随机算法课程

实验报告

实验五：基于 JL 变换的文章聚类

姓名：袁野、陈思莹、詹儒彦

学号：1190200122、1190202327、1190202307

班级：1903102

评分表：（由老师填写）

最终得分：	
对实验题目的理解是否透彻：	
实验步骤是否完整、可信 ：	
代码质量 ：	
实验报告是否规范 ：	
趣味性、难度加分 ：	
特 色：	1
	2
	3

一、实验题目概述

在计算机科学应用中，我们经常会对一些向量进行聚类，而在聚类问题中，我们要处理的向量的维数可能会非常多，高维向量的聚类时会有很多意外的出现，例如在高维空间中，向量的分布会具有稀疏性，并不便于提取其特征，会导致分类效果极差。而且高位空间的向量聚类所需要的时间复杂度通常情况下是难以忍受的。

降维就是一种对高维度特征数据预处理方法。降维是将高维度的数据保留下最重要的一些特征，去除噪声和不重要的特征，从而实现提升数据处理速度的目的。在实际的生产和应用中，降维在一定的信息损失范围内，可以为我们节省大量的时间和成本。降维也成为应用非常广泛的数据预处理方法。

这次实验中，我们用一个词向量代表一篇文章，对于这些向量进行聚类，从而使想有关于文章主题的聚类操作，而词向量的维数可能会十分巨大，因此我们会采用 JOHNSON-LINDENSTRAUSS TRANSFORMATION 的方法对高维度的的词向量进行降维，并且探究不同降维矩阵下的实际效果。

在本次实验中我们的分工为：

袁野：文章数据的选择爬取、实验内容步骤的设计，文献资料的查找。

詹儒彦：两种 JL 变换矩阵的生成，以及对原有词向量的降维部分代码编写。

陈思莹：分词并生成原有词向量，以及 K-Means 部分代码的编写与调试。

二、对实验步骤的详细阐述

1、文章的爬取

首先利用通过抓取网页请求并对网页源代码进行解析的方式爬取文章，分别在中国经济网金融证券频道（http://finance.ce.cn/2015home/jj）、CNMO 网互联网板块（https://internet.cnmo.com/tech/）、中国新闻网娱乐新闻（https://www.chinanews.com.cn/entertainment.shtml）三个网站共爬取了 625 篇文章作为我们聚类的对象。

2、分词操作

利用 python 中的 jieba 库进行**分词**，把爬取得到的文章中的句子划分成词（划分后的形式如下图所示）：

之后，设置一些停用词，我们采用了百度的停用词表，同时，针对我们爬取的数据，又自己加入了一些**停用词**（部分停用词如下图所示）：

将停用词删去之后，我们采用 **TF-IDF 算法**来进行向量生成。

该方法时一种统计方法，用来评估一个字词对于一个文件集或者一个语料库中其中一份文件的重要程度。TF 是文章中的词频、IDF 是逆词频公式如下：

$$W_{x,y} = tf_{x,y} \times log\left(\frac{N}{df_x}\right)$$

其中，$tf_{x,y}$是词x在文章y中的频率；df_x是词x在几篇文章中出现过；N是文章总数；

$$Idf = \log\left(\frac{N}{df}\right)$$

那么，**分词转化为词向量的 TF-IDF 算法流程**可以修改如下：

(1)求所有文本的分词结果的并集$|U|$

(2)对于$word_i \in U, i = 1,2,$计算$w_{word_i,\ text} = tf_{word_i,\ text} * log\left(\frac{N}{df_{word_i}}\right)$，文本 text 对应的向量即$\left(w_{word_1,\ text}, w_{word_2,\ text}, \cdots, w_{word_{|U|},\ text}\right)$。

然后对词向量进行筛选，考察所有对应的词，将df_{word_i}高于 90%或者低于 0.5%对应的 word 从词向量中删去，进行初步地降维。因为出现频率太低或太高的词会影响聚类地效率以及最终地聚类效果。

3、简单介绍 JL 变换：

通俗的说 JL 变换，即只需要$log(n)$维的空间就可以塞下n个向量。而最严谨的 JL 变换矩阵需要向量都是模长为 1 的正交向量，然而现实生活中并不能很好的找到这样的向量。因此我们采用：

马尔可夫不等式：如果x是非负随机变量，$a > 0$，那么

$$P(x \geq a) \leq \frac{\mathrm{E[x]}}{\mathrm{a}}$$

切比雪夫不等式：

$$P((\mathrm{x}-\mathrm{E}[\mathrm{x}])^2 \geq \mathrm{a}^2) \leq \frac{\mathrm{E}[(\mathrm{x}-\mathrm{E}[\mathrm{x}])2]}{\mathrm{a}^2} = \frac{\mathrm{Var}[\mathrm{x}]}{a^2}$$

