

实验报告

设计思路：文本相似度计算方法有 2 个子任务, 即文本表示模型和相似度度量方法，文本表示模型将文本表示为可以计算的数值向量，根据特征构建词向量；相似度度量方法可以根据词向量计算文本之间的相似度。本文实现了 tfidf 生成权重向量、simhash 等构建词向量的方法、分别使用余弦相似度、海明距离进行计算相似度。并利用多进程方法进行优化，最后对各种方法进行了小结。

全文目录

- 一、数据预处理.....2
- 二、文本表示模型-词向量.....4
 - 2.1 TF-IDF 值作为词向量的权重4
 - 2.2 利用 scipy 库处理稀疏矩阵.....5
 - 2.3 利用 gensim 库生成词向量7
 - 2.4 利用 Simhash 生成词向量并计算相似度.....7
- 三、计算向量的相似度.....9
 - 3.1 计算使用 tf-idf 加权的向量相似度.....9
 - 3.2 对 simhash 计算海明距离..... 11
- 四、利用多进程计算优化计算时间..... 12
- 五、总结： 15

一、数据预处理

1. 观察给定的语料集可以发现, 该语料是已经分词和标注好的新闻文章语料, 首先需要对该数据集进行预处理, 处理成规范化使用的格式。

19980101-01-001-001/m 迈向/v 充满/v 希望/n 的/u 新/a 世纪/n ——/w 一九九
19980101-01-001-002/m 中共中央/nt 总书记/n 、/w 国家/n 主席/n 江/nr 泽民/
19980101-01-001-003/m (/w 一九九七年/t 十二月/t 三十一日/t) /w
19980101-01-001-004/m 12月/t 31日/t , /w 中共中央/nt 总书记/n 、/w 国
。/w (/w 新华社/nt 记者/n 兰/nr 红光/nr 摄/Vg) /w
19980101-01-001-005/m 同胞/n 们/k 、/w 朋友/n 们/k 、/w 女士/n 们/k 、/w
19980101-01-001-006/m 在/p 1998年/t 来临/v 之际/f , /w 我/r 十分/m 高
 , /w 向/p 全国/n 各族/r 人民/n , /w 向/p 香港/ns 特别/a 行政区/n 同胞/n 、/
挚/a 的/u 问候/vn 和/c 良好/a 的/u 祝愿/vn ! /w

2. 读取停用词表，剔除与表示文章特征无关的符号和相关词汇，部分停用词如下：

```
In [1]: #读取停用词
stopwords=[]
with open("E:/pythonwork/停用词.txt") as g:
    while True:
        line=g.readline()
        if line=="":
            break
        line=line.strip()
        stopwords.append(line)
print(stopwords)
```

3. 读取数据集，利用正则表达式提取分词，并剔除停用词，生成的不重复的词袋中词语的数量为 54033。

```
In [2]: ##构建词袋
import re
wordbags=[]
with open("E:/pythonwork/199801_clear.txt") as f:
    file=f.read()
    word=re.findall("[^a-zA-Z].*?)/",file) #匹配斜线前的词语
    for str in word:
        str=str.strip()
        if (str not in stopwords) and len(str)!=19 and str!="":
            wordbags.append(str)
wordbags=list(set(wordbags))
print("词袋中的数量是:",len(wordbags))
print(wordbags)
```

词袋中的数量是: 54033

['核能', '2 2 5. 0 3 亿', '发财梦', '催缴', '公映', '低毒', '银色', '汇集', '官僚资本', '证', '圣但尼油', '续', '失足者', '密林', '牢记', '村子', '2: 0', '地方军', '小褂儿', '电缆桥', '1 1 8 1 2', '熙' '两侧', '4 1. 1', '虚名', '石南', '八运', '神经中枢', '爱诗', '汽配', '谈笑', '拔河', '零点六九一六' '物', '笨', '激增', '倒行逆施', '肥西县', '尔均', '3 8. 6 0', '塔里木', '回头客', '襄阳', '薄壳核桃', '发', '一把', '生辉', '床铺', '道教', '平度市', '远科', '福斯特', '为言者', '如为', '餐饮', '8 5. 8 万' '链', '步调一致', '老工人', 'GB 9 2 5 4', '工本费', '1 · 1 万', '男双', '偷税额', '恩望', '醒世', '将' '国桐', '烟客', '庆新', '销售税', '怒吼', '言谈', '围栏', '达夫', '建设者', '熙坤', '张掖', '高出一筹' '雷州', '乐仁', '拉·甘地', '热', '急迫', '乔迁之喜', '里道', '沂蒙', '在场', '清河径', '士华', '鹤壁' '张贴者', '零点几', '兄弟', '抛光', '玉工', '富集', '是之', '病禽', '萍踪浪迹', '绽开', '亚非拉', '知

4. 观察数据集发现一篇文章被分成了不同文段，进行文章的拼接和标准化处理，拼接后的文档数量为 3147 篇。标准化的文档样例如下：

19980101-01-001-001/m 迈向/v 充满/v 希望/n 的/u 新/a 世纪/n ——/w 一九九
19980101-01-001-002/m 中共中央/nt 总书记/n 、/w 国家/n 主席/n 江/nr 泽民/
19980101-01-001-003/m (/w 一九九七年/t 十二月/t 三十一日/t) /w
19980101-01-001-004/m 1 2月/t 3 1日/t , /w 中共中央/nt 总书记/n 、/w 国
。/w (/w 新华社/nt 记者/n 兰/nr 红光/nr 摄/Vg) /w
19980101-01-001-005/m 同胞/n 们/k 、/w 朋友/n 们/k 、/w 女士/n 们/k 、/w
19980101-01-001-006/m 在/p 1 9 9 8年/t 来临/v 之际/f , /w 我/r 十分/m 高
、/w 向/p 全国/n 各族/r 人民/n , /w 向/p 香港/ns 特别/a 行政区/n 同胞/n 、/
挚/a 的/u 问候/vn 和/c 良好/a 的/u 祝愿/vn ! /w
19980101-01-001-007/m 1 9 9 7年/t , /w 是/v 中国/ns 发展/vn 历史/n 上/f :
的/u 遗产/n , /w 继续/v 把/p 建设/v 有/v 中国/ns 特色/n 社会主义/n 事业/n 持
制/r " /w 、/w 、/w " /w 港人治港/l " /w 、/w 高度/d 自治/v 的/u 方针/n 保持/v i
会/n , /w 高举/v 邓小平理论/n 伟大/a 旗帜/n , /w 总结/v 百年/m 历史/n , /w
19980101-01-001-008/m 在/p 这/r 一/m 年/q 中/f , /w 中国/ns 的/u 改革/vn
低/a 通胀/j " /w 的/u 良好/a 发展/vn 态势/n 。/w 农业/n 生产/vn 再次/d 获得
经济/n 技术/n 合作/vn 与/c 交流/vn 不断/d 扩大/v 。/w 民主/a 法制/n 建设/vn
最近/t 一个/m 时期/n 一些/m 国家/n 和/c 地区/n 发生/v 的/u 金融/n 风波/n ,
况/n 会/v 逐步/d 得到/v 缓解/vn 。/w 总的来说/c , /w 中国/ns 改革/v 和/c 发

```
print(documents_st[0])##查看第一篇文章格式
```

