实验报告

实验一：数据仓库与数据挖掘

实验内容

基于互联网数据，对巴以两国及中东局势进行热点话题的建模、提取和舆情分析，从而学习文本主题分析基本原理、基本方法和基本流程。主要分解为以下 4 部分内容：

1. 文本数据的预处理基本方法和基本流程；
2. 基于 LDA/OLDA 主题模型的文本的表示方法；
3. 基于聚类方法的主题发现和跟踪方法及基本流程；
4. 基于所提取的热点话题，对两国舆情及对全球政治经济的影响进行分析。
5. 基于新的文本数据，对热点话题进行预测分析。

实验目的

通过本实验，掌握以下技能：

1. 学习文本数据预处理的基本方法和流程；
2. 理解并应用 LDA/OLDA 主题模型进行文本表示；
3. 掌握基于聚类方法的主题发现和跟踪技术；
4. 能够分析热点话题对舆情及全球政治经济的影响；
5. 学会利用新的文本数据进行热点话题的预测分析。

实验原理

本实验的原理主要包括以下几个方面：

1. **文本数据获取**：
   - 使用爬虫技术从微博获取数据：编写爬虫脚本，自动抓取微博上的文本数据。
2. **文本数据预处理**：
   - 数据清洗：去除噪声数据，如HTML标签、特殊字符等。
   - 分词：将文本数据分割成单独的词语或短语。
   - 停用词过滤：去除常见但无实际意义的词语，如“的”、“是”等。
   - 词干提取：将词语还原为其基本形式。
3. **LDA/OLDA 主题模型**：

- LDA（Latent Dirichlet Allocation）：一种生成模型，用于发现文档集合中潜在的主题分布。
- OLDA（Online LDA）：LDA的在线变体，适用于大规模数据集的增量学习。

4. **聚类方法**：
   - K-means 聚类：一种常用的无监督学习算法，用于将数据点分配到K个簇中。
   - 层次聚类：一种基于树状结构的聚类方法，通过不断合并或分裂簇来构建层次树。
5. **舆情分析**：
   - 热点话题建模
     - 提取热点话题的主题词，并以词云、标签云等形式进行可视化展示。
   - 热点话题的追踪
     - 对于新的文本数据，将其表示为主题分布，依据已有的热点话题，对新文本数据基于分类的方法进行判定，判别该话题是否是属于已有话题。如果是已有话题，则将其划归为旧话题，否则将视为是新的话题。
6. **预测分析**：
   - 时间序列分析：利用历史数据，预测未来的趋势和变化。
   - 机器学习模型：训练和应用各种机器学习模型，对新数据进行预测和分类。

这些原理共同构成了本实验的理论基础，帮助我们理解和应用数据仓库与数据挖掘技术进行热点话题的建模和分析。

实验步骤

1. **获取微博数据**
   - 使用 `weibo-search` 下的由 `scrapy` 框架搭建的微博爬虫获取微博数据。
   - 编写爬虫脚本，设置关键词，启动爬虫，自动抓取相关微博文本数据。
2. **热点话题建模**
   - 数据预处理：
     - 读取 CSV 文件中的数据，并进行文本清洗，移除所有的非中文字符。
     - 使用 jieba 库对清洗后的文本进行分词。
     - 加载停用词表，并去除分词结果中的停用词。
   - TF-IDF 转换：
     - 初始化 TfidfVectorizer，对清洗后的文本进行 TF-IDF 转换，生成 TF-IDF 矩阵。
     - 将稀疏矩阵转换为密集矩阵，并获取词汇表。
   - LDA 模型训练：
     - 使用 Gensim 库创建词典，并将 TF-IDF 矩阵转换为 Gensim 格式的文档词袋表示。
     - 检查是否存在已有的 LDA 模型文件，如果存在则加载模型，否则新建一个 LDA 模型并保存。输出每个主题的前几个关键词。
   - 主题分布获取：
     - 获取每篇文档的主题分布，并将稀疏表示转换为密集向量。
   - K-means 聚类：

    - 定义 K-means 模型，并拟合模型，获取每个文档的聚类标签和聚类中心。
    - 将聚类标签加入到原始数据框中，并保存结果为新的 CSV 文件。
  - PCA 降维和可视化：
    - 使用 PCA 将主题分布降维到 2D，并绘制聚类结果图，保存为 `kmeans_clusters.png` 。
3. **热点话题的追踪**
  - 热点话题关键词提取：
    - 打印每个聚类的热点话题关键词，获取每个簇的前 5 个最重要的主题。
    - 汇总每个簇的关键词，并生成词云，保存为 `cluster_{i}_wordcloud.png` 。
  - 新文本数据的分类：
    - 对于新的文本数据，使用相同的预处理和 TF-IDF 转换步骤，将其表示为主题分布。
    - 使用训练好的 LDA 模型和 K-means 模型，对新文本数据进行分类，判别该话题是否属于已有话题。如果是已有话题，则将其划归为旧话题，否则将视为是新的话题。

