VSM-LSI 分析

1. 工作概要

第二次作业是对 VSM 和 LSI 的简单使用，并体会它们的用处和不同，感谢前面几位同学的作业，从而了解到可以使用 gensim 这个开源库，并在此基础上进行了 VSM-LSI 的相似度计算和比较

2. 语料预处理

首先，由于老师给的文档不是连续的，如下图，一篇文档分成了若干部分，并以空格隔开，所以首要任务就是把原来是一个文档的数据放到一起。

```
19980101-01-001-001/m 迈向/v 充满/v 希望/n 的/u 新/a 世纪/n ——/w 一九九八年/t 新年/t 讲话/n （/w 附/v 图片/n 1/r
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/nr
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 国家/n 主席/n 江/nr 泽民/nr 发表/v １９９８年
19980101-01-001-005/m 同胞/n 们/k 、/w 朋友/n 们/k 、/w 女士/n 们/k 、/w 先生/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 高兴/a 地/u 通过/p 中央/n 人民/n 广播/vn 电
19980101-01-001-007/m １９９７年/t ，/w 是/v 中国/ns 发展/vn 历史/n 上/f 非常/d 重要/a 的/u 很/d 不/d 平凡/a 的/u
19980101-01-001-008/m 在/p 这/r 一/m 年/q 中/f ，/w 中国/ns 的/u 改革/vn 开放/vn 和/c 现代化/vn 建设/vn 继续/v 向
19980101-01-001-009/m 在/p 这/r 一/m 年/q 中/f ，/w 中国/ns 的/u 外交/n 工作/vn 取得/v 了/u 重要/a 成果/n 。/w 通
19980101-01-001-010/m １９９８年/t ，/w 中国/ns 人民/n 将/d 满怀信心/l 地/u 开创/v 新/a 的/u 业绩/n 。/w 尽管/c 我
19980101-01-001-011/m 实现/v 祖国/n 的/u 完全/a 统一/vn ，/w 是/v 海内外/s 全体/n 中国/ns 人/n 的/u 共同/b 心愿/n
19980101-01-001-012/m 台湾/ns 是/v 中国/ns 领土/n 不可分割/l 的/u 一/m 部分/n 。/w 完成/v 祖国/n 统一/vn ，/w 是/v
19980101-01-001-013/m 环顾/v 全球/n ，/w 日益/d 密切/a 的/u 世界/n 经济/n 联系/vn ，/w 日新月异/i 的/u 科技/n 进步
19980101-01-001-014/m 中国/ns 政府/n 将/d 继续/v 坚持/v 奉行/v 独立自主/i 的/u 和平/n 外交/n 政策/n ，/w 在/p 和平
19980101-01-001-015/m 在/p 这/r 辞旧迎新/l 的/u 美好/a 时刻/n ，/w 我/r 祝/v 大家/r 新年/t 快乐/a ，/w 家庭/n 幸福/a
19980101-01-001-016/m 谢谢/v ！/w （/w 新华社/nt 北京/ns １２月/t ３１日/t 电/n ）/w

19980101-01-002-001/m 在/p 十五大/j 精神/n 指引/vn 下/f 胜利/vd 前进/v ——/w 元旦/t 献辞/n
19980101-01-002-002/m 我们/r 即将/d 以/p 丰收/vn 的/u 喜悦/an 送/v 走/v 牛年/t ，/w 以/p 昂扬/a 的/u 斗志/n 迎来/v
19980101-01-002-003/m 刚刚/d 过去/v 的/u 一/m 年/q ，/w 大气磅礴/i ，/w 波澜壮阔/i 。/w 在/p 这/r 一/m 年/q ，/w
和/c 国际/n 威望/n 进一步/d 提高/v 。/w 实践/v 使/v 亿万/m 人民/n 对/p 邓小平理论/n 更加/d 信仰/v ，/w 对/p 以/p 江/
19980101-01-002-004/m １９９８年/t ，/w 是/v 全面/ad 贯彻/v 落实/v 党/n 的/u 十五大/j 提出/v 的/u 任务/n 的/u 第一/r
```