['迈向', '充满', '希望', '新', '世纪', '一九九八年', '新年', '讲话', '附', '图片', '张', '中共中央', '总书记', '国家', '主席', '江', '泽民', '一九九七年', '十二月', '三十一日', '1 2月', '3 1日', '中共中央', '总书记', '国家', '主席', '江', '泽民', '发表', '1 9 9 8年', '新年', '讲话', '迈向', '充满', '希望', '新', '世纪', '新华社', '记者', '兰', '红光', '摄', '同胞', '朋友', '女士', '1 9 9 8年', '来临', '之际', '中央', '广播', '电台', '中国', '国际', '广播', '电台', '中央', '电视台', '全国', '各族', '香港', '特别', '行政区', '同胞', '澳门', '台湾', '同胞', '海外', '侨胞', '世界', '各国', '朋友', '致以', '诚挚', '问候', '祝愿', '1 9 9 7年', '中国', '发展', '历史', '平凡', '年', '中国', '决心', '继承', '邓', '小平', '同志', '遗产', '建设', '中国', '特色', '事业', '推向', '中国', '政府', '顺利', '恢复', '香港', '行使', '主权', '一国两制', '港人治港', '高度', '自治', '方针', '香港', '繁荣', '稳定', '中国', '共产党', '成功', '第十五', '次', '全国', '代表大会', '高举', '邓小平理论', '旗帜', '百年', '历史', '展望', '新', '世纪', '制定', '中国', '跨', '世纪', '发展', '纲领', '年', '中', '中国', '改革', '开放', '现代化', '建设', '向前', '迈进', '国民经济', '高', '增长', '低', '通胀', '发展', '态势', '农业', '生产', '收成', '企业', '改革', '深化', '生活', '进一步', '改善', '对外', '经济', '技术', '合作', '交流', '民主', '法制', '建设', '精神文明', '建设', '各项', '事业', '新', '进展', '关注', '时期', '国家', '地区', '发生', '金融', '风波', '国家', '地区', '努力', '国际', '合作', '情况', '缓解', '中国', '改革', '发展', '全局', '稳定', '年', '中', '中国', '外交', '工作', '成果', '高层', '互访', '中国', '美国', '俄罗斯', '法国', '日本', '大国', '关系', '未来', '发展', '目标', '指导', '方针', '中国', '周边', '国家', '发展中国家', '友好', '合作', '进一步', '中国', '参与', '亚', '太', '经合', '组织', '活动', '参加', '东盟', '中', '日', '韩', '中国', '东盟', '首脑', '非正式', '会晤', '外交', '活动', '符合', '和平', '发展', '时代', '主题', '顺应', '世界', '走向', '多极化', '趋势', '国际', '社会', '友好', '合作', '发展', '作出', '贡献', '1 9 9 8年', '中国', '满怀信心', '开创', '新', '业绩', '经济', '社会', '发展', '中', '面临', '困难', '邓小平理论', '指引', '改革', '开放', '2 0', '年', '成就', '积累', '经验', '条件', '克服', '困难', '稳步前进', '进一步', '解放思想', '实事求是', '抓住', '机遇', '开拓进取', '建设', '中国', '特色', '道路', '越', '走', '越', '宽广', '祖国', '统一', '海内外', '中国', '心愿', '中', '葡', '合作', '努力', '一国两制', '方针', '澳门', '基本法', '1 9 9 9年', '1 2月', '澳门',

二、文本表示模型-词向量

2.1 TF-IDF 值作为词向量的权重

1. 考虑程序的计算效率。因为字典的查询时间复杂度为 $O(1)$, 所以我们用字典保存每篇文档的词频和反文档频率来优化查询的效率。首先计算所有文章的 TF。Coutfreq 方法中我们加入了一个 num 参数, 根据课堂 ppt 上的 tf 变式公式。Num 值为 0 时为计算文章的普通词频、num 值为 1 时为词频/当前文章的最大词频(TF/TFmax)、num 值为 2 时计算词频/当前文档长度。后两种方法实际上是对 tf 词频进行标准化而消除不同文本长度对关键词计算的影响。

2.1.1 实现tf-idf (词频、最大词频标准化、文章长度标准化)

```
[8]: #tf-idf算法
    ##1. 统计tf词频
    def coutfreq(doc, num=0): #统计该文档内的词频 num参数可以选模式, 默认0为普通词频, 1为除以最大词频, 2为除以文本长度
        worddict={}
        for word in doc:
            if word in worddict:
                worddict[word]+=1
            else:
                worddict[word]=1
        if(num==0):
            return worddict
        elif(num==1):
            maxtf= max(worddict.values())
            for k in worddict.keys():
                worddict[k]=(worddict[k]/maxtf)
            return worddict
        elif(num==2):
            for k in worddict.keys():
                worddict[k]=(worddict[k]/len(worddict))
            return worddict
```

计算词频所需的时间如下:

```
: %%time
# ##测试tfreq
# wd1=coutfreq(documents_st[0], 1)
# print(wd1)
tflist_1=[] ##标准化tf
tflist_0=[] ##词频tf
tflist_2=[] ##文章长度修正tf
for doc in documents_st:
    tflist_1.append(coutfreq(doc, 1))
    tflist_2.append(coutfreq(doc, 2))
    tflist_0.append(coutfreq(doc, 0))
```

Wall time: 368 ms

2. 同理计算每篇文章中不同词汇的反文档频率, 计算时间为 8min37s (可以看出由于数据规模较大, 运算时间比较长, 后续会做优化)

```

1: import math
def countfreq_idf(wordbags, documents):
    dict_idf={}
    for word in wordbags:
        num=0
        for doc in documents:
            if word in doc:
                num+=1
            idf = math.log(len(documents)/(num+1))
            dict_idf[word]=idf
    return dict_idf

```

```

1: %%time
dict_idf=countfreq_idf(wordbags, documents_st)

```

Wall time: 8min 37s

3. 计算每篇文章的所有词的 TF-IDF 值，这里我们打印第一篇文章中 TF-IDF 值前五的关键词：

```

1: def count_tfidf(word, doc_index, tflist, dict_idf):
    '''word 查询词
    doc_index 查询文章在文档集中索引
    tflist_tf tf词频字典列表
    dict_idf idf字典
    '''
    dict_tf=tflist[doc_index]
    tf=dict_tf[word]
    idf=dict_idf[word]
    return tf*idf

```

```

1: ##第一篇文章中标准化后前五的关键词
d={}
for i in documents_st[0]:
    d[i]=count_tfidf(i,0,tflist_1,dict_idf)
d_order=sorted(d.items(),key=lambda x:x[1],reverse=True)
print(d_order[:5])

```

```

[('中国', 1.215799696217063), ('世纪', 0.7398860001749124), ('国际', 0.7211169028429748), ('各国', 0.6930868913233472), ('发展', 0.68290
01640957153), ('和平', 0.616761395053676), ('澳门', 0.6095188833395576), ('台湾', 0.6038980163701821), ('世界', 0.5900235786645349),
('合作', 0.5826549487780065), ('两岸', 0.5337519078932682), ('同胞', 0.5202980748692225), ('经济', 0.5063212086835828), ('东盟', 0.47135
9135355503), ('充满', 0.43754100068298324), ('外交', 0.4311293388845394), ('友好', 0.41044003407484275), ('交流', 0.40125222632572694),
('迈向', 0.3996756965869104), ('方针', 0.3779708719589964), ('新年', 0.3708260702290029), ('一国两制', 0.3702739153683485), ('公正', 0.3
68051561444351), ('希望', 0.3660452904301579), ('发展中国家', 0.3658869375778967), ('香港', 0.36406766757021836), ('建设', 0.3626368439
1402323), ('1 2 月', 0.36046511275328186), ('进一步', 0.3604267808731059), ('早日', 0.3558346052621788), ('开创', 0.3539618672745073),
('电台', 0.3521302917146093), ('3 1 日', 0.3503381079461957), ('新', 0.3481161908047018), ('改革', 0.32904528366146946), ('1 9 8 年',
0.3248604727133382), ('总书记', 0.3207926898061201), ('广播', 0.3159604183352576), ('越', 0.31479489817069595), ('民心所向', 0.306710738
1876859), ('秩序', 0.30497964519933063), ('稳定', 0.30021446459946954), ('机遇', 0.2990247315341519), ('环顾', 0.289816358683179), ('弱
势', 0.289816358683179), ('朋友', 0.2857693403150113), ('邓小平理论', 0.28178133855020837), ('道路', 0.28100613916168227), ('葡', 0.2778
2960566435483), ('一帆风顺', 0.27782960566435483)]