项目文件夹介绍

本项目文件夹包含以下文件和目录，每个文件和目录的作用如下：

根目录

- 巴以冲突**.csv**： 包含巴以冲突相关的微博数据。
- **hit_stopwords.txt**： 停用词表文件，包含需要过滤掉的无意义词语。
- **lda_model.gensim**： 训练好的LDA模型文件。
- **lda_model.gensim.expElogbeta.npy**： LDA模型的中间结果文件。
- **lda_model.gensim.id2word**： LDA模型的词典文件。
- **lda_model.gensim.state**： LDA模型的状态文件。
- **process.py**： 数据处理脚本，包含数据预处理、LDA模型训练、K-means聚类等步骤。
- **processed_data.csv**： 处理后的数据文件，包含聚类标签等信息。
- **test.py**： 测试脚本，用于生成中文词云图。
- **tfidf_matrix.npz**： 存储TF-IDF矩阵的文件。

weibo-search 目录

- **.gitignore**： Git忽略文件配置。
- **README.md**： 项目说明文档，包含功能介绍、使用说明等。
- **requirements.txt**： 项目依赖包列表。
- **scrapy.cfg**： Scrapy配置文件。

weibo-search/weibo 目录

- **__init__.py**： 初始化文件。

- **items.py**： 定义Scrapy爬取的Item结构。
- **middlewares.py**： 定义Scrapy中间件。
- **pipelines.py**： 定义数据处理管道，包括CSV、图片、视频、MongoDB、MySQL等处理。
- **settings.py**： Scrapy项目的配置文件。

weibo-search/weibo/spiders 目录

- **__init__.py**： 初始化文件。
- **search.py**： 定义Scrapy爬虫，负责从微博抓取数据。

weibo-search/weibo/utils 目录

- **__init__.py**： 初始化文件。
- **region.py**： 包含地区代码和城市信息。
- **util.py**： 包含各种实用函数，如微博类型转换、关键词列表获取、区域筛选等。

实验结果

1. **数据预处理结果**:
   - 清洗后的文本数据示例:

```
0    俄乌局势俄乌局势新进展以色列境内所有客运铁路均已暂停运营以便利部队和装备的转移与此同时至
1    巴以冲突有三战的味道了大毛终于开窍了斯拉夫人互砍确实太单调拉上阿拉伯人一起玩更热闹嘿嘿中
2                          巴以冲突以色列的客户还挺多的打仗了不知道还能不能收到包裹世界和平
3    巴以问题很多年了但本次事件发生在这个时间上感觉是希望油价不跌希望油价上涨的国家在背后撺掇
4                                过段时间哈马斯又躲进巴勒斯坦人民中去了巴以冲突
Name: cleaned_content, dtype: object
```

   - 分词后的文本数据示例:

```
0    俄乌 局势 俄乌 局势 新进展 以色列 境内 所有 客运 铁路 均 已 暂停 运营 以 便利...
1    巴 以 冲突 有三战 的 味道 了 大毛 终于 开窍 了 斯拉夫人 互砍 确实 太 单调 拉...
2    巴 以 冲突 以色列 的 客户 还 挺 多 的 打仗 了 不 知道 还 能 不能 收到 包裹...
3    巴 以 问题 很多年 了 但 本次 事件 发生 在 这个 时间 上 感觉 是 希望 油价 不...
4                    过段时间 哈马斯 又 躲 进 巴勒斯坦 人民 中去 了 巴以 冲突
Name: segmented_content, dtype: object
```

   - 去停用词后的文本数据示例:

```
0    俄乌 局势 俄乌 局势 新进展 以色列 境内 所有 客运 铁路 均 已 暂停 运营 便利 部...
1    巴 冲突 有三战 味道 大毛 终于 开窍 斯拉夫人 互砍 确实 太 单调 拉上 阿拉伯人 一...
2                巴 冲突 以色列 客户 还 挺 打仗 不 知道 还 不能 收到 包裹 世界 和平
3    巴 问题 很多年 本次 事件 发生 时间 上 感觉 希望 油价 不 跌 希望 油价 上涨 国...
4                          过段时间 哈马斯 躲 进 巴勒斯坦 人民 中去 巴以 冲突
Name: filtered_content, dtype: object
```