即如第一个同学分享的 txt：

```
NO：0 Title:迈向充满希望的新世纪——一九九八年新年讲话（附图片１张）/Author:中共中央总书记、国家主席江泽民/main Body:（一
共同心愿。通过中葡双方的合作和努力，按照“一国两制”方针和澳门《基本法》，１９９９年１２月澳门的回归一定能够顺利实现。台湾

NO：1 Title:在十五大精神指引下胜利前进——元旦献辞/Author: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
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

NO：2 Title:北京举行新年音乐会/Author:江泽民李鹏乔石朱镕基李瑞环刘华清尉健行李岚清与万名首都各界群众和劳动模范代表一起辞旧
留念。李铁映、贾庆林、曾庆红等领导同志也出席了今晚音乐会。

NO：3 Title:李鹏在北京考察企业/Author:向广大职工祝贺新年，对节日坚守岗位的同志们表示慰问/main Body:新华社北京十二月三十一
，当李鹏了解到老闫退休前一直都是厂里的先进工作者、曾经被评为北京市“五好职工”，退休后仍然为改善职工的住房而奔波时，十分高

NO：4 Title:挂起红灯迎新年（图片）/Author:元旦来临，安徽省合肥市长江路悬挂起３３００盏大红灯笼，为节日营造出“千盏灯笼凌空

NO：5 Title:全总致全国各族职工慰问信/Author:勉励广大职工发挥工人阶级主力军作用，为企业改革发展建功立业/main Body:本报北京

NO：6 Title:忠诚的共产主义战士，久经考验的无产阶级革命家刘澜涛同志逝世/Author:（附图片１张）/main Body:新华社北京１２月３

NO：7 Title:党中央国务院关心西藏雪灾救灾工作/Author:灾区各级政府全力组织抗灾力争降低灾害损失/main Body:据新华社北京１２月

NO：8 Title:明天气象预报（１９９８年元月１日２０时—元月２日２０时）/Author:天气趋势分析/main Body:受暖湿气流影响，今天晚
```

不过这里没必要弄成如上的样子，不然还得划分，上面的可以用来对结果比对时分析，下面的代码将相同文章的词进行了合并，得到词袋：

```
"""
函数说明：语料预处理
输入：要打开的文件名
输出：
documents - 处理过后的词空间列表
"""
def load_data(filename ='199801_clear.txt'):
    # 文档分词列表
    documents = []
    #文档分词
    doc = []
    # 打开语料库文件，逐行读取语料，将同一篇文章的不同章节合并，并过滤掉无意义的词和符号
    fr_open = open(filename, 'r', encoding='GBK')

    for line in fr_open:
        if line.strip():   # 跳过空行,这里 strip（）无参数是去掉首尾空格
            words = line.split()
            # 过滤掉无意义的词和符号
            for i in range(len(words)):
                word = words[i]
                if 'w' not in word and 'y' not in word and 'u' not in word \
                        and 'c' not in word and 'm' not in word:
doc.append(word)
        else: #出现空行，即将开始下一篇文档读入，所以把上一篇文档添加进 documents，并将 doc 置为初始
            documents.append(doc)
            doc = []
    return documents
```

3. TF-IDF 权重

再对文本进行重组后就可以统计它的词频了，并根据词袋计算它们的 tfidf 值，这里使

用 gensim 的 corpora 进行词频统计和计算权重，代码注释已经很详细了

```
"""
函数说明：统计并计算 tf_idf 权重值
输入：documents - 处理过后的词空间列表
输出：corpus_tfidf - tfidf 的词向量空间
"""
def tf_idf(documents):
    # 将 documents 放入一个字典,这个字典默认按照一定的顺序排列 documents
    dictionary = corpora.Dictionary(documents)
    #转换成词频，会去掉每个文本的重复词，并自动生成词向量空间
    corpus = [dictionary.doc2bow(text) for text in documents]
    # 将 corpus 作为初始化的参数，得到 tfidf 关于 corpus 的模型
    tfidf = models.TfidfModel(corpus)
    # 通过 tfidf【】将词频转换成 tfidf 值
    corpus_tfidf = tfidf[corpus]
    return corpus_tfidf
```

