“数据科学与工程”实验一

一. 实验内容

基于互联网数据，针对当前国内、国际局势，围绕中、日、韩及美国、欧盟的社交媒体数据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. 数据（文本、图像、视频）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，对国内外舆情及对全球政治经济影响进行分析。
5. 基于新的文本数据，对热点话题进行预测分析。

二. 实验过程

实验内容

第一部分：文本挖掘的基本流程

1.1 基本的文本主题挖掘流程

文本主题挖掘的研究框架如图 1 所示，主要分为三个部分，分别是数据采集层、数据预处理层和舆情分析层。

- 数据采集层主要使用爬虫脚本获取社交媒体的数据（如微博数据），然后存储在本地。
- 数据预处理层包含数据清洗、Jieba 分词、去停用词 、TF-IDF（term frequency-inverse document frequency）提取关键词。
- 舆情分析层一方面将预处理后的结果输入到 word2vec 模型，进行训练，另一方面对预处理后的数据按照时间顺序划分时间片，输入到 OLDA 模型，将从 OLDA 模型输出的主题词利用词向量表示、聚类，发现热点聚类簇，并对热点簇中的主题使用词性标注的方式进行分析。

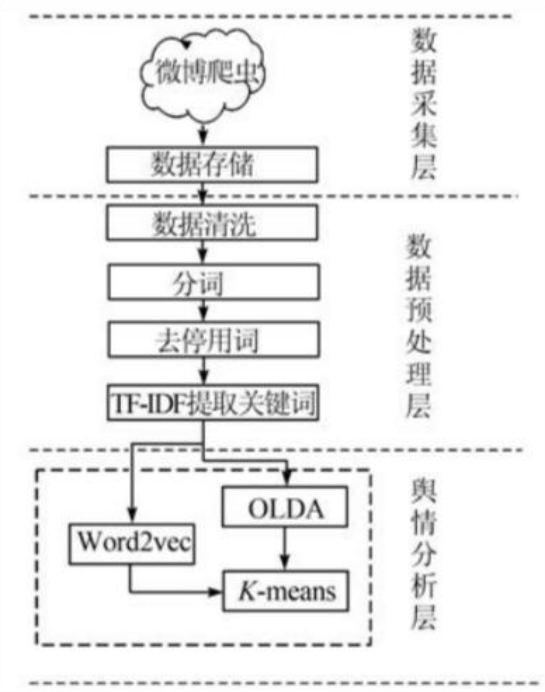

图 1 基本的文本主题挖掘框架

1.2 基于多源-时空数据特性的文本主题挖掘框架

1. 多源数据集的引入：

当前，国际、国内发生了很多重大事件，引起了世界各国极大的社会关注 ，造成了巨大的社会影响。

全球新闻媒体对其广泛报道 ，大量网民在微博、知乎、微信公众平台和 Twitter 等国内外社交媒体上发表了对此次事件的看法。本实验要求引入国内外各个社交平台、 网络新闻和电视新闻等多源数据，形成全方位的主题挖掘及舆情分析 ，深入探索国际、国内局势发展趋势及对国际局势造成的影响。

2. 时序 OLDA 模型的引入：

社交媒体数据集中的数据呈现出了时空特性。其中，空间特性可通过地区进行区分。此时，在进行热点话题建模时，还需考虑时间特性，因而需要利用在线潜在狄利克雷分布（online latent Dirichlet allocation， OLDA）模型，按照时间序列分解数据，进行主题演化进行发现、提取和分析。