```

2.2 利用 scipy 库处理稀疏矩阵

1. 首先手动构建词向量，遍历词袋中的词汇，若文章中出现该词汇则计算该词汇的 tf-idf 值作为权重，若未出现过则设为 0，这样会产生非常稀疏的向量空间，因此需要对稀疏矩阵进行处理。

```

: def build_matrix(documents_st, tflist, dict_idf):
    matlist=[]
    for i in range(len(documents_st)):
        print("r", end="")
        print("Process progress: {}% ".format((i*100)/len(documents_st)), "■" * (int)((i*100)/len(documents_st) // 2), end="")
        sys.stdout.flush()
        time.sleep(0.05)
        m=[count_tfidf(word, i, tflist, dict_idf) if word in documents_st[i] else 0 for word in wordbags]
        sp_mat=np.array(m)
        matlist.append(sp_mat)
    return matlist

```

```
Download progress: 99.96822370511599% ██████████
Wall time: 12min 1s
```

COO 优点: :容易构造,比较容易转换成其他的稀疏矩阵存储格式, 时间复杂度 $O(1)$ 并且支持相同的(row,col)坐标上存放多个值。

因此 COO 矩阵的适用场景为加载数据文件时使用，可以快速构建稀疏矩阵，然后调用 `to_csr()`、`to_csc()`、`to_dense()` 把它转换成 CSR 或稠密矩阵。

- 1: 高效地按行切片。
- 2: 快速地计算矩阵与向量的内积。
- 3: 高效地进行矩阵的算术运行, $CSR + CSR$ 、 $CSR * CSR$ 等。

- 1: 按列切片很慢。
- 2: 一旦构建完成后, 再往里面添加或删除元素成本很高
- 3: CSR 格式在存储稀疏矩阵时非零元素平均使用的字节数(Bytes per Nonzero Entry)最为稳定(float 类型约为 8.5, double 类型约为 12.5)。CSR 格式常用于读入数据后进行稀疏矩阵计算。

考虑到我们存储的形式，首先我们构建后无需再添加或删除元素，并且向量没有表现出很规律的可对角化的特征，所以我们旋转 CSR 进行存储。需要注意的地方 `csc_matrix.multiply(X,Y)`为矩阵对应元素相乘，而 `csc_matrix.dot(X,Y)`不是点乘而是矩阵相乘。

2.3 利用 gensim 库生成词向量

```
#利用gensim生成词向量
from gensim import corpora, similarities
dictionary = corpora.Dictionary([wordbags])
#TypeError: doc2bow expects an array of unicode tokens on input,
# not a single string -->需要unicode编码, wordbags外面加【】变成list
dictionary.doc2bow(documents_st[0])
print(dictionary.token2id)
corpus = [dictionary.doc2bow(doc) for doc in documents_st]
print(len(corpus))
```

Gensim 库首先可以利用 `dictionary` 构建词典, `doc2bow` 方法生成的是编码后的词向量, 该向量具体是以元组列表的形式给出的。例如(1,23)元组中第一个位置为该词汇的位置, 第二个位置是该词汇在 `dictionary` 中的 `id`。通过阅读源码可以看出 `Token2id` 方法和 `id2token` 方法分别是词和编码的相互映射, `gensim` 包中有直接计算 `tfidf` 值的方法, 也可以使用 `word2vec` 模型利用神经网络 `skip-gram` 或者 `CBOW` 模型生成不同维度的词向量, 该模型是基于语义进行判断。

2.4 利用 Simhash 生成词向量并计算相似度

`simhash` 是一种局部敏感 `hash`。假定 A、B 具有一定的相似性, 在 `hash` 之后, 仍然能保持这种相似性, 就称之为局部敏感 `hash`。我们首先利用 `TFIDF` 找出文档的关键词, 取得一篇文章关键词集合, 并通过 `hash` 的方法, 把上述得到的关键词集合 `hash` 成一串二进制。通过对比二进制数, 看其相似性来得到两篇文档的相似性。此时计算相似性的时候采用海明距离, 即二进制位数的不同。

- 1.计算文档 `TF-IDF` 值, 取 `TF-IDF` 权重最高的前 20 个词 (`feature`) 和权重 (`weight`)。即一篇文档得到一个长度为 20 的 (`feature: weight`) 的集合。
- 2.对其中的词 (`feature`), 进行普通的哈希之后得到一个 64 为的二进制, 得到长度为 20 的 (`hash : weight`) 的集合。根据 (`hash`) 中相应位置是 1 是 0, 对相应位置取正值 `weight` 和负值 `weight`。
- 3.对 2 中 20 个列表进行列向累加得到一个列表。并对新生成列表中每个值进行判断, 当为负值的时候去 0, 正值取 1。保存到 `numpy` 格式, 方便后续计算。


```

157 def string_hash(source):
158     if source == "":
159         return 0
160     else:
161         x = ord(source[0]) << 7
162         m = 1000003
163         mask = 2 ** 128 - 1
164         for c in source:
165             x = ((x * m) ^ ord(c))
166         x ^= len(source)
167         if x == -1:
168             x = -2
169         x = bin(x).replace('0b', '').zfill(64)[-64:]
170         print(source, x)
171         return str(x)

```

```

12 s_h = []
13
14 for doc in range(len(documents_st)):
15     t = {}
16     for i in documents_st[doc]:
17         t[i] = count_tfidf(i, doc, tflist_1, dict_idf)
18     t_order = sorted(t.items(), key=lambda x: x[1], reverse=True)
19     # print(t_order[:20])
20     keyList = []
21     klen=min(20,len(t_order))
22     for m in range(klen):
23         feature = string_hash(t_order[m][0])
24         weight = t_order[m][1]
25         temp = []
26         for i in feature:
27             if i == '1':
28                 temp.append(weight)
29             else:
30                 temp.append(-weight)
31         keyList.append(temp)
32     list1 = np.sum(np.array(keyList), axis=0)
33     sim_hash = ''
34     if not keyList:
35         sim_hash = '00'
36     for i in list1:
37         if i > 0:
38             sim_hash = sim_hash + '1'
39         else:
40             sim_hash = sim_hash + '0'

```


三、计算向量的相似度

3.1 计算使用 tf-idf 加权的向量相似度

为了两两计算 3147 篇文档的相似度，首先建立 3147*3147 的 Dataframe,为了避免重复计算，可以首先计算出每篇文档向量的模。

3.1 余弦相似度

```

: %%time
import pandas as pd
dict={}
for i in range(len(documents_st)):
    dict[i]=[0]*len(documents_st)
data=pd.DataFrame(dict, dtype='float')
dim=len(documents_st)
dotlist=[]
for i in range(len(documents_st)):
    dotlist.append(scipy.sort((csr_matrix.multiply(csrlist[i],csrlist[i])).sum()))