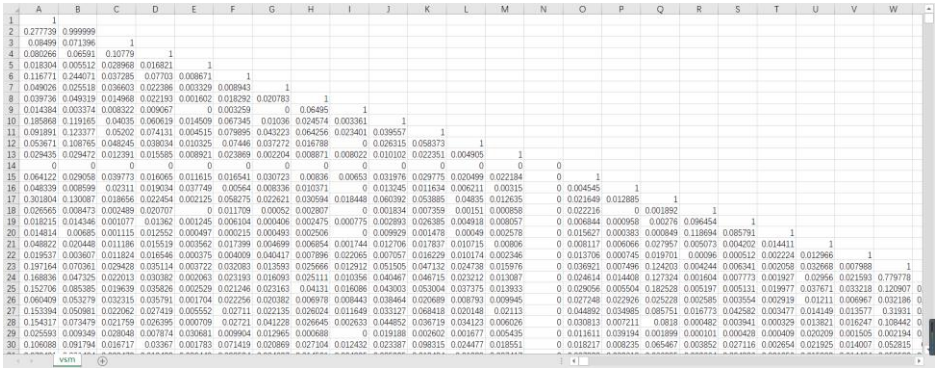
4. VSM 及结果分析

将上一步得到的词向量空间计算它们的相似度，这里不能直接用 similarities.MatrixSimilarity（）函数，会出现内存不够，应该先将其持久化到磁盘，再计算相似度，相当于扩大了虚拟内存。然后再将相似度输出到 csv 文件里面，因为相似度矩阵是一个对称阵，所以我输出了一个下三角型矩阵（节约存储空间），最后可以查找某一个文本和其他文本的相似度，我这里返回前 3 个最大相似度的文本的序号

```
"""
函数说明：进行 LSI 分析，并计算各个文档之间的相似度
输入：corpus_tfidf - tfidf 的词向量空间
输出：CSV 格式的 VSM 下三角相似度矩阵
"""
```

```
def VSM(corpus_tfidf):
    # 必须先把 corpus_tfidf 写入磁盘，因为矩阵太大，直接在内存计算相似度会导致内存不够
    corpora.MmCorpus.serialize('./vsm_test.mm', corpus_tfidf)
    corpus_tfidf = corpora.MmCorpus('./vsm_test.mm')
    # 用文档向量初始化一个相似度计算的对象
    # index = similarities.MatrixSimilarity(corpus_tfidf)
    index = similarities.SparseMatrixSimilarity(corpus_tfidf)
    #输出下三角矩阵
    with open('./vsm.csv', 'w') as fr:
        i = 1
        for sims in index[corpus_tfidf]:
            a = sims[:i]
            fr.write(','.join(map(str, a)) + '\n')
            i += 1
    num = 1 #查找第 num 篇文档的相似度，可自行修改
    query = sorted(enumerate(index[corpus_tfidf[num]]),
key=lambda x: x[1], reverse=True)
    #返回最相似的前三个文档的序号
    print(query[:4])
```

输出的下三角形矩阵：



可以再 excel 里面查找相似度较高的文章，这里我搜寻了大于 0.8 的值，如 101 行 BH 列的值为 0.823，即第 101 篇和第 60 篇的相似度为 0.823，它们是：

NO：59　Title:居民生活如常事业蒸蒸日上香港特区政府发表回归半年大事录/Author:据新华社香港 1 月 1 日电香港特区政府在 1 9 9 7 年的最后一天，发表了《香港特别行政区———首六个月大事录》，认为香港回归后的情况"粉碎了一些对 7 月 1 日后香港变得灰暗的预言"。/main Body:大事录详细记录了香港回归以来政治、经济、法治、人权、教育、基本建设等方面的重大事项。包括：香港成为中国