3. 基于多源-时空特性的文本主题挖掘框架

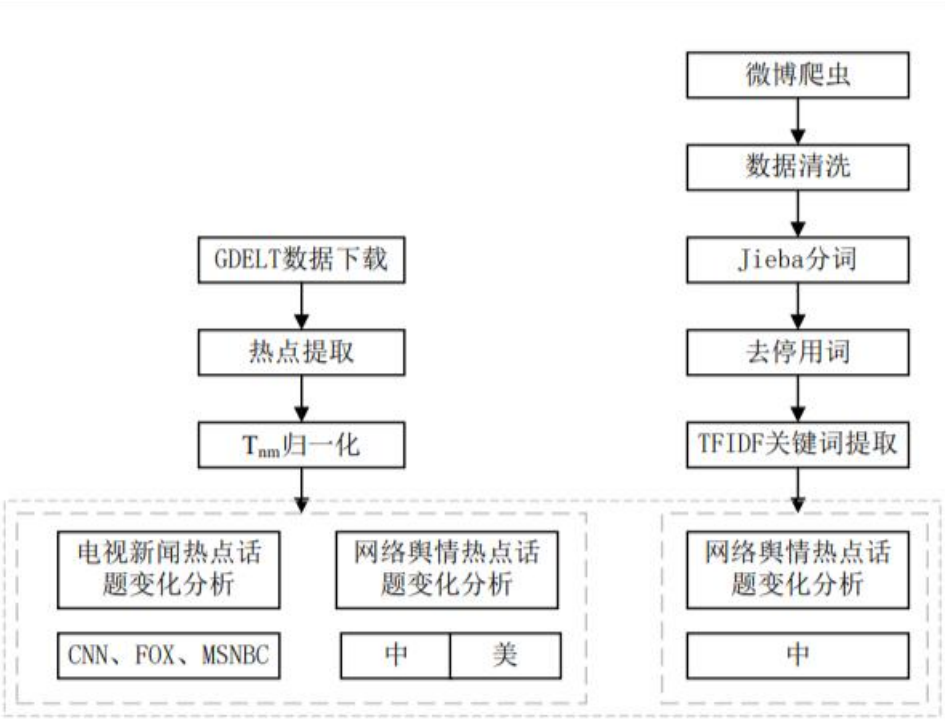

图 2 基于多源-时空特性的文本主题挖掘框架

1.3 文本数据集的预处理方法和流程

1.文本数据集的预处理流程：

① 输入文档数据集；

② “分词”（例如，使用 Iieba 软件包实现）；

③ 取名词；

④ 在分词结果上去停用词，利用一个停用词表，编写代码，利用这个 stop_words 表把不需要的停用词（如啊，了，的等）去掉。

⑤ 建词典（就是把所有文档经过前面 4 步处理过的单词汇总）

⑥ 用词典将每一篇文档利用生成的词典表示成向量的形式,即使用词袋（BOW-Bag of Words）方法。

2. 常用文本分词包：

① 斯坦福大学自然语言处理的网页，能够下载能够处理英文、中文等语义的分词工具。http://nlp.stanford.edu/software/tagger.shtml

② 中科院的能够处理分词的工具——Ictcas

③ Python 中的 Jieba 分词包。

3. 基于 CHI-TFIDF 算法改进特征词的选取

从 CHI 计算公式可以看出，CHI 值只与词语出现的文档数量有关，这样就会出现如下问题：如果 A 和 B 两个词语同时在同样的文档中出现，这样 A 和 B 计算的 CHI 值应该相等，但如果词语 B 在文档中出现频率比 A 大，那么 B 词语的重要性应该要比词语 A 高，但在使用 CHI 进行特征词提取时无法对这种情况做有效处理。

TF-IDF 是由 Salton 在 1988 年提出的，用于评估一个词对一个文章的重要程度，其中：TF 称为词频，既该词汇在文章中的次数比重；IDF 称为逆文章频率，用于表现该词在所有文章中的频率，该方法可以辨别词汇是否具有区别不同类别的能力。