```

编写余弦相似度公式两两计算文档的向量相似度：计算的总时间为 **1h9min**。(计算时间比较长，在算法时间复杂度已经不能提升的基础下，需要想新的优化方式)。

```
%%time
for i in range(dim):
    print("\r", end="")
    print("Process progress: {}% ".format((i*100)/len(documents_st)), "\█" * (int)((i*100)/len(documents_st) // 2), end="")
    sys.stdout.flush()
    time.sleep(0.01)
    for j in range(dim):
        data[i][j]=csr_matrix.multiply(csrlist[i],csrlist[j]).sum()/(dotlist[i]*dotlist[j])
```

```
Process progress: 99.96822370511599%: ██████████  
Wall time: 1h 9min 47s
```

```
: data.head()
```

	0	1	2	3	4	5	6	7	8	9 ...	3137	3138	3139	3140	3141	
1	0.266844	1.000000	0.066927	0.060077	0.005150	0.238377	0.026339	0.032312	0.001578	0.118101	...	0.013472	0.018287	0.007426	0.002936	0.007303
47	0.525525	0.462329	0.055546	0.071160	0.008736	0.229525	0.046846	0.040431	0.006851	0.213928	...	0.003068	0.005663	0.010279	0.001268	0.008345
2960	0.544460	0.428053	0.060080	0.114017	0.018109	0.157164	0.037008	0.021104	0.009774	0.143538	...	0.009058	0.033775	0.004742	0.002029	0.018363
426	0.258891	0.403633	0.043914	0.059283	0.005183	0.176630	0.051211	0.028558	0.002842	0.111166	...	0.006795	0.008647	0.008563	0.000000	0.014014
1513	0.155869	0.361963	0.071511	0.059089	0.000636	0.177694	0.040554	0.053625	0.003141	0.086661	...	0.007653	0.023712	0.006642	0.000962	0.005969
5 rows × 3147 columns																

对矩阵按文章 1 相似度进行排序，打印与文章 1 相似度最高的两篇文章：

```
print(documents_pri[1])
```

在十五大精神指引下胜利前进——元旦献辞我们即将以丰收的喜悦送走牛年，以昂扬的斗志迎来虎年。我们伟大祖国在新的一年里，将是充满生机、充满希望的一年。刚刚过去的一年，大气磅礴，波澜壮阔。在这一年，以江泽民同志为核心的党中央，继承邓小平同志的遗志，高举邓小平理论的伟大旗帜，领导全党和全国各族人民坚定不移地沿着建设有中国特色社会主义道路阔步前进，写下了改革开放和社会主义现代化建设的辉煌篇章。顺利地恢复对香港行使主权，胜利地召开党的第十五次全国代表大会——两件大事办得圆满成功。国民经济稳中求进，国家经济实力进一步增强，人民生活继续改善，对外经济技术交流日益扩大。在国际金融危机的风浪波及许多国家的情况下，我国保持了金融形势和整个经济形势的稳定发展。社会主义精神文明建设和民主法制建设取得新的成绩，各项社会事业全面进步。外交工作取得可喜的突破，我国的国际地位和国际威望进一步提高。实践使亿万人民对邓小平理论更加信仰，对以江泽民同志为核心的党中央更加信赖，对伟大祖国的光耀前景更加充满信心。1998年，是全面贯彻落实党的十五大提出的任务的第一年，各条战线改革和发展的任务都十分繁重，有许多深层次的矛盾和问题有待克服和解决，特别是国有企业改革已经进入攻坚阶段。我们必须进一步深入学习和掌握党的十五大精神，统揽全局，精心部署，狠抓落实，团结一致，艰苦奋斗，开拓前进，为夺取今年改革开放和社会主义现代化建设的新胜利而奋斗。今年是党的十一届三中全会召开20周年，是我们党和国家实现伟大的历史转折、进入改革开放历史新时期的20周年。在新的一年里，大力发扬十一届三中全会以来我们党所恢复的优良传统和在新的历史条件下形成的优良作风，对于完成好今年的各项任务具有十分重要的意义。我们要更好地坚持解放思想、实事求是的思想路线。解放思想、实事求是，是邓小平理论的精髓。实践证明，只有解放思想、实事求是，才能冲破各种不切合实际的或者过时的观念的束缚，真正做到尊重、认识和掌握客观规律，勇于突破，勇于创新，不断开创社会主义现代化建设的新局面。党的十五大是我们党解放思想、实事求是的新的里程碑。进一步认真学习 and 掌握十五大精神，解放思想、实事求是，我们的各项事业就能结出更加丰硕的成果。我们要更好地坚持以经济建设为中心。各项工作必须以经济建设为中心，是邓小平理论的基本观点，是党的基本路线的核心内容，近20年来的实践证明，坚持这个中心，是完全正确的。今后，我们能否把建设有中国特色社会主义伟大事业全面推向21世纪，关键仍然要看能否把经济工作搞上去。各级领导干部要切实把精力集中到贯彻落实好中央关于今年经济工作的总体要求和各项重要任务上来，不断提高领导经济建设的能力和水平。我们要更好地坚持“两手抓、两手都要硬”的方针。在坚持以经济建设为中心的同时，积极推进社会主义精神文明建设和民主法制建设，是建设富强、民主、文明的社会主义现代化国家的重要内容。实践证明，经济建设顺利地进行，离不开精神文明建设和民主法制建设的保证。党的十五大依据邓小平理论和党的基本路线提出的党在社会主义初级阶段经济、政治、文化的基本纲领，为“两手抓、两手都要硬”提供了新的理论根据，提出了更高要求，现在的关键是认真抓好落实。我们要更好地发扬求真务实、密切联系群众的作风。这是把党的方针、政策落到实处，使改革和建设取得胜利的重要保证。在当前改革进一步深化，经济不断发展，同时又出现一些新情况、新问题和新困难的形势下，更要发扬这样的好作风。要尊重群众的意愿，重视群众的首创精神，关心群众的生活疾苦。江泽民同志最近强调指出，要大力倡导说实话、办实事、鼓实劲、讲实效的作风，坚决制止追求表面文章，搞花架子等形式主义，坚决杜绝脱离群众、脱离实际、浮躁虚夸等官僚主义。这是非常重要的。因此，各级领导干部务必牢记全心全意为人民服务的宗旨，在勤政廉政、艰苦奋斗方面以身作则，当好表率。1998年，瞩目的中华。新的机遇和挑战，催人进取；新的目标和征途，催人奋发。英雄的中国人民在以江泽民同志为核心的党中央坚强领导和党的十五大精神指引下，更高地举起邓小平理论的伟大旗帜，团结一致，扎实工作，奋勇前进，一定能够创造出更加辉煌的业绩！

2]: print(documents_pri[47])