Cramér-Chernoff 方法：

理论上，我们可以找到使得最右端最小的 λ，以获得最高的估计精度：

$$P(\mathrm{x} \geq \mathrm{a}) \leq \min_{\lambda>0} \mathrm{e}^{(-\lambda \mathrm{a})} E\left[\mathrm{e}^{\lambda \mathrm{x}}\right]$$

通过使用 Cramer-Chernoff 法则，可以得到**单位模引理**：设$u \in \mathrm{R}^{\mathrm{n}}$是独立重复采样自$N\left(0,\frac{1}{\mathrm{n}}\right)$的向量，$\varepsilon \in (0,1)$是给定常数，那么我们有

$$P(|\| \mathrm{u} \|^2 - 1| \geq \varepsilon) \leq 2exp\left(-\frac{\varepsilon^2 \mathrm{n}}{8}\right)$$

由单位模原理可以得到：当n足够大的时候，u的模长明显偏离 1 的概率是非常小的（给定ε后，将以n的指数形式递减至 0），所以从$N(0,1/n)$采样出来的n维向量将会非常接近单位向量。因此，我们可以使用**随机采样的高斯矩阵**作为 JL 的变换矩阵。

利用以上工具我们可以推出：

JL 引理：令$\in (0,1)$，给定R^{d}中的包含 n 个点的点集P，$k = \frac{1}{\varepsilon^2}$（与 d 无关），那么存在一个从$\mathrm{R}^{\mathrm{d}}$到$\mathrm{R}^{k}$的 **Lipshcitz** 映射 f，对于$\forall u, v \in P$有：

$$(1-\varepsilon) \| u-v \|^2 \leq \| f(\mathrm{u}) - f(\mathrm{v}) \|^2 \leq (1+\varepsilon) \| u-v \|^2 \quad (公式 1)$$

其中，lipshcitz 映射 f 为：构造一个矩阵$B \in \mathrm{R}^{\mathrm{k}\times\mathrm{d}}$，得到$f(u) = Bu$。映射的构造方法：$B = \frac{1}{\sqrt{k}}A$，其中$A$是一个$k \times d$维矩阵，且它的每一个元素均独立同分布于高斯分布 $N(0,1)$。

而根据 Matousek Theorem 与 Hoeffding's inequality 可以得到较为稀疏的两种亚高斯变换矩阵：

$$\mathrm{A}_{\mathrm{ij}} = \begin{cases} -1\ with\ prob\frac{1}{6} \\ 0\ with\ prob\frac{2}{3} \\ 1\ with\ prob\frac{1}{6} \end{cases} \quad \mathrm{A}_{\mathrm{ij}} = \begin{cases} -1\ with\ prob\frac{1}{2} \\ 1\ with\ prob\frac{1}{2} \end{cases}$$

我们采用第一种来进行设计（binaryJL）

值得注意的一点是，这里给出的k是一个"最坏的"下界，以保持X中任意两点距离。同时，我们可以验证一下k的大小：

在$\varepsilon = 0.1, n = 15$时，$k \geq 2321$

在$\varepsilon = 0.5, n = 15$时，$k \geq 130$

在$\varepsilon = 0.8, n = 15$时，$k \geq 73$

在$\varepsilon = 0.1, n = 2$时，$k \geq 595$

可以发现，这与我们想要的降维效果相去甚远，甚至两个向量降维的维度都需要五百多维，实际上，JL 引理仅仅是给出了一个充分条件，一个最坏的情况，即**“**k在满足这个下界的时候我们一定能找到这样的保距映射**”**，但k不满足下界的时候，我们并不能得到任何关于这个映射是否存在的信息。

JL 引理的结论中之所以能够出现$log\ N$，本质上是因为**“**单位模引理**”**中的概率项$2exp\left(-\frac{\varepsilon^2 n}{8}\right)$是指数衰减的，而我们可以放宽这个衰减速度，让其变成多项式衰减，从而出现了$log\ N$。

总的来说，JL 引理告诉我们，以误差ε塞下N个向量，只需要$O\left(\frac{log\ N}{\varepsilon^2}\right)$维的空间，至于$\frac{log\ N}{\varepsilon^2}$前面的常数是多少，其实不大重要。因为事实上 JL 引理是一个非常充分的条件，实际情况中条件往往更加宽松。比如，在 JL 引理的证明中如果我们将条件改为$n > \frac{16log\ N}{\varepsilon^2}$，那么引理成立的概率就不小于

$$1 - N(N-1)exp\left(-\frac{\varepsilon^2 n}{8}\right) \geq 1 - N(N-1)N^{-2} = \frac{1}{N}$$

注意$\frac{1}{N}$虽然小，但终究是大于0的，所以此时依然是存在A使得公式一成立，只不过寻找A的成本更大罢了（每次命中的概率只有$\frac{1}{N}$），而如果我们只关心存在性，那么这也够了。

