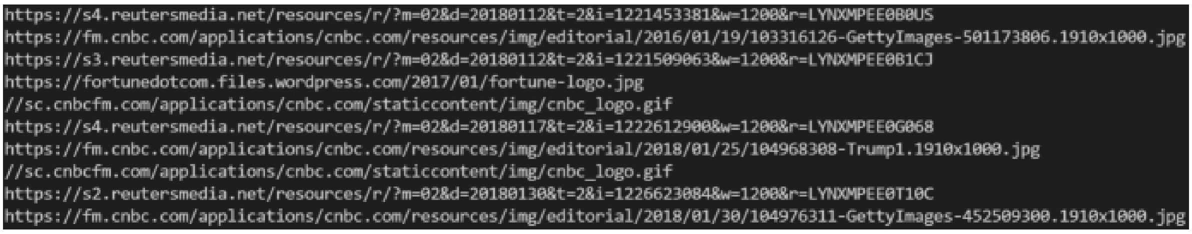
测试结果

1. bool 查询



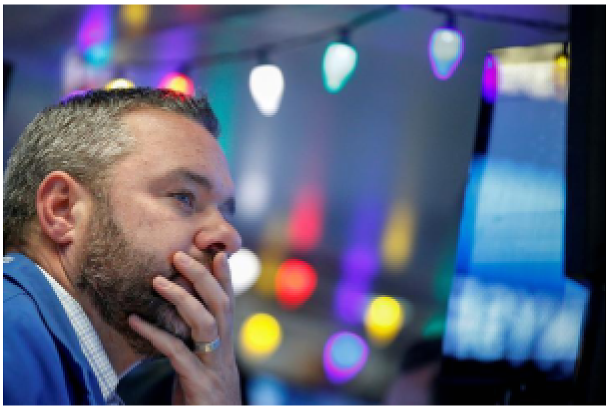
```
'income AND (business OR share) AND china'
```

结果得到10个图片的链接

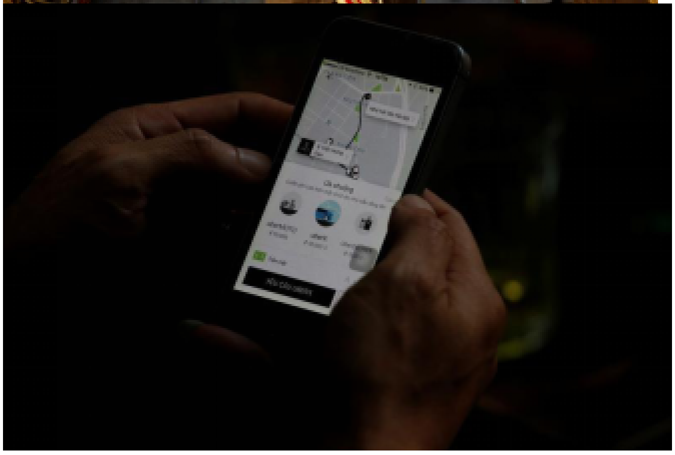


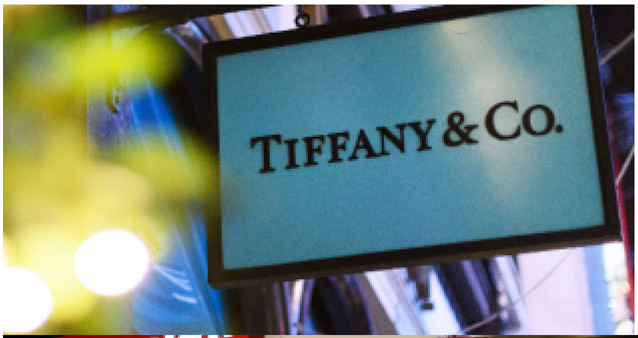
```
https://s4.reutersmedia.net/resources/r/?m=02&d=20180112&t=2&i=1221453381&w=1200&r=LYNXMPEE0B0US
https://fm.cnbc.com/applications/cnbc.com/resources/img/editorial/2016/01/19/103316126-GettyImages-501173806.1910x1000.jpg
https://s3.reutersmedia.net/resources/r/?m=02&d=20180112&t=2&i=1221509063&w=1200&r=LYNXMPEE0B1CJ
https://fortunedotcom.files.wordpress.com/2017/01/fortune-logo.jpg
//sc.cnbcfm.com/applications/cnbc.com/staticcontent/img/cnbc_logo.gif
https://s4.reutersmedia.net/resources/r/?m=02&d=20180117&t=2&i=1222612900&w=1200&r=LYNXMPEE0G068
https://fm.cnbc.com/applications/cnbc.com/resources/img/editorial/2018/01/25/104968308-Trump1.1910x1000.jpg
//sc.cnbcfm.com/applications/cnbc.com/staticcontent/img/cnbc_logo.gif
https://s2.reutersmedia.net/resources/r/?m=02&d=20180130&t=2&i=1226623084&w=1200&r=LYNXMPEE0T10C
https://fm.cnbc.com/applications/cnbc.com/resources/img/editorial/2018/01/30/104976311-GettyImages-452509300.1910x1000.jpg
```