2. **TF-IDF 矩阵**:
   - 生成的 TF-IDF 矩阵形状为 `(文档数量, 词汇数量)`，例如 `(10099, 28993)`。
3. **LDA 模型训练结果**:
   - 每个主题的前几个关键词:

```
(0, '0.009*"国际" + 0.008*"文章" + 0.007*"头条" + 0.007*"支持" + 0.007*"黄金"')
(1, '0.018*"局势" + 0.014*"表示" + 0.014*"升级" + 0.013*"进一步" + 0.012*"停火"')
(2, '0.029*"冲突" + 0.021*"以色列" + 0.019*"微博" + 0.018*"视频" + 0.017*"宣布"')
(3, '0.050*"冲突" + 0.017*"世界" + 0.013*"和平" + 0.013*"战争" + 0.010*"国家"')
(4, '0.034*"特使" + 0.031*"中东问题" + 0.028*"联合国" + 0.028*"微博" + 0.027*"视频"')
(5, '0.044*"视频" + 0.043*"微博" + 0.040*"冲突" + 0.026*"以色列" + 0.020*"曝光"')
(6, '0.033*"冲突" + 0.029*"加沙" + 0.021*"视频" + 0.020*"微博" + 0.019*"巴勒斯坦"')
(7, '0.028*"冲突" + 0.022*"哈马斯" + 0.021*"以色列" + 0.014*"视频" + 0.013*"微博"')
(8, '0.084*"冲突" + 0.056*"视频" + 0.053*"微博" + 0.046*"以色列" + 0.029*"巴勒斯坦"')
(9, '0.027*"加沙" + 0.026*"冲突" + 0.021*"死亡" + 0.020*"以色列" + 0.019*"地带"')
```

4. **K-means 聚类结果**:
   - 每个文档的聚类标签:

```
[2 2 0 ... 1 0 0]
```

   - 每个簇的中心:

```
Cluster 0: [0.03364185 0.03940007 0.03263206 0.6122275  0.01746702 0.03027065
0.05475445 0.07941668 0.06224021 0.03794952]
Cluster 1: [0.03541241 0.03066715 0.03833424 0.08903068 0.03068767 0.03833615
0.06619668 0.06455738 0.54293096 0.06384668]
Cluster 2: [0.02540039 0.05545196 0.09440226 0.0805777  0.03521213 0.04922051
0.16693021 0.24690935 0.05786454 0.18803091]
```

5. **聚类结果可视化**:
   - 使用 PCA 将主题分布降维到 2D，并绘制聚类结果图，保存为kmeans_clusters.png
6. **热点话题关键词提取**:
   - 每个聚类的热点话题关键词:

```
Cluster 0 Hot Topics:
Topic 3: 0.050*"冲突" + 0.017*"世界" + 0.013*"和平" + 0.013*"战争" + 0.010*"国家"
Topic 7: 0.028*"冲突" + 0.022*"哈马斯" + 0.021*"以色列" + 0.014*"视频" + 0.013*"微博"
Topic 8: 0.084*"冲突" + 0.056*"视频" + 0.053*"微博" + 0.046*"以色列" + 0.029*"巴勒斯坦"
Topic 6: 0.033*"冲突" + 0.029*"加沙" + 0.021*"视频" + 0.020*"微博" + 0.019*"巴勒斯坦"
Topic 1: 0.018*"局势" + 0.014*"表示" + 0.014*"升级" + 0.013*"进一步" + 0.012*"停火"

Cluster 1 Hot Topics:
Topic 8: 0.084*"冲突" + 0.056*"视频" + 0.053*"微博" + 0.046*"以色列" + 0.029*"巴勒斯坦"
Topic 3: 0.050*"冲突" + 0.017*"世界" + 0.013*"和平" + 0.013*"战争" + 0.010*"国家"
Topic 6: 0.033*"冲突" + 0.029*"加沙" + 0.021*"视频" + 0.020*"微博" + 0.019*"巴勒斯坦"
Topic 7: 0.028*"冲突" + 0.022*"哈马斯" + 0.021*"以色列" + 0.014*"视频" + 0.013*"微博"
Topic 9: 0.027*"加沙" + 0.026*"冲突" + 0.021*"死亡" + 0.020*"以色列" + 0.019*"地带"

Cluster 2 Hot Topics:
Topic 7: 0.028*"冲突" + 0.022*"哈马斯" + 0.021*"以色列" + 0.014*"视频" + 0.013*"微博"
Topic 9: 0.027*"加沙" + 0.026*"冲突" + 0.021*"死亡" + 0.020*"以色列" + 0.019*"地带"
Topic 6: 0.033*"冲突" + 0.029*"加沙" + 0.021*"视频" + 0.020*"微博" + 0.019*"巴勒斯坦"
Topic 2: 0.029*"冲突" + 0.021*"以色列" + 0.019*"微博" + 0.018*"视频" + 0.017*"宣布"
Topic 3: 0.050*"冲突" + 0.017*"世界" + 0.013*"和平" + 0.013*"战争" + 0.010*"国家"
```