而且，JL 引理只考虑了在随机线性投影下的降维，就已经得到$n > \frac{16log\ N}{\varepsilon^2}$了，如果是其他更精细的降维，比如基于 SVD 的降维，是有可能得到更好的结果的（前面的系数更小）；如果非线性的降维方法也考虑进去，那么结果又能变得更优了。所以说，不需要太关心$\frac{log\ N}{\varepsilon^2}$前面的常数是多少，我们只需要知道$O\left(\frac{log\ N}{\varepsilon^2}\right)$的量级，如果真要用到它，通常还需要根据实际情况确定前面的常数，而不是调用理论结果。

具体实验中，我们采用 binaryJL 和 GaussJL 两种投影矩阵的生成方式进行实验对比。设定误差率ϵ为 0.1，设定降维后维数 k 为$\frac{\log(n)}{\epsilon^2}$的 0.5、1、1.5、2、3 倍。然后创建数组从文件中读取并存储的文章的词向量形成的 datamat 及其规格 rows、max_len。利用计算 jl 变换后的维度 k 创建 JL 变换矩阵，主要应

用了 gaussJL 与 binaryJL 两种变换矩阵：

1） GaussJL 主要利用高斯向量在高维中通常正交的性质来设计转换矩阵，矩阵中的每个数从N(0,1)中抽取，并利用系数$\frac{1}{\sqrt{k}}$来进行归一化

2） 亚高斯 JL（binary JL）利用放松高斯矩阵的限制将矩阵稀疏化。矩阵中的每个元素以$\frac{2}{3}$、$\frac{1}{6}$、$\frac{1}{6}$的概率设定为 0,1，-1，以达到稀疏化的目的，使计算简便。

通过将 datamat 中原本的词向量与得到的两种矩阵分别相乘进行降维，得到了 k 维向量，并输出到文件中进行后续聚类。

通过产生 10 次不同的转换矩阵，并生成降维之后的低维矩阵，然后计算所有向量对的欧氏距离与降维之前的误差率平均值并输出。

4、聚类

对于降维之后的向量聚类，这里我们利用 k-Means 算法，对向量结果进行聚类，对比聚类结果与真实结果之间的差距，评估不同降维方法的效果。其中聚类是的距离函数采用的是余弦函数。对于 k-Means 算法内细节的调整，我们将在第五部分提到。