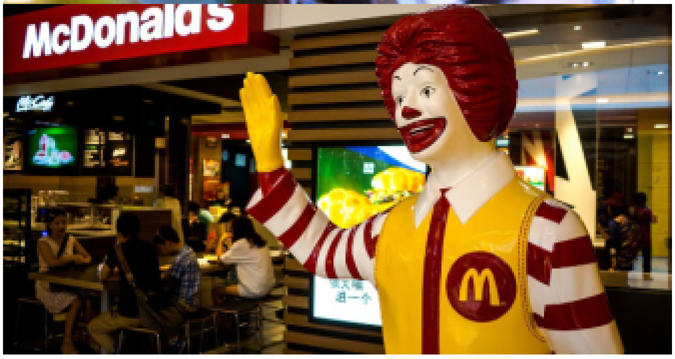
我们打开这些图片



McDonald's

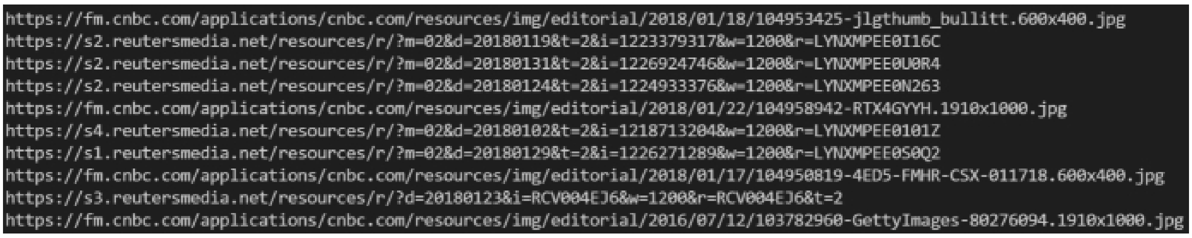






2. 语义查询`company loss billion`

结果得到如下10个图片链接



```
https://fm.cnbc.com/applications/cnbc.com/resources/img/editorial/2018/01/18/104953425-jlgthumb_bullitt.600x400.jpg
https://s2.reutersmedia.net/resources/r/?m=02&d=20180119&t=2&i=1223379317&w=1200&r=LYNXMPEE0I16C
https://s2.reutersmedia.net/resources/r/?m=02&d=20180131&t=2&i=1226924746&w=1200&r=LYNXMPEE0U0R4
https://s2.reutersmedia.net/resources/r/?m=02&d=20180124&t=2&i=1224933376&w=1200&r=LYNXMPEE0N263
https://fm.cnbc.com/applications/cnbc.com/resources/img/editorial/2018/01/22/104958942-RTX4GYYH.1910x1000.jpg
https://s4.reutersmedia.net/resources/r/?m=02&d=20180102&t=2&i=1218713204&w=1200&r=LYNXMPEE0101Z
https://s1.reutersmedia.net/resources/r/?m=02&d=20180129&t=2&i=1226271289&w=1200&r=LYNXMPEE0S0Q2
https://fm.cnbc.com/applications/cnbc.com/resources/img/editorial/2018/01/17/104950819-4ED5-FMHR-CSX-011718.600x400.jpg
https://s3.reutersmedia.net/resources/r/?d=20180123&i=RCV004EJ6&w=1200&r=RCV004EJ6&t=2
https://fm.cnbc.com/applications/cnbc.com/resources/img/editorial/2016/07/12/103782960-GettyImages-80276094.1910x1000.jpg
```

展示其中的图片：

JAY LENO'S GARAGE
CNBC

D·R·HORTON
America's Builder

Ford

CSX
(CSX)
CSX
57.82
-0.31
[-0.53%]
INTRA DAY
58.13
CLOSE
56.5
9:30A
12P
2P
4P
CNBC MARKET FLASH
CSX: TAX REFORM WILL MEAN MORE CASH FOR SHAREHOLDERS
CNBC

CDU

JPMorgan