的一个特别行政区后，在“一国两制”方针政策下运作；基本法保障了香港资本主义制度及原有生活方式五十年不变；发行自己的货币；保持原有税制、独立关税区地位和公务员制度等。这份特区政府的报告长达２０页。文中说，香港顺利回归祖国，香港居民生活如常，正如基本法所保证的一样。报告说，回归两个月后，香港成功举行了世界银行和国际货币基金组织理事会年会，数千名来自各国和地区的财长、央行行长及官员和２０００多传媒人士，在香港得到了第一手的特区运作观感，香港回归后的情况赢得了人们的信心。报告认为，虽然亚洲出现了金融风波，但香港经济仍然跨步向前。报告重申了特区政府维持联系汇率制度的决心，指出香港拥有多种有利因素，包括：比邻近地区通胀率低、香港银行业采取审慎贷款及维持高资本充足比率的政策、租金和楼价下调、联系汇率可稳定汇率等。因此，“假若情况好转，本港的复原速度亦会较快”。报告还记录了行政长官董建华去年９至１１月的外访活动和１０月发表首份施政报告。

NO：100　Title:谱写脱贫致富新篇章——吕梁扶贫开发与计划生育相结合工作纪实/Author:赖林嵩周美林范双民张振明/main Body:山西省吕梁地区是个贫困大区。全区１３个县市中有９个是贫困县，其中国家重点扶持的贫困县有７个。１９７１年建区时，总人口２２６．６万，出生率高达３３．１０‰，死亡率８．４‰，人口自然增长率２４．７０‰。１９７６年开始推行计划生育，无计划生育势头有所遏制。近几年来，全区上下结合实施“八七扶贫攻坚计划”，按“经济要上去，人口要下来”的方针，计划生育工作连年取得好成绩。１９９６年，全区总人口控制在３２４万人，出生率下降到１４．５５‰，自然增长率９．４３‰。贫困人口由１９９２年的１０５万人减少到１９９６年的８０．４万人。吕梁地区的计划生育工作坚持服从、服务于经济建设这个中心，服从、服务于扶贫攻坚这个历史性任务，努力实现计划生育工作的“两个转变”，走出了一条扶贫开发与计划生育相结合的希望之路、成功之路。党政重视，步入人口发展的良性循环轨道吕梁地区坚持不懈、持之以恒地严格控制人口增长。他们采取的主要措施有：———明确一把手负总责。１１年来，一直把计划生育工作作为“一把手工程”来抓，地委、行署每年召开计划生育工作专题会议都在三次以上，每次会议都要切实解决几个实质性问题。每年还要定期召开一把手“交账会议”，既交经济账，又交计划生育账。———加强基层基础工作不松劲。全区各县市都建立了计划生育服务站；２３６个乡镇全部建立了计划生育服务所；４６２６个村配齐了计划生育服务员，人员、经费、待遇得到落实。———确保经费投入逐年增加。地区计划生育经费已由１９９０年前每年２０万元逐年增加到１９９７年的３００万元。县市财政用于计划生育的经费由１９９０年的１１７．６万元（人均０．６０元），增长到１９９６年的５９０万元（人均１．８５元）。穷则思变，走扶贫开发与计生相结合之路吕梁的特点是贫困人口占全省的１／３左右。吕梁的干部群众认识到，扶贫开发与计划生育相结合，是解决贫困地区人口与发展问题的必由之路。１９９０年，地委、行署在总结经验的基础上正式出台《关于扶贫开发与计划生育工作相结合的意见》，在全区实施。结合学习外地经验，吕梁地区扶贫开发与计划生育相结合工作开始形成党政牵头、部门配合、计生协调、群众参与的格局。柳林县下三交镇坪上村原是个有名的老大难村，党支部、村委会班子瘫痪，多年来人均收入徘徊在一二百元之间。１９８６年全村２４２名已婚妇女，有半数多超计划生育。这一年新党支部书记刘步飞上任，他带头只生了一个女孩，拿出自己放电影补助费捐赠给计划生育贫困户，垫支７０００元兴修水利、引水浇地，还带领群众创办了“少生快富甜蜜公司”

看标题 NO：59Title:居民生活如常事业蒸蒸日上香港特区政府发表回归半年大事录和 NO：100　Title:谱写脱贫致富新篇章，知两个都是与人民生活越来越好相关。还是比较相关。