在全国政协新年茶话会上的讲话（一九九八年一月一日）（附图片1张）江泽民江泽民在茶话会上发表重要讲话。（新华社记者王新庆摄）同志们、朋友们，在这辞旧迎新的喜庆时刻，我代表中共中央、国务院、中央军委，向各民主党派、全国工商联和无党派爱国人士，向全国广大工人、农民、知识分子和干部，向人民解放军指战员、武警官兵、公安干警，向香港特别行政区同胞、澳门同胞、台湾同胞和海外侨胞，向关心和帮助中国现代化建设的国际友人，表示良好的祝愿！祝大家新年好！一九九七年，是我们党和国家历史上非常重要而又极不平凡的一年。年初，敬爱的邓小平同志离开了我们，全党全军全国各族人民紧密团结在党中央周围，毫不动摇地坚持党的基本理论和基本路线，把建设有中国特色社会主义事业继续推向前进。我国政府顺利恢复对香港行使主权，保持了香港的繁荣稳定，在毫无祖国统一大业的道路上迈出了重要的一步。我们胜利召开党的十五大，高举邓小平理论伟大旗帜，总结历史，展望未来，制定了党在社会主义初级阶段的基本纲领，对改革开放和现代化建设跨世纪发展作出了全面部署。这两件大事，极大地鼓舞全党和全国各族人民更加紧密地团结起来，为把我国建成富强民主文明的社会主义现代化国家，完成祖国统一大业而奋斗。一九九七年，我国改革开放和现代化建设继续全面推进。国民经济实现了“高增长、低通胀”，主要经济指标基本达到宏观调控的预期目标，保持着良好的发展态势。农业生产再次获得好收成，企业改革继续深化，结构调整步伐加大，财政收入增长较快，金融形势稳定。城乡人民生活进一步改善。对外经济技术交流与合作继续扩大，国际收支状况良好。总的来看，去年的经济形势是好的，“九五”计划的良好开局得到巩固和发展。党的建设、民主法制建设和精神文明建设取得显著成就。人民解放军革命化、现代化、正规化建设得到进一步加强。我们伟大的祖国，保持社会政治、经济、文化协调发展和全面进步的兴盛局面。一九九七年，我国外交工作取得了重要成果。我国同周边国家的睦邻友好关系继续加强，同广大发展中国家的团结合作进一步巩固，同西方发达国家的关系得到改善和发展。通过高层互访，我国与美、俄、法、日等国确立了面向二十一世纪发展双边关系的目标和指导方针。今天，我国已同南非共和国实现关系正常化，正式建立外交关系。我国还积极参与了区域性、洲际性经济贸易和技术合作与交流。我国的国际地位和国际影响进一步提高。国际舆论普遍认为，中国在促进世界和地区的和平、稳定与发展中发挥着日益重要的作用。一九九八年，是全面贯彻落实十五大提出的各项任务的第一年，也是完成“九五”计划的关键一年。我们要高举邓小平理论伟大旗帜，以党的十五大精神为指导，统揽全局，精心部署，团结一致，艰苦奋斗，开拓前进。要继续贯彻稳中求进的方针，抓住影响经济工作的关键环节，全面推进改革开放和经济建设的各项工作；加强农业基础地位，加快国有企业为重点的各项改革，积极调整和优化经济结构，进一步扩大对外开放，继续推进经济体制和增长方式的根本转变，保持国民经济持续快速健康发展；加强民主法制建设，推进依法治国，抓紧进行机构改革；加强精神文明建设，促进教育科学文化事业发展和社会全面进步；加强党的建设，坚持不懈地开展反腐败斗争。当前，特别要注意组织和调动各方面的力量，切实安排好群众的工作和生活，维护城乡社会稳定。在“一国两制”、“港人治港”、高度自治的方针指导下，继续保持香港的繁荣稳定。要切实做好澳门回归的各项准备工作。一九九八年，将召开第九届全国人民代表大会第一次会议和中国人民政治协商会议第九届全国委员会第一次会议，地方人大和政协也要换届。这是我国政治生活中的大事。各级党委和政府要加强领导，精心组织，做好工作，进一步坚持和完善人民代表大会制度、共产党领导的多党合作和政治协商制度。改革开放近二十年来，我们取得了举世瞩目的成就，这为我们的事业取得新的胜利奠定了坚实的基础。但也要清醒地看到，国际竞争日益激烈，在我们这样一个有十二亿人口的国家中进行现代化建设，任重而道远。要承担和完成艰巨繁重的改革和建设任务，各级干部特别是领导干部，一定要坚持学习马列主义、毛泽东思想特别是邓小平理论，并在实践中创造性地加以运用，进一步解放思想，实事求是，抓住机遇，开拓进取；一定要努力吸收新知识，研究新情况，思考新问题，善于把中央的路线方针政策同本地本单位的实际紧密结合起来，依靠人民群众，脚踏实地、扎扎实实地工作。这样，我们就一定能克服前进中的困难，把改革和发展的各项事业不断推向前进。实现祖国完全统一，是海内外一切爱国的中华儿女的共同心愿。澳门将于一九九九年十二月回到祖国怀抱。此时此刻，我们更加思念台湾同胞。我们将继续坚持“和平统一、一国两制”的基本方针和发展两岸关系、推进祖国和平统一进程的八项主张，大力发展两岸经济、科技、文化领域的交流与合作，推动实现两岸直接“三通”。我们希望台湾当局以民族大义为重，尽早回应我们提出的在一个中国的原则下两岸进行谈判的郑重呼吁。我们将坚持独立自主的和平外交政策，继续在和平共处五项原则基础上发展同各国的友好合作关系，同世界各国人民共同努力，为推动建立公正合理的国际新秩序，为促进世界和平与发展的崇高事业和开创人类美好的未来，作出积极的贡献。人民政协在过去的一年中，认真履行自己的职能，积极反映社情民意，为促进国家的改革和建设提供了许多好的建议和意见。在新的一年里，希望人民政协继续发挥爱国统一战线组织的作用，继续推进政治协商、民主监督、参政议政的规范化和制度化，使之成为我们党团结各界的重要渠道。要继续发挥各民主党派、人民团体和各爱国人士在政协中的作用，为促进改革开放和现代化建设，实现中华民族振兴和完成祖国统一，不断贡献自己的智慧和力量。同志们、朋友们，让我们在邓小平理论的指引下，更加紧密地团结起来，为把建设有中国特色社会主义事业全面推向二十一世纪而努力奋斗！（新华社北京1月1日电）

: print(documents_pri[2960])