本课程依据以往的教学实践，建议使用一种“CHI 改进算法”，将特征词提取算法中加入词频（TF）因素。改进后 CHI 计算公式如下：

$$\mathrm{CHI}(t,c_i)=\frac{N*(AD-BC)^2}{(A+C)*(B+D)*(A+B)*(C+D)}*f(TF_{ci})$$

4. 基于词空间-词袋模型（Bag of words, BoW）的文本表示

基于词袋模型，在建立了数据词典后，将文本初步表示成基于词的文档向量的形式，后续将输入到 OLDA 模型中，进行主题建模，将文档由词空间表示为主题空间。

1.4 基于 OLDA 的文本主题建模

1.LDA 模型

使用 LDA 概率主题模型生成文本的具体过程描述如下：

S1：根据抽取出的 LDA 模型，读入文档的主题分布 θ 以及每个主题上面单词的分布 φ；

S2：统计文本的长度 N;

S3：$d=\varnothing$；

S4：根据主题分布 θ，使用轮盘赌算法随机选择一个主题 z;

S5：根据主题 z 上的单词概率分布 $\varphi(z)$，使用轮盘赌算法随机选择一个单词 w 作为文本的一个新词，令 $d=d\cup\{w\}$；

S6：重复 S4 和 S5，直到 $d=N$。

其中 θ 是一个 $1\times k$ 的随机行向量，它的具体函数形式就是 Dirichlet 分布，这一分布保证 θ 的 k 个分量 $\theta\_1, \theta\_2, \ldots, \theta\_k$ 都取连续的非负值，且 $\theta\_1 + \theta\_2 + \ldots + \theta\_k = 1$；$z$ 是离散随机变量，在主题 T 中取 k 个离散值， w 是离散随机变量,在词汇表 V 中取 $|V|$ 个离散值， $\varphi(z)$ 是给定 z 时 w 的条件分布，可以把它看作 $k\times|V|$ 的矩阵。

2.OLDA 模型

网络上的文本信息随着事件的发酵逐渐变多，这给文本处理工作带来了新的挑战。文本的增多使计算量增加，却不能检测到文本中事件主体随着时间的变化趋势。

为了解决这个问题，采用 OLDA 模 型，引入文本的时间信息到 LDA 模型。

OLDA 模型与 LDA 模型的最大不同是加入了时间维度信息。将文本按照时间顺序进行排序，划分为若干个时间片，OLDA 模型用来处理该时间片内的信息。这样既减轻了模型处理数据的计算量又提高了模型对事件主体检测的灵敏度。

由于同一事件在时间上不会发生很大的改变， 因此在用 OLDA 模型处理数据的过程中，先验参数继承了上一个时间片模型的参数。主题的先验参数保存在 B^t_k 中，B^t_k 表示主题 k 在 δ 个时间片上的演化矩阵，δ 为历史时间片的长度。时间片 t 主题 k 的先验参数表达式如下：

$$\beta_k^t = B_k^{t-1}\omega^{\delta}$$

其中，ω^{δ} 为不同时间片上的权重，权重是用来衡量不同时间片对当前时间片的影响程度。

第二部分 通过聚类对热点话题进行提取分析

1. 基于 K-means 热点话题聚类分析

将 OLDA 获得的当前时间窗口的主题词，删除概率较小的主题词，使用词向量对主题词进行表示，如：

主 题 k =[w1，w2，w3，…，wh],

主 题 向 量 $x = \frac{w_1 + w_2 + w_3 + \cdots + w_h}{h}$。

K-means 是一种无监督 聚类算法，该算法输入的是无标签的数据集，输出的是 l 个簇 C，C = C1，C2，…，Cl。

K-means 聚类算法的损失函数为：

$$E = \sum_{i=1}^{l}\sum_{x\in C_i}\|x-\mu_i\|^2$$

其中 μ_i 为簇 Ci 的中心点：

$$\mu_i = \frac{1}{|C_i|}\sum_{x \in C_i} x$$

将主题使用词向量进行表示，K-means 通过随 机选择 l 个主题作为簇的中心，计算每个主题和各 个簇中心主题的距离，将样本点划入距离最近的簇 中心主题，根据已经划分的簇，重新计算簇中心主题的位置。通过增加迭代次数来提高主题聚类的准确率。

2.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。

3. 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。