5. LSI 及结果分析

前面的数据处理过程都一样，就是得到 tfidf 权重值。LSI 主要是对矩阵用 SVD 进行了一个降维，根据 SVD，任何 m*n 的矩阵 M，都存在以下 SVD 分解：

$$M_{m\times n} = U_{m\times m}\Sigma_{m\times n}(V_{n\times n})^{\cdot}$$

通过几何意义对其证明如下：

假设存在 M*N 矩阵 A，事实上，A 矩阵将 n 维空间中的向量映射到 k（k<=m）维空间中，k=Rank(A)。现在的目标就是：在 n 维空间中找一组正交基，使得经过 A 变换后还是正交的。假设已经找到这样一组正交基：

$$\{v_1, v_2, \dots, v_n\}$$

则 A 矩阵将这组基映射为：

$$\{Av_1, Av_2, \dots, Av_n\}$$

如果要使他们两两正交，即

$$\mathrm{A}v_i . Av_j = (\mathrm{A}v_i)^T Av_j = v_i{}^T A^T \mathrm{A}v_j = 0$$

根据假设，存在

$$v_i{}^T v_j = v_i \cdot v_j = 0$$

所以如果正交基 v 选择为A^TA的特征向量的话，由于A^TA是对称阵，v 之间两两正交，那么

$$\begin{aligned} v_i{}^T A^T \mathrm{A}v_j &= v_i{}^T \lambda_j v_j \\ &= \lambda_j v_i{}^T v_j \\ &= \lambda_j v_i . v_j = 0 \end{aligned}$$

这样就找到了正交基使其映射后还是正交基了，现在，将映射后的正交基单位化：

因为

$$\mathrm{A}v_i . Av_i = \lambda_i v_i . v_i = \lambda_i$$

所以有

$$|\mathrm{A}v_i|^2 = \lambda_i \geq 0$$

所以取单位向量

$$u_i = \frac{Av_i}{|Av_i|} = \frac{1}{\sqrt{\lambda_i}} Av_i$$

由此可得

$$Av_i = \sigma_i u_i \ ，\sigma_i(\text{奇异值}) = \sqrt{\lambda_i} \ ，0 \le i \le k \ ，k = \text{Rank}(A)$$

当 k < i <= m 时，对 u1，u2，…，uk 进行扩展 u(k+1),…,um，使得 u1，u2，…，um 为 m 维空间中的一组正交基，即

$$\text{将}\{u_1, u_2, \dots, u_k\}\text{正交基扩展成}\{u_1, u_2, \dots, u_m\}R^m\text{空间的单位正交基}$$

同样的，对 v1，v2，…，vk 进行扩展 v(k+1),…,vn（这 n-k 个向量存在于 A 的零空间中，即 Ax=0 的解空间的基），使得 v1，v2，…，vn 为 n 维空间中的一组正交基，即

$$\text{在A的零空间中选取}\{v_{k+1}, v_2, \dots, v_n\}\text{使得}Av_i = 0 \ ，\ i > k \ ，\text{并取}\sigma_i = 0$$

则可得到

$$A[v_1\, v_2\, \cdots\, v_k | v_{k+1}\, \cdots\, v_n] = [u_1\, u_2\, \cdots\, u_k | u_{k+1}\, \cdots\, u_m] \left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_k & \\ \hline & 0 & & 0 \end{array}\right]$$

继而可以得到 A 矩阵的奇异值分解：

$$A = U\Sigma V^T$$

Σ是一个对角矩阵，其中的元素值就是奇异值，并且按照从大到小的顺序排列。在很多情况下，前 10%甚至更少的奇异值的平方和就占了全部奇异值平方和的 90%以上了，因此可以用前 k 个奇异值来近似描述矩阵：