在春节团拜会上的讲话（附图片1张）李鹏（1998年1月27日）李鹏总理在春节团拜会上讲话。（新华社记者兰红光摄）朋友们，同志们，五谷丰登的牛年即将过去，生机勃勃的虎年就要来临，全国人民沉浸在节日的喜悦之中。在这辞旧迎新之际，我们欢聚一堂，共庆传统的新春佳节。1997年是我国历史上非常重要而又极不平凡的一年。全国各族人民继承邓小平同志的遗志，把建设有中国特色社会主义事业继续推向前进。香港顺利回到祖国怀抱，并保持繁荣稳定。这是中华民族的大事，也是全世界瞩目的大事，中国人民为此感到无比自豪。中国共产党成功地召开了第十五次全国代表大会，高举邓小平理论伟大旗帜，总结百年历史，提出跨世纪的纲领，将指引着中国人民在社会主义现代化的大道上，不断夺取新的胜利。在过去的一年里，我国改革开放和各项建设取得新的成就。国民经济稳中求进，实现了高增长、低通胀，主要经济指标达到宏观调控的预期目标，保持着良好的发展态势。农业在连续几年丰收的基础上又获得好收成。国有企业改革继续深化。经济结构调整步伐加快。重点工程建设进展顺利。对外经济技术交流日益扩大。在东南亚金融危机波及许多国家的情况下，我国保持了金融形势和整个经济形势的稳定。城乡市场丰富多彩，人民生活进一步改善。社会主义精神文明建设和民主法制建设取得新成绩，各项社会事业蓬勃发展。全国政通人和，民族团结，社会稳定，百业兴旺。各族人民对伟大祖国的美好未来充满信心和希望。1998年是全面贯彻落实党的十五大精神的第一年，也是实现了伟大历史转折的十一届三中全会召开20周年。我们要坚持解放思想、实事求是的思想路线，继续推进经济体制和经济增长方式的根本转变，加强农业基础地位，加快以国有企业为重点的各项改革，加大经济结构调整力度，提高对外开放水平，促进国民经济持续快速健康发展。继续加强社会主义民主法制建设和精神文明建设，积极发展教育、科技、文化等各项事业，坚持不懈地开展反腐败斗争，巩固团结稳定的政治局面，促进社会全面进步。实现祖国的完全统一和民族振兴，是海峡两岸全体中国人的共同愿望。我们将一如既往地，按照“和平统一、一国两制”的基本方针，努力发展两岸的广泛合作与交流。希望台湾当局以民族大义为重，采取实际行动，促进两岸直接通邮、通航、通商早日实现，并尽早回应我们发出的在一个中国的原则下两岸进行谈判的郑重呼吁。当今世界两极化的格局日益明显。和平与发展仍然是时代的主题。大国关系相互调整，发展中国家总体实力增长，国际经济联系日益密切，科技进入日新月异。所有这些，都对各个国家经济与社会的发展，对世界政治和经济格局的演变，产生着重大而深刻的影响。在新的国际形势下，中国将始终不渝地奉行独立自主的和平外交政策，进一步发展与世界各国的友好合作关系，加强同联合国和其他国际组织的协调与合作。中国永远是维护世界和平与稳定的重要力量。我们愿意与各国政府和人民一道，为建立公正合理的国际政治经济新秩序，促进人类和平与发展的崇高事业而贡献自己的力量。朋友们，同志们！1998年将是虎虎有生气、充满希望的一年。我们相信，在以江泽民同志为核心的党中央领导下，全国人民团结奋斗，一定能够取得改革开放和社会主义现代化建设的新成就！（新华社北京1月27日电）

能够看出相似度还是比较高的。

3.2 对 simhash 计算海明距离

```
def hamming_distance(x, y):  
    x=int(x,2)  
    y=int(y,2)  
    return bin(x ^ y).count('1')  
  
def comp_doc(x, y, s_h):  
    return hamming_distance(s_h[x], s_h[y])
```

```
1 import numpy as np  
2 from simhash import hamming_distance, comp_doc  
3 if __name__ == '__main__':  
4     mat_list = np.load("E:/pythonwork/result_hash.npy")  
5     m = []  
6     for i in mat_list:  
7         m.append(str(i))  
8     a = m[1]  
9     sim = {}  
10    for i in range(len(m)):  
11        sim[i] = (comp_doc(1, i, m))  
12    t_order = sorted(sim.items(), key=lambda x: x[1], reverse=False)  
13    print(t_order)
```

```
hashsimilarity ×  
E:\python\python.exe C:/Users/lmh/PycharmProjects/pythonProject/hashsimilarity.py  
[(1, 0), (1353, 14), (2618, 14), (47, 17), (825, 17), (84, 18), (841, 18), (1292, 1  
  
Process finished with exit code 0
```