- 生成的词云图示例：
  cluster_0_wordcloud.png
  cluster_1_wordcloud.png
  cluster_2_wordcloud.png

7. **新文本数据的分类**：

   i. **使用 OLDA 对新文档进行预测**

   如果想使用训练好的 OLDA 模型对新文档进行主题预测，可以按照以下步骤：

```python
# 假设有一篇新的文档，我们首先对其进行 TF-IDF 预处理
new_document = "这是一篇新的新闻文章，内容涉及巴以冲突和中东局势。"
new_document_cleaned = clean_text(new_document)  # 清洗文本
new_document_segmented = jieba_cut(new_document_cleaned)  # 分词
new_document_filtered = remove_stopwords(new_document_segmented)  # 去停用词

# 将新文档转换为词袋表示
new_document_bow = dictionary.doc2bow(new_document_filtered.split())

# 使用 LDA 模型进行主题预测
new_document_topic = lda_model[new_document_bow]
print(f"新文档的主题分布: {new_document_topic}")
```

`new_document_topic` 会返回该文档在各个主题上的分布，形式如下：

```
[(0, 0.75), (1, 0.25)]
```

这表示新文档属于主题 0 的概率为 75%，属于主题 1 的概率为 25%。

ii. **OLDA 模型的增量训练**

如果有新的文档数据，可以继续训练 OLDA 模型，采用增量学习的方式来适应新数据：

```
# 假设新的文档数据也已经被处理成 TF-IDF 矩阵
new_tfidf_matrix = sp.load_npz('new_tfidf_matrix.npz')
new_tfidf_matrix_dense = new_tfidf_matrix.toarray()

# 将新数据转换为 Gensim 格式
new_corpus = [dictionary.doc2bow(doc) for doc in new_tfidf_matrix_dense]

# 增量训练
lda_model.update(new_corpus)
```

增量训练是 OLDA 的一个重要特点，它能在不重新训练整个模型的情况下，根据新数据调整主题分布。

实验总结

1. **数据预处理**：
   - 通过数据清洗、分词和去停用词等步骤，有效地提取了文本中的关键信息，为后续的主题建模和聚类分析打下了坚实的基础。
2. **主题建模**：
   - 使用 LDA 模型对文本数据进行了主题建模，成功提取了多个具有代表性的主题。每个主题的关键词能够较好地反映该主题的核心内容。
3. **聚类分析**：
   - 通过 K-means 聚类算法，将文档分配到不同的簇中，并使用 PCA 对聚类结果进行了可视化。聚类结果图展示了不同簇之间的分布情况，有助于理解文档的主题分布。
4. **热点话题关键词提取**：
   - 提取了每个聚类的热点话题关键词，并生成了词云图。词云图直观地展示了每个簇的主要关键词，有助于快速了解每个簇的主题内容。
5. **新文本数据的分类**：
   - 使用训练好的 LDA 模型和 K-means 模型，对新文本数据进行了分类。通过将新文本数据表示为主题分布，能够有效地判别新文本是否属于已有话题。
6. **OLDA 模型的增量训练**：

- 采用增量学习的方式，对 OLDA 模型进行了更新。增量训练能够在不重新训练整个模型的情况下，根据新数据调整主题分布，提高了模型的适应性和灵活性。

7. **实验结果的可视化**：
   - 通过生成聚类结果图和词云图，直观地展示了实验结果。这些可视化结果有助于更好地理解和分析文本数据的主题分布和热点话题。
8. **总结**：
   - 本实验通过数据预处理、主题建模、聚类分析和热点话题提取等步骤，成功实现了对巴以冲突相关微博数据的舆情分析。实验结果表明，LDA 模型和 K-means 聚类算法能够有效地提取和分类文本数据的主题内容。通过增量训练，模型能够适应新数据的变化，提高了分析的准确性和实时性。