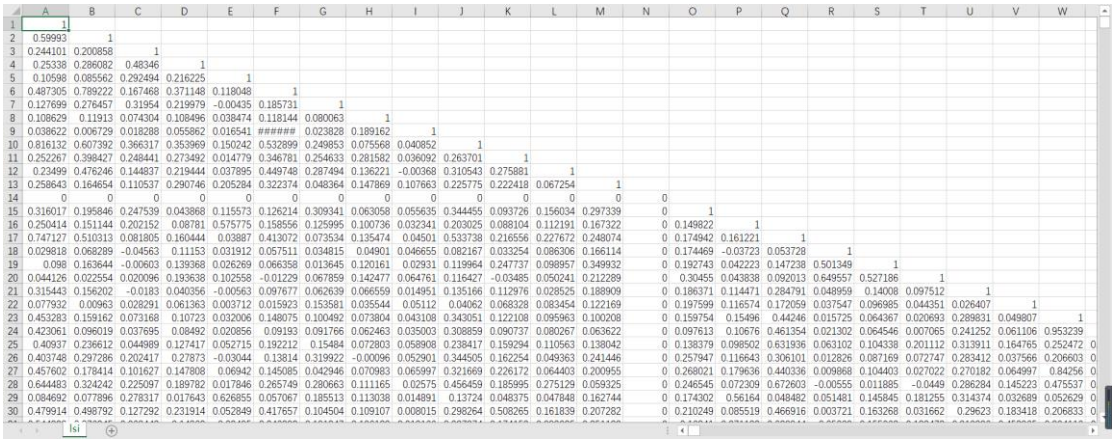
$$M_{m\times n} \approx U_{m\times k}\Sigma_{k\times k}(V_{n\times k})^*$$

使用 models.LsiModel()函数构造 LSI 模型并计算相似度：

```
"""
函数说明：进行 LSI 分析，并计算各个文档之间的相似度
输入：corpus_tfidf - tfidf 的词向量空间
输出：CSV 格式的 LSI 下三角相似度矩阵
```

```
"""
def LSI(corpus_tfidf):
    # 构造 LSI 模型并将待检索的 query 和文本转化为 LSI 主题向量
    # 转换之前的 corpus 是 BOW 向量
    lsi_model = models.LsiModel(corpus_tfidf,num_topics=80)
    corpus_lsi = lsi_model[corpus_tfidf]
    # 转换成潜语义文档向量列表并持久化
    corpora.MmCorpus.serialize('./lsi_test.mm', corpus_lsi)
    corpus_lsi = corpora.MmCorpus('./lsi_test.mm')
    # 构建文档相似度矩阵索引用于查询，再使用文档列表本身进行相似度查询（默认使用 Cosine）
    index = similarities.MatrixSimilarity(corpus_lsi)
    # 输出下三角矩阵
    with open('./lsi.csv', 'w') as fr:
        i = 1
        for sims in index[corpus_lsi]:
            a = sims[:i]
            fr.write(','.join(map(str, a)) + '\n')
            i += 1
```

在降维后，再来计算它们的相似度，可以得到:



下面是从中选取的两篇相似度较高的文档的分析:

NO：0　Title:迈向充满希望的新世纪——一九九八年新年讲话（附图片１张）/Author:中共中央总书记、国家主席江泽民/main Body:

（一九九七年十二月三十一日）１２月３１日，中共中央总书记、国家主席江泽民发表１９９８年新年讲话《迈向充满希望的新世纪》。（新华社记者兰红光摄）同胞们、朋友们、女士们、先生们：在１９９８年来临之际，我十分高兴地通过中央人民广播电台、中国国际广播电台和中央电视台，向全国各族人民，向香港特别行政区同胞、澳门和台湾同胞、海外侨胞，向世界各国的朋友们，致以诚挚的问候和良好的祝愿！１９９７年，是中国发展历史上非常重要的很不平凡的一年。中国人民决心继承邓小平同志的遗志，继续把建设有中国特色社会主义事业推向前进。中国政府顺利恢复对香港行使主权，并按照“一国两制”、“港人治港”、高度自治的方针保持香港的繁荣稳定。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，展望新的世纪，制定了中国跨世纪发展的行动纲领。在这一年中，中国的改革开放和现代化建设继续向前迈进。国民经济保持了“高增长、低通胀”的良好发展态势。农业生产再次获得好的收成，企业改革继续深化，人民生活进一步改善。对外经济技术合作与交流不断扩大。民主法制建设、精神文明