以第一篇文章为例，查看相似度：

	documents_pri[1]	
		’ 在十五大精神指引下胜利前进——元旦献辞我们即将以丰收的喜悦送走牛年，以昂扬的斗志迎来虎年。我们伟大祖国在新的一年里，将是充满生机、充满希望的一年。刚刚过去的一年，大气磅礴，波澜壮阔。在这一年，以江泽民同志为核心的党中央，继承邓小平同志的遗志，高举邓小平理论的伟大旗帜，领导全党和全国各族人民坚定不移地沿着建设有中国特色社会主义道路阔步前进，写下了改革开放和社会主义现代化建设的辉煌篇章。顺利地恢复对香港行使主权，胜利地召开党的第十五次全国代表大会——两件大事办得圆满成功。国民经济稳中求进，国家经济实力进一步增强，人民生活继续改善，对外经济技术交流日益扩大。在国际金融危机的风浪波及许多国家的情况下，我国保持了金融形势和整个经济形势的稳定发展。社会主义精神文明建设和民主法制建设取得新的成绩，各项社会事业全面进步。外交工作取得可喜的突破，我国的国际地位和国际威望进一步提高。实践使亿万人民对邓小平理论更加信仰，对以江泽民同志为核心的党中央更加信赖，对伟大祖国的光辉前景更加充满信心。1998年，是全面贯彻落实党的十五大提出的任务的第一年，各条战线改革和发展的任务都十分繁重，有许多深层次的矛盾和问题有待克服和解决，特别是国有企业改革已经进入攻坚阶段。我们必须进一步深入学习和掌握党的十五大精神，统揽全局，精心部署，狠抓落实，团结一致，艰苦奋斗，开拓前进，为夺取今年改革开放和社会主义现代化建设的新胜利而奋斗。今年是党的十一届三中全会召开20周年，是我们党和国家实现伟大的历史转折、进入改革开放历史新时期的20周年。在新的一年里，大力发扬十一届三中全会以来我们党所恢复的优良传统和在新的历史条件下形成的优良作风，对于完成好今年的各项任务具有十分重要的意义。我们要更好地坚持解放思想、实事求是的思想路线。解放思想、实事求是，是邓小平理论的精髓。实践证明，只有解放思想、实事求是，才能冲破各种不切合实际的或者过时的观念的束缚，真正做到尊重、认识和掌握客观规律，勇于突破，勇于创新，不断开创社会主义现代化建设的新局面。党的十五大是我们党解放思想、实事求是的新的里程碑。进一步认真学习和掌握十五大精神，解放思想、实事求是，我们的各项事业就能结出更加丰硕的成果。我们要更好地坚持以经济建设为中心。各项工作必须以经济建设为中心，是邓小平理论的基本观点，是党的基本路线的核心内容，近20年来的实践证明，坚持这个中心，是完全正确的。今后，我们能否把建设有中国特色社会主义宏伟事业全面推向21世纪，关键仍然要看能否把经济工作搞上去。各级领导干部要切实把精力集中到贯彻落实好中央关于今年经济工作的总体要求 and 各项任务上，不断提高领导经济建设的能力和水平。我们要更好地坚持“两手抓、两手都要硬”的方针。在坚持以经济建设为中心的同时，积极推进社会主义精神文明建设和民主法制建设，是建设富强、民主、文明的社会主义现代化国家的重要内容。实践证明，经济建设的顺利进行，离不开精神文明建设和民主法制建设的保证。党的十五大依据邓小平理论和党的基本路线提出的党在社会主义初级阶段经济、政治、文化的基本纲领，为“两手抓、两手都要硬”提供了新的理论根据，提出了更高要求，现在的关键是认真抓好落实。我们要更好地发扬求真务实、密切联系群众的作风。这是把党的方针、政策落到实处，使改革和建设取得胜利的重要保证。在当前改革进一步深化，经济不断发展，同时又出现一些新情况、新问题和新困难的形势下，更要发扬这样的好作风。要尊重群众的意愿，重视群众的首创精神，关心群众的生活疾苦。江泽民同志最近强调指出，要大力倡导说实话、办实事、鼓实劲、讲实效的作风，坚决制止追求表面文章，搞花架子等形式主义，坚决杜绝脱离群众、脱离实际、浮躁虚夸等官僚主义。这是非常重要的。因此，各级领导干部务必牢记全心全意为人民服务的宗旨，在勤政廉政、艰苦奋斗方面以身作则，当好表率。1998，瞩目中华，新的机遇和挑战，催人进取，新的目标和征途，催人奋发。英雄的中国人民在以江泽民同志为核心的党中央坚强领导和党的十五大精神指引下，更高地举起邓小平理论的伟大旗帜，团结一致，扎实工作，奋勇前进，一定能够创造出更加辉煌的业绩！’
	documents_pri[1353]	
		’ 致读者伴随着香港回归祖国的福音，伴随着党的十五大胜利召开的鼓点，我们与广大读者一起匆匆却充实地走过了一九九七年。如果说过去的一年我们还算取得了一点成绩的话，那要归功于来自方方面面的支持和帮助，同时，这也将进一步激发我们的工作热情，在新的一年里尽力将我们的工作做得更好。今天读者朋友们看到的是我们新年的第一块专版，也许细心的读者朋友已注意到，我们的专版增加了新的内容，那就是『侨』。中国的发展和富强，离不开广大海外同胞的关心帮助，离不开广大归侨、侨眷的积极参与。中华民族的统一和振兴，需要海内外炎黄子孙的携手奋斗。因此，在版面上充分反映『侨』，我们责无旁贷。我们想此举定会获得广大读者朋友的理解与支持，并且读者朋友将一如既往地同我们携起手来，反映侨情，表达侨声，为侨服务。最后，祝愿广大读者朋友在新的一年里百福并臻、千祥云集。’
	documents_pri[2618]	
		’ 闹市熙攘排队人本报驻科威特瓦记者杨贵兰自觉排队、谦恭礼让，已成为西非科威特瓦经济首都阿比让市人们社会交往中的一种时尚。阿比让私车拥有者虽不少，但对大多数工薪阶层的人来说，上下班仍然主要依靠公共交通工具。每天下午五六点的下班高峰期，公共汽车站上时常排成长蛇阵，尤其在首发站，每队三五十人的几条长龙静静地排列在那里，恭候着汽车的到来。每当汽车来到时，几条长队便有条不紊地同时向着汽车收缩。在那里，虽然没有专门人员来维持秩序，但登车的秩序始终良好，既没有加塞儿现象，也看不到蜂拥而上的场面，更没有车到后最后来的人跑到车门去挤的镜头。在某些较小的汽车站上，一个牌子上有时写着好几个号码，这显然是说，有好几路车在此停靠。在这样的车站，排队上车是不容易的，因为你不知道乘哪路车，但每当一辆汽车进站时，后到者总是让先来者登车，极少发生争抢、推挤的现象。在水、电、电话等费用缴纳处，在超市的收款台前，也常见人们都自觉地排队等候。在高楼大厦的电梯门口，人们更是排着长队。当电梯门口打开、里面的人走出后，等候的人才按顺序迅速跨进电梯内，进不去的便继续在电梯门口静候，他们既不会因排队出口出怨言，更不会去“捷足先登”。记者常常到非洲其他国家采访，阿比让人这种文明礼让、自觉遵守公共秩序的现象也是随处可见，给人留下深刻的印象。（本报阿比让电）’
	documents_pri[47]	
		’ 在全国政协新年茶话会上的讲话（一九九八年一月一日）（附图片1张）江泽民江泽民在茶话会上发表重要讲话。（新华社记者王新庆摄）同志们、朋友们，在这辞旧迎新的喜庆时刻，我代表中共中央、国务院、中央军委，向各民主党派、全国工商联和无党派爱国人士，向全国广大工人、农民、知识分子和干部，向人民解放军指战员、武警官兵、公安干警，向香港特别行政区同胞、澳门同胞、台湾同胞和海外侨胞，向关心和帮助中国现代化建设的国际友人，表示良好的祝愿！祝大家新年好！一九九七年，是我们党和国家历史上非常重要而又极不平凡的一年。年初，敬爱的邓小平同志离开了我们，全党全军全国各族人民紧密团结在党中央周围，毫不动摇地坚持党的基本理论和基本路线，把建设有中国特色社会主义事业继续推向前进。我国政府顺利恢复对香港行使主权，保持了香港的繁荣稳定，在完成祖国统一大业的道路上迈出了重要的一步。我们胜利召开党的十五大，高举邓小平理论伟大旗帜，总结历史，展望未来，制定了党在社会主义初级阶段的基本纲领，对改革开放和现代化建设跨世纪发展作出了全面部署。这两件大事，极大地鼓舞全党 and 全国各族人民更加紧密地团结起来，为把我国建成富强民主文明的社会主义现代化国家，完成祖国统一大业而奋斗。一九九七年，我国改革开放和现代化建设继续全面推进。国民经济实现了“高增长，低通胀”，主要经济指标基本达到宏观调控的预期目标，保持着良好的发展态势。农业生产再次获得丰收成，企业改革继续深化，结构调整步伐加大，财政收入增长较快，金融形势稳定。城乡人民生活进一步改善。对外经济技术交流与合作继续扩大，国际收支状况良好。总的来看，去年的经济形势是好的，“九五”计划的良好开局得到巩固和发展。党的建设、民主法制建设和精神文明建设取得显著成就。人民解放军革命化、现代化、正规化建设得到进一步加强。我们伟大的祖国，保持社会政治、经济、文化协调发展和全面进步的兴盛局面。一九九七年，我国外交工作取得了重要成果。我国同周边国家的睦邻友好关系继续加强，同广大发展中国家的团结合作进一步巩固，同西方发达国家的关系得到改善和发展。通过高层互访，我国与美、俄、法、日等大国确立了面向二十一世纪发展双边关系的目标和指导方针。今天，我国已同南非共和国实现关系正常化，正式建立外交关系。我国还积极参与了区域性、洲性经济贸易和技术合作与交流。我国的国际地位和国际影响进一步提高。国际舆论普遍认为，中国在促进世界和地区的和平、稳定与发展中发挥着日益重要的作用。一九九八年，是全面贯彻落实十五大提出的各项任务的第一年，也是完成“九五”计划的关键一年。我们要高举邓小平理论伟大旗帜，以党的十五大精神为指导，统揽全局，精心部署，狠抓落实，团结一致，艰苦奋斗，开拓前进。要继续贯彻稳中求进的方针，抓住影响经济工作的关键环节，全面推进改革开放和经济建设的各项工作；加强农业基础地位，加快国有企业为重点的各项改革，积极调整和优化经济结构，进一步扩大对外开放，继续推进经济体制和增长方式的根本转变，保持国民经济持续快速健康发展；加强民主法制建设，推进依法治国，抓紧进行机构改革；加强精神文明建
	documents_pri[825]	
		’ 全国清理公款电话收回资金4.84亿元据新华社北京1月8日电记者从中央纪委、监察部了解到，自去年10月15日召开全国纪检监察系统电视电话会议之后，各地各部门清理党政机关公款配置移动电话和住宅电话取得了明显的阶段性成效。据不完全统计，截至去年11月底，31个省市区清理出党政机关公款购买的移动电话4.7万多部，已处理14.7万多部；清理出公款安装的住宅电话67.9万多部，已处理22.7万多部。截至去年12月中旬，中央国家机关有1188个单位上报了清理进展情况，有955个单位清理出公款配置的移动电话近3000部，已处理近百部；清理出公款安装住宅电话2万多部，已处理5000多部。通过处理移动电话和住宅电话，全国已收回资金4.84多亿元。在清理工作中，从整体看各地各部门态度积极，行动迅速，采取了一系列有效的措施。’