对于聚类效果的评估，我们采用了公式

$$score = \frac{类别划分正确的文章数}{文章数}$$

易知，$score \in [0,1]$且 score 越高，效果越好。

三、实验数据

1．实验设置

实验环境：

袁野：Ubuntu 20.04.4 LTS (GNU/Linux 5.4.0-109-generic x86_64) IDE：PyCharm

陈思莹：Windows11 IDE：PyCharm

詹儒彦：Windows11 IDE：VsCode

数据：

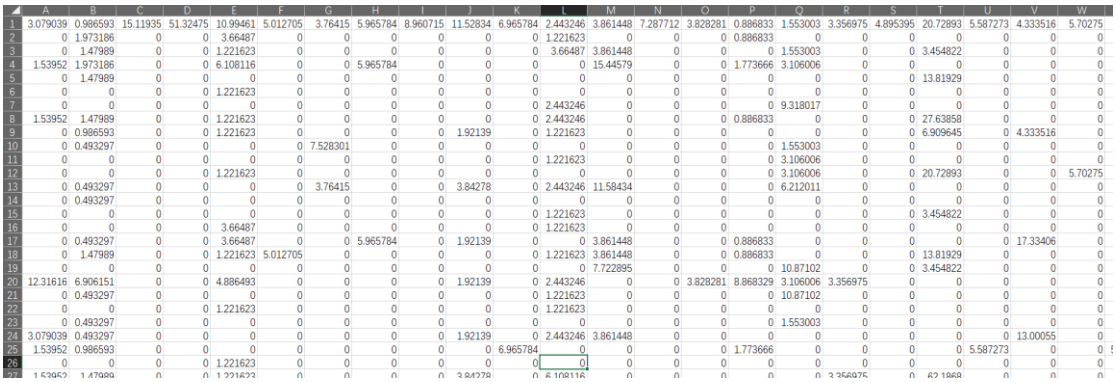
从三个网站爬取下来的文章分别为经济、科技、娱乐三类，这三个类别即为后面聚类是的标准分类，共 625 篇。

```
drwxr-xr-x 2 root root 98304 Jun 14 18:19  ./
drwxr-xr-x 5 root root  4096 Jun 16 17:55  ../
-rw-r--r-- 1 root root  2430 Jun 14 18:08  娱乐0_中国新闻网联合出品，毛不易与你奔赴《小世界》.txt
-rw-r--r-- 1 root root  4730 Jun 14 18:08  娱乐100_北美电影仍在摸索“混搭发行”.txt
-rw-r--r-- 1 root root  6252 Jun 14 18:08  娱乐101_万象为宾客，《经典咏流传》第八期唱响“今朝更好看”的风貌之美.txt
-rw-r--r-- 1 root root 12621 Jun 14 18:08  娱乐102_60岁杨紫琼的多重宇宙.txt
-rw-r--r-- 1 root root  5706 Jun 14 18:08  娱乐103_老炮儿“转场”的变与不变.txt
-rw-r--r-- 1 root root  3672 Jun 14 18:08  娱乐104_《三生三幸》讲述患难爱情令人泪目.txt
-rw-r--r-- 1 root root  2311 Jun 14 18:08  娱乐105_慢综艺展现乡村真实美好生活.txt
-rw-r--r-- 1 root root  5433 Jun 14 18:08  娱乐106_听懂罗大佑的人都老了？那又怎样！.txt
-rw-r--r-- 1 root root  2009 Jun 14 18:08  娱乐107_首场线上演唱会为何叫《童年》？罗大佑独家回应.txt
-rw-r--r-- 1 root root  3334 Jun 14 18:08  娱乐108_谁是“匿名用户”？《奔跑吧》第十季开启“悬疑烧脑”模式！.txt
-rw-r--r-- 1 root root  3605 Jun 14 18:08 '娱乐109_昔日队友再见成敌手 《了不起！舞社》与这街惊喜联动.txt'
-rw-r--r-- 1 root root  3660 Jun 14 18:08  娱乐10_痛都痛了，那就绽放吧.txt
-rw-r--r-- 1 root root  5710 Jun 14 18:08  娱乐110_潘周聃：走红不意外，不拒绝做公众人物.txt
-rw-r--r-- 1 root root  1111 Jun 14 18:08  娱乐111_《警察荣誉》张若昀白鹿演绎实习民警.txt
```

```
-rw-r--r-- 1 root root  1475 Jun 14 17:57  科技106_新发现！原来秦兵马俑是用鸡蛋动物胶成分粘接的.txt
-rw-r--r-- 1 root root  1620 Jun 14 17:57  科技107_苹果手表制成！这就是世界最小的电视？网友表示不服.txt
-rw-r--r-- 1 root root  1379 Jun 14 17:57 '科技108_连续两年超台湾省 2020年大陆半导体全球销售份额达9%.txt'
-rw-r--r-- 1 root root  1221 Jun 14 17:57 '科技109_好转？今年首季整体NAND Flash价格跌幅收敛至8-13%.txt'
-rw-r--r-- 1 root root  1672 Jun 14 17:53 '科技10_吉利卫星公司CEO：和马斯克相比 我们的野心小得多.txt'
-rw-r--r-- 1 root root  1311 Jun 14 17:57  科技110_“纸片人”高阶进化！明星偶像会被虚拟人物取代吗？.txt
-rw-r--r-- 1 root root  8894 Jun 14 17:57 '科技111_触摸屏行业新风向 超声波触控打造人机交互新体验.txt'