建设和其他各项事业都有新的进展。我们十分关注最近一个时期一些国家和地区发生的金融风波，我们相信通过这些国家和地区的努力以及有关的国际合作，情况会逐步得到缓解。总的来说，中国改革和发展的全局继续保持了稳定。在这一年中，中国的外交工作取得了重要成果。通过高层互访，中国与美国、俄罗斯、法国、日本等大国确定了双方关系未来发展的目标和指导方针。中国与周边国家和广大发展中国家的友好合作进一步加强。中国积极参与亚太经合组织的活动，

NO：47　Title:北国有寒趣南疆闻花香各地群众喜庆元旦/Author:据新华社北京１月１日电１９９８年元旦之际，从寒峭的北国大地，到如春的海南宝岛；从美丽的东海之滨，到莽莽的西部边陲，全国各地各族人民纷纷喜庆佳节，决心在新的一年里，高举邓小平理论的伟大旗帜，在党的十五大精神指引下胜利前进。/main Body:元旦的北京，节日气氛格外浓厚。清晨，来自祖国各地及港澳台地区的４万多人，冒着零下７摄氏度的严寒，自发地聚集到了祖国的心脏———天安门广场，参加了庄严的元旦天安门广场升国旗仪式。在长安街沿线及大街小巷的高大建筑物上，国旗、彩旗飘扬。京城各大公园、商场及各种娱乐场所也都充满了欢乐气氛。巨大的氢气红灯笼和各种颜色的彩球飘荡在冬日京城的天空中。今天，哈尔滨市的最低气温达到零下２５摄氏度，但市民们欢庆新年的热情却没有丝毫减弱。在冰雪覆盖的松花江上，刚刚开放的江上游乐园成了人们游玩的好去处。一座座冰雕构成了另一个崭新的世界，人们穿梭其间，尽情地享受冰雪带来的无穷乐趣。与寒冷的哈尔滨形成鲜明对照，南国春来早，花城今更艳。'９８广州风情贺年缤纷大巡游今天在广州隆重举行，为广州新年第一天增添了喜庆的节日气氛，同时也拉开了'９８华夏城乡游（广州）活动的序幕。来自香港特区１０多个团体的４００多位香港同胞首次来广州参加此次贺年活动。海口市阳光明媚。

这两篇文档一个是喜庆新年，一个喜庆元旦，具有一定的相似度。

6. VSM 和 LSI 比较

首先看运算时间，这里因为前面的数据预处理和，tfidf 权重计算都一样，所以就没加入，只考虑了后面 VSM 和 LSI 处理的计算时间

```
"""
函数说明：进行测试
"""
def test():
    #加载数据并生成 tfidf 词向量空间
    documents = load_data('199801_clear.txt')
    corpus_tfidf = tf_idf(documents)
    #对词向量空间进行 VSM 的相应处理并计时
    VSMstart_time = time.time()
    VSM(corpus_tfidf)
    VSMend_time = time.time()
    print('VSM 耗时：' + str(VSMend_time - VSMstart_time) + 's')
    # 对词向量空间进行 LSI 的相应处理并计时
    LSIstart_time = time.time()
```

```
    LSI(corpus_tfidf)
    LSIend_time = time.time()
    print('LSI 耗时：' + str(LSIend_time - LSIstart_time) + 's')
```

测试结果:

```
VSM耗时：11.93734097480774s
LSI耗时：22.029853582382202s
```

可以看出 LSI 大概比 VSM 多用 10S，基本上是矩阵进行 SVD 分解，和奇异值计算造成的。但原本的向量空间太过于稀疏，有很多的 0，进行降维后压缩了矩阵，也去除了噪音，使得结果更加精确。

并且 VSM 无法处理文本中隐含的潜语言和一次多意，而 LSI 在压缩矩阵时，减轻了此问题，但 SVD 主动舍去了一部分不重要的特征，所以在降维的同时也要承担数据损失的风险。

7. 参考

https://www.cnblogs.com/lzllovesyl/p/5243370.html SVD 推导

https://blog.csdn.net/u011630575/article/details/80159100 SVD 推导

https://blog.csdn.net/zhongkejingwang/article/details/43053513 gensim