可以看出相似度计算的准确率方差很大，对于较短的文本（提取出的关键词以及相同编码较少）并没有很好的识别能力，对于长文本的相似度计算能力较好。

四、利用多进程计算优化计算时间

Python 多进程库 multiprocessing 和 concurrent.future 的官方使用文档如下：

[multiprocessing --- 基于进程的并行 — Python 3.10.0 文档](#)

[concurrent.futures --- 启动并行任务 — Python 3.10.0 文档](#)

由于 python 只有一个 GIL，同一时间只会会有一个获得 GIL 线程在跑，其他线程都处于等待状态。所以并不能利用多线程来使用多核 CPU，所以考虑 python 的多进程。

multiprocessing 是一个支持使用与 threading 模块类似的 API 来产生进程的包。multiprocessing 包同时提供了本地和远程并发操作，通过使用子进程而非线程有效地绕过了全局解释器锁。因为 jupyter notebook 无法进行多进程运行，所以使用 pycharm 进行方法的优化。

1. 对计算 idf 时间进行优化：

首先改写 idf 函数，利用 multiprocessing 建立进程池，我们的电脑实验环境为 8 核 CPU，开启全功率模式进行计算，将原来的数据集等分十份并开启十个新的进程同步进行运算，并将生成的字典以 json 格式存储，方便下次运行时直接加载。

其中 pool 的 map 函数可以直接通过传入一个参数和可迭代的对象进行快速的调用，但是缺点为只能传入一个参数，因此选取 apply_async 方法，该方法支持多个参数传入。

```
def countfreq_idf(wordbags, documents, dict_idf):
    for word in wordbags:
        num=0
        for doc in documents:
            if word in doc:
                num+=1
        idf = math.log(len(documents)/(num+1))
        dict_idf[word]=idf
    return dict_idf

# 多进程计算优化计算idf
start = time.clock()
print("开始计算idf=====")
pool = mp.Pool(10)
results = []
for wordset in wordbag_splitlist:
    results.append(pool.apply_async(countfreq_idf2, (wordset, documents_st)))
pool.close() # 关闭进程池，表示不能再往进程池中添加进程，需要在join之前调用
pool.join() # 等待进程池中的所有进程执行完毕
end = time.clock()
print('Running time: %s Seconds' % (end - start))
idf_dict={}
print(len(results))
for i in results:
    idf_dict.update(i.get())
info_json = json.dumps(idf_dict, sort_keys=False, indent=4, separators=(',', ': '))
f = open('E:/pythonwork/idf_dict.json', 'w')
f.write(info_json)
f.close()
# 显示数据类型
print(type(info_json))
```

从下图可以看出，运行的时间由原来的 8min37s 降到了 174s，计算速度有了明显的提升。同理我们将生成词向量的函数也多进程并行运算。

```

E:\python\python.exe C:/Users/lmh/PycharmProjects/pythonProject/similarity.py
词袋中的数量是: 54033
Running time: 24.9129557 Seconds
拼接后文章数: 3147 个
开始计算idf=====
Running time: 174.7432007 Seconds
10

Process finished with exit code 0

```

2. 并行生成词向量

并行计算词向量，改写函数并保存为 npy 格式，方便直接加载，提高运行速度。

```

def build_matrix(num, documents_st, tflist, dict_idf, wordbags):
    matlist=[]
    for index in num:
        m=[count_tfidf(word,index,tflist,dict_idf) if word in documents_st[index] else 0 for word in wordbags]
        sp_mat=np.array(m)
        matlist.append(sp_mat)
    return matlist

start = time.clock()
print("====生成文章向量====")
pool = mp.Pool(4)
results = []
for num in nums:
    results.append(pool.apply_async(build_matrix, (num,documents_st,tflist_1,idf_dict,wordbags)))
pool.close() # 关闭进程池，表示不能再往进程池中添加进程，需要在join之前调用
pool.join() # 等待进程池中的所有进程执行完毕
end = time.clock()
print('Running time: %s Seconds' % (end - start))
mat_list=[]
print(len(results))
for i in results:
    mat_list.extend(i.get())
np.save("E:/pythonwork/number.npy", mat_list)

```

3. 并行计算相似度

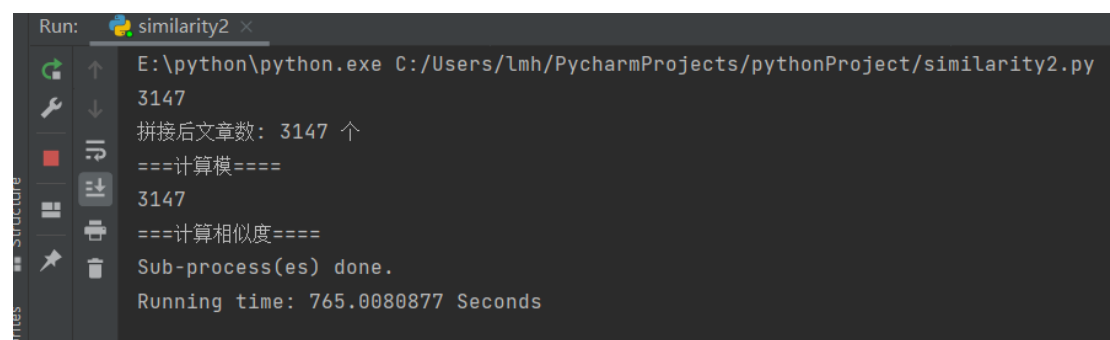
为了能够让所有的进程都具有修改值的权限，我们可以开辟一块资源让所有的进程共享内存。Mp.value 可以将一个值或列表设为共享的内存中的属性，但是该方法中列表只支持一维的列表。所以使用 manager 进行进程的管理，建立共享的字典对象来填入所有文档相似度的结果。

```

print("===计算模===")
for i in range(len(documents_st)):
    dotlist.append(scipy.sqrt((csr_matrix.multiply(csrlist[i], csrlist[i])).sum()))
numset=[x for x in range(len(documents_st))]
print(len(numset))
nums=splitdataset(numset,8)
print("===计算相似度===")
start = time.clock()
with mp.Manager() as manager:
    d = manager.dict()
    pool = mp.Pool()
    for num in nums:
        pool.apply_async(count, (num, dim, csrlist, dotlist, d))
pool.close() # 关闭进程池，表示不能再往进程池中添加进程，需要在join之前调用
pool.join() # 等待进程池中的所有进程执行完毕
print("Sub-process(es) done.")
end = time.clock()
print('Running time: %s Seconds' % (end - start))
info_json2 = json.dumps(dict(d))
f = open('E:/pythonwork/similarity.json', 'w')
f.write(info_json2)

```

程序的运行结果如下:对比之前优化前的 1 小时 9 分的运行时间，并行计算只用了 765 秒，性能提升了五倍以上。证明了多核运算的效率。



```

Run: similarity2 x
E:\python\python.exe C:/Users/lmh/PycharmProjects/pythonProject/similarity2.py
3147
拼接后文章数: 3147 个
===计算模===
3147
===计算相似度===
Sub-process(es) done.
Running time: 765.0080877 Seconds

```

五、总结：

1. 不同的文本表征方法都有最适合自己的相似度度量方法
2. 对于不同的文本算法的适应性不同，以 TF-IDF 算法为例，在现有论文中还有许多变式，例如不同的词性在 tf 值中的权重应有不同。Idf 的标准化还有更合理的构建方法。
3. 在算法时间复杂度不能降低的基础上，多进程并行运算可以显著的提升运算的效率、运算的提升效果取决于机器 CPU 的核数。
4. Simhash 和海明距离计算文本相似度的方法在短文本上没有很好的特征提取能力，效果较差。