-rw-r--r-- 1 root root  1394 Jun 14 17:57  科技112_什么原理？最新研究称全球变暖令新生儿体重快速增长.txt
-rw-r--r-- 1 root root  1070 Jun 14 17:57  科技113_油管首个播放量破百亿视频！《鲨鱼之家》你看过吗？.txt
-rw-r--r-- 1 root root  1191 Jun 14 17:57  科技114_大疆年度总结来袭！累积飞行里程相当于登月93次.txt
-rw-r--r-- 1 root root  1254 Jun 14 17:57 '科技115_字节跳动投资未斯科技 后者涵盖RPA机器人技术等业务.txt'
-rw-r--r-- 1 root root  1327 Jun 14 17:57 '科技116_螺蛳粉都能“闻”出来！铁蛋解锁嗅觉新技能 你心动吗？.txt'
-rw-r--r-- 1 root root  4090 Jun 14 17:57 '科技117_字节跳动研发医疗AI抗癌 可帮助实时筛查结直肠肿瘤.txt'
-rw-r--r-- 1 root root  1250 Jun 14 17:57  科技118_媒体报道：英特尔计划5年内将晶圆厂产能扩增三成.txt
-rw-r--r-- 1 root root  1442 Jun 14 17:57  科技119_花5.6万元就能游太空？日企推出乘气球游太空服务.txt
-rw-r--r-- 1 root root  1406 Jun 14 17:53 '科技11_黑科技上线！世界首例活组织制成3D打印耳朵移植成功 .txt'
```

```
-rw-r--r-- 1 root root  5357 Jun 14 18:19  经济138_基础设施REITs多元化扩容机构动作频频强化布局.txt
-rw-r--r-- 1 root root  2827 Jun 14 18:19  经济139_机构内控和信息安全事件频发监管持续加码.txt
-rw-r--r-- 1 root root  6049 Jun 14 18:19  经济13_外资年内大幅加仓A股银行保险股对10股持股量增超50%.txt
-rw-r--r-- 1 root root  4343 Jun 14 18:19  经济140_ETF纳入互联互通资本市场开放再进一步.txt
-rw-r--r-- 1 root root  9296 Jun 14 18:19  经济141_揭秘雍禾医疗业绩增长“第二曲线”：养固发业务披上医疗外衣？.txt
-rw-r--r-- 1 root root  5726 Jun 14 18:19  经济142_差异化房贷利率促楼市平稳发展.txt
-rw-r--r-- 1 root root  3807 Jun 14 18:19  经济143_央行“降息”十日：17城首套房贷利率低至4.25%专家称后续仍有下调空间.txt
-rw-r--r-- 1 root root  2632 Jun 14 18:19  经济144_年内小微企业专项金融债发行规模1375亿元专家称短期看发行热情受两因素影响.txt
-rw-r--r-- 1 root root  2933 Jun 14 18:19  经济145_处罚落地！药明康德股东违规套现被罚2亿剩余持股市值近4亿.txt
-rw-r--r-- 1 root root 10337 Jun 14 18:19  经济146_中国“海王”去年暴赚近900亿，却只分红139亿！有股民不干了：太少，反对！公司详细
-rw-r--r-- 1 root root  4769 Jun 14 18:19  经济147_人民日报评景甜被罚：明星代言广告，应自醒自律自重.txt
-rw-r--r-- 1 root root  4282 Jun 14 18:19  经济148_原茅台电商系列酒负责人受贿448万元不如实交代赃款去向，退赃数额仅一半，一审被判
-rw-r--r-- 1 root root  5310 Jun 14 18:19  经济149_活久见！上市公司伪造审计报告被发现，竟说是被审计师陷害？真相是….txt
-rw-r--r-- 1 root root  5724 Jun 14 18:19  经济14_科创板三大成果凸显多元指数产品吸引长线资金.txt
-rw-r--r-- 1 root root  4261 Jun 14 18:19  经济150_突发！3家公司同一天公告退市！这公司打造“第二个宁王”梦碎股价仅剩0.59元.txt
```

分词并去除停用词之后的向量共有 35685 维，在去除df过低或者过高的词维度之后剩余的维度为 7012 维，这 625 个 7012 维的向量就作为我们降维之前的原始向量。

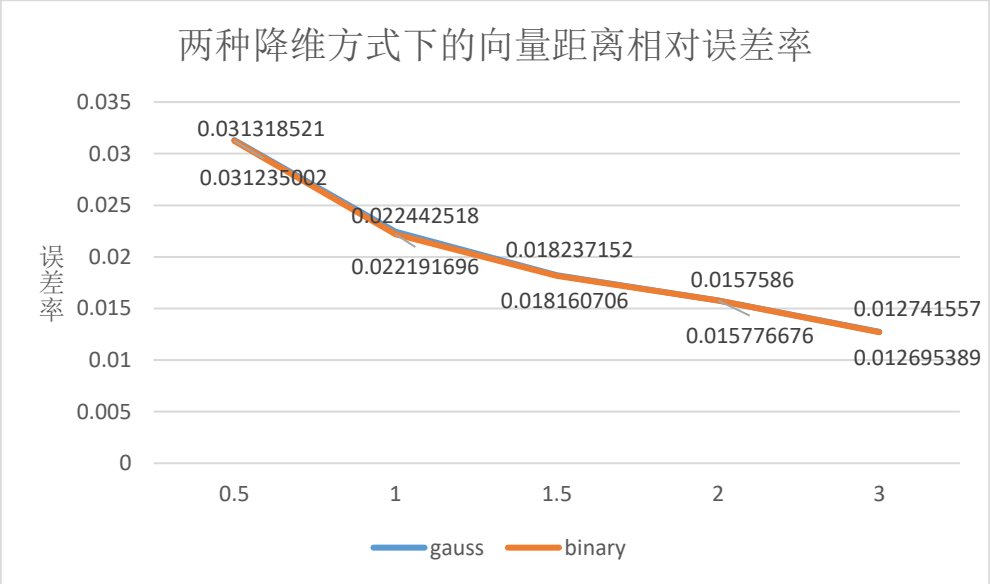


2．实验结果

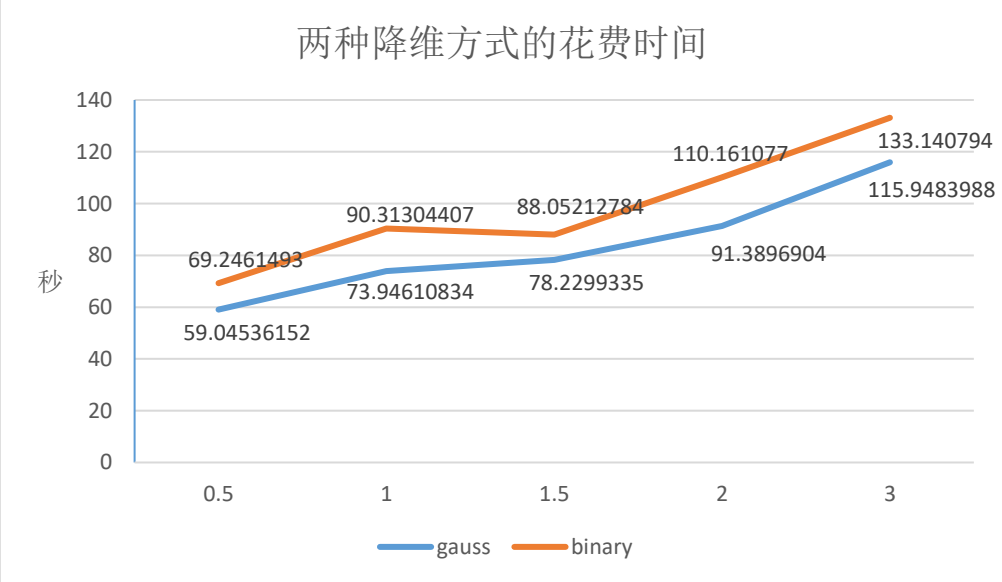
我们将目标维度$O\left(\frac{log\ N}{\varepsilon^2}\right)$系数分别取值为 0.5，1，1.5，2，3 进行降维。

放大系数	向量维数
0.5	322
1	644
1.5	966
2	1288
3	1932
（原词向量）	7012

每组重复进行 10 次并计算任意点对之间欧氏距离的平均误差率，结果如下：

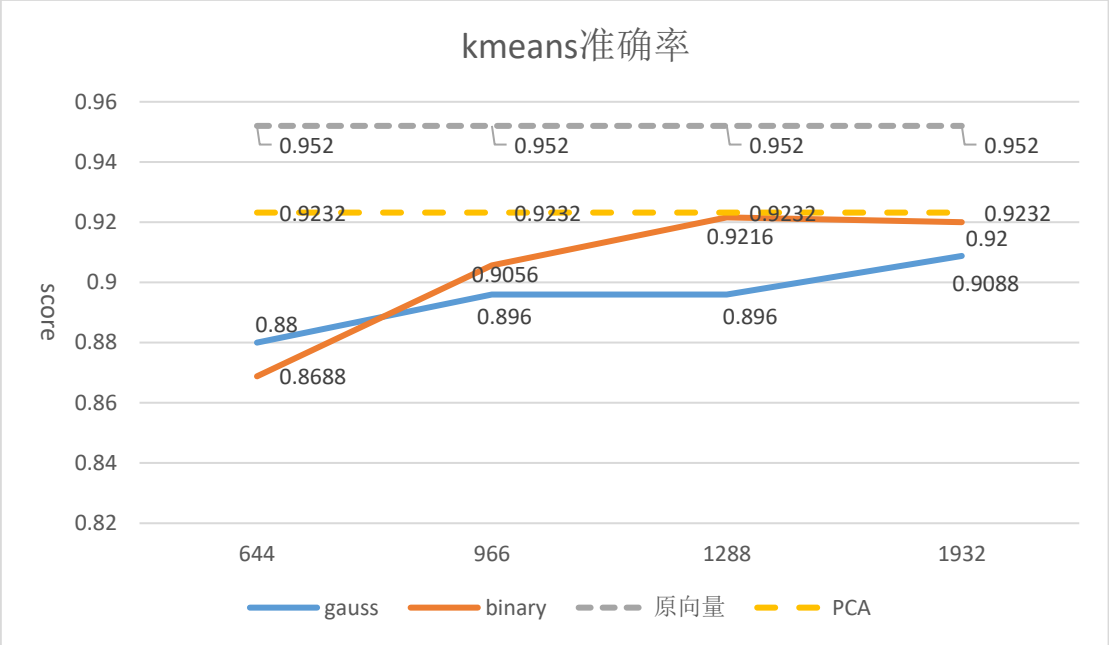


不同系数下两种降维方式所用的时间为

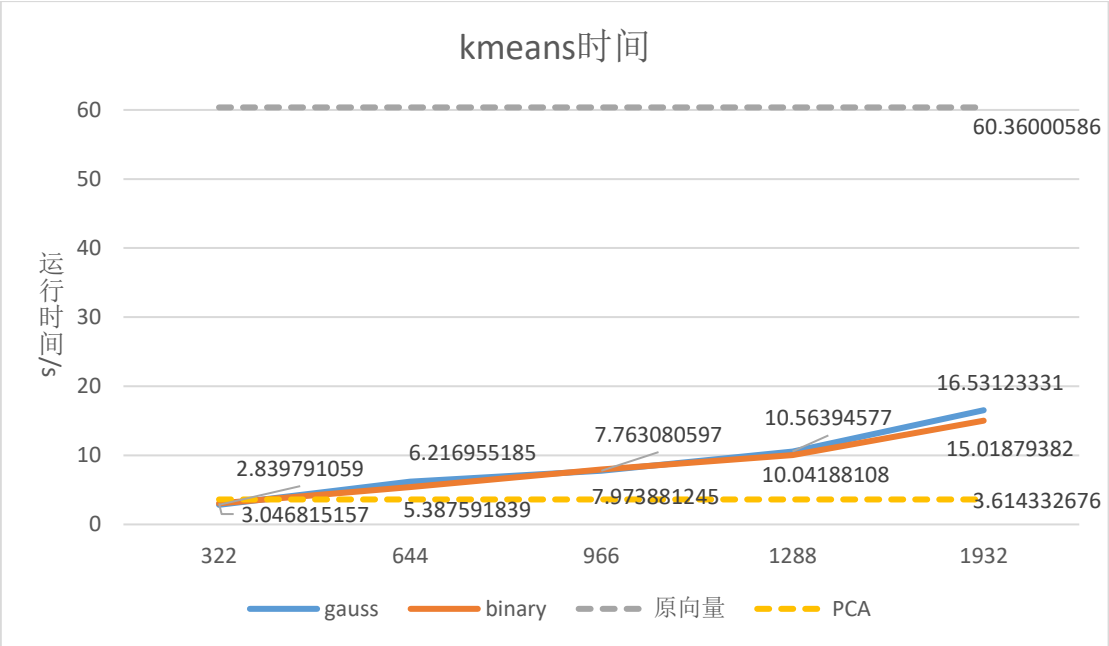


在聚类时，我们采用了 PCA 降维做效果对比，降维的结果维度为 624

不同系数下两种降维方式以及 PCA 降维、未降维所得到的聚类结果的准确度为



不同系数下两种降维方式以及 PCA 降维、未降维所得到的聚类结果所需的时间



四、结果观察与分析

1、 首先我们观察到，任意两个向量之间距离在降维之后与之前的相对误差值，两种降维方式的结果所差无几，而随着降维维数的增加，其误差率在逐渐变小。这是很好理解的，映射得到维数增加，会使得向量的特征信息保留的更多，相对误差就会更小。总的来看这些误差不超过 4%，特征信息保留较好。

2、 在降维速率方面，高斯矩阵的生成要快于 01 矩阵，高斯矩阵生成代码如下

```
transformerGauss = np.random.normal(0.0, 1.0, size=(k, max_len))
```

01 矩阵生成代码如下

```
for i in range(k):
    for j in range(max_len):
        t = random.random()
        if (t < 2 / 3):
            continue
        elif (t < 5 / 6):
            transformerBinary[i, j] = -1
        else:
            transformerBinary[i, j] = 1
```

我们猜测是 random 包中生成一个高斯分布的随机矩阵底层代码较为优秀，而我们生成 01 矩阵的过程基本都是手写实现，效率可能会略差。

3、 我们实现了一个 624 维的 PCA 降维作为对比，可以发现的是，实际上还是降维之前的向量准确度较高，能达到 95%之多，PCA 降维其次，能达到 92%。随着系数的不断增加，两种 JL 降维的准确率也在不断增加，但均为超过 PCA 的准确率，相比之下 01 矩阵降维且系数为 2 时其聚类准确率最接近 PCA 的准确率，且它随着系数增加效果能有很明显的提升。而高斯矩阵降维的聚类准确率最差，只有系数达到 3 的情况下勉强上 90%，更小的降维都使得准确率不到 90%。

其实我们虽然选择了三类不同的文章，但是经过后边的仔细阅读发现，科技类和经济类的文章共有的词是很多的，科技类的文章会讲述其对经济的意义，经济类的文章也会讨论到股市中的科技板块，导致这两类的降维时很容易出现特征性丢失的情况，因此如果以 TF-IDF 算法建立文章向量的话这种情况确实是无法

避免的。因此降维之后的聚类准确率会出现小幅度的下降。

而原向量聚类效果十分优秀，并未出现预想的纬度灾难现象。这个也是可以理解的，首先我们通过筛词，尤其是筛掉低频词，使得向量维度又原先的三万多下降到 7012，这样的降维会增强向量的特征性，使得向量的簇更为明显。再加上我们在聚类中使用的是余弦距离，相比较于欧式距离更能减轻高维度下随机向量纬度灾难效应，这一点会在第五部分提及。

4、运行时间上来看，整体都与向量维数成正比。而我们期待的 01 矩阵降维比高斯矩阵降维的效果好这一现象并未发生，二者时间相差无几，这和 python 的计算机制有关，其并没有稀疏矩阵优化，因此更为稀疏的 01 矩阵并未有任何优势。

5、综上来看，JL 降维（尤其是 01 矩阵降维）可以以损失小幅度准确率的代价，大大减少聚类所需时间，甚至在维数十分庞大时提升准确性。

五、实验过程中最值得说起的几个方面

1、

关于文章的选择，原本选择了中国新闻网的滚动新闻，但是发现滚动新闻中的所有新闻类别过散，并不好对不同类型的新闻进行标记，而且聚类效果也千奇百怪，因此在三个不同的网站选择了三类不同的新闻。

2、

另外关于文章向量的选择，最初我们设想的是简单的 01 向量，流程如下：（1）求所有文本的分词结果的并集，设并集大小为 m，（2）求所有文本的分词结果的并集，设并集大小为 m。

但是初步进行聚类时发显，每次聚类结果都不一样，相差很多。在与骆老师的交流中我们意识到，简单的存在与否并不能真实体现一个词对于一篇文章特征性的描述，其出现的频率、以及其在不同文章中出现的频次都对文章特性十分重要。因此骆老师建议我们将词语的频次考虑进去。在经过查找文献后，我们了解到了 TF-IDF 算法，这种算法可以通过公式很好地描述单个此于对文章特征的影

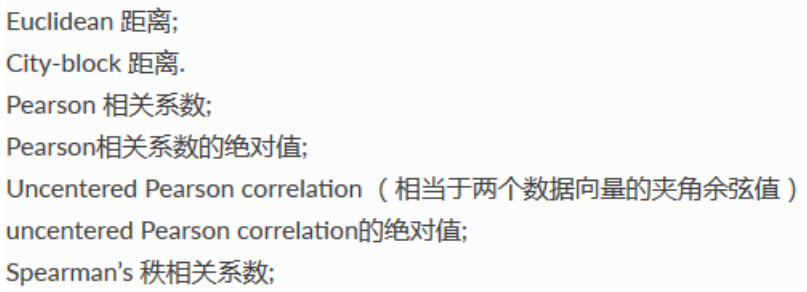
响，并且其剔除低频词和高频词的思想也很好地提升了后边聚类算法的效率。

3、

在初次实现降维时，我们发现向量之间相对距离的误差率达到了一倍之多，但是聚类效果也能很可观，这是十分反常的现象，查阅了更多资料后我们才发现降维之后实际上是需要归一化的，向量在降维之后，模长会等比例发生变化，因此我们需要对 gauss 矩阵上除以$\sqrt{k}$，k 是降维之后的维度，01 矩阵在除以$\sqrt{k}$的基础上还需要乘$\sqrt{3}$，因为 01 矩阵每一行都平均有三分之一的数字不是 0，因此最终的模长会受此影响变成原来的$\frac{1}{\sqrt{3}}$。

4、

刚开始，我们采用的是欧式距离，但是分类效果很不明显，会出现绝大多数样本都被聚类到同一类的情况，针对这种情况分析，应该是向量之间的距离不明显，找不到分界面分开他们，在实验课的时候也与老师交流过，可能是高维空间中向量之间的距离不明显，所以我们决定改换距离函数。我们查阅了很多距离函数（如下图所示），对比发现其中**效果最明显的是余弦函数**。



在改换余弦函数后，降维之前的聚类效果有了明显的提高，但是在降维之后的向量上进行聚类，效果还是不尽如人意，准确率在 0.4 左右，难以令人接受，于是我们选择改进 kmeans。

5、

初始，我们是自己实现的 kmeans 函数，发现效果不好，尝试使用 kmeans++ 算法中的初始平均点选择方法，但是改进之后，效果并没有很大的提升。

后又改进了迭代条件，之前的迭代条件是前一次的分类结果和本次一样就停止迭代，考虑这种情况可能导致迭代次数不足，才出现聚类效果分散、不准确的情况。所以现在将迭代次数直接设置为固定值（该值经多次调试，得到了较好的效果），效果有了明显提升，也是我们最终实现的版本。

```
times = 0
while times < 100:
    times += 1
```

6、整体来说这个实验达到了目的，实现了一个“反直觉”的降维方法，虽然效果不如更加常用的 PCA，但是其理论的作用更大于其实际作用，它很好地证明了随机线性投影的可行性，在现代机器学习中，JL 引理也为我们理解、调试模型维度等相关参数提供了重要的理论支撑。