爱文本 AI texts

智能文本分析系统
Intelligent text analysis system

产品手册
PRODUCT MANUAL

上海经禾信息技术有限公司
Shanghai Jinghe Information Technology Co.,Ltd

目　录

1. 文本数据 1
    1.1. 公司公告 2
    1.2. 新闻报道 2
    1.3. 研究报告 2
    1.4. 公司研究 2
    1.5. 综合文本 3
    1.6. 文本信息 3
        1.6.1. 基本信息 3
        1.6.2. 词频统计 3
        1.6.3. 可读性 4
2. 自由搜索 4
3. 文本分析 5
    3.1. 基本信息 5
    3.2. 情感分析 5
    3.3. 可读性 5
4. 文本挖掘 6
    4.1. 文本处理 6
        4.1.1. 数据读取 6
        4.1.2. 分词 6
        4.1.3. 去停用词 7
        4.1.4. 特征词库 7
        4.1.5. 文本向量化 8
    4.2. 机器学习 9
        4.2.1. 相似度 10
        4.2.2. 文本分类 11
        4.2.3. 文本聚类 17
        4.2.4. 主题模型 18
参考文献 20

随着文本分析技术的进步，越来越多的财经领域学者选择利用文本分析开展研究。文本分析技术已经成为财经领域重要的研究方法。爱文本智能文本分析系统致力于为财经领域研究者提供便利的可视化文本分析工具。系统的文本挖掘过程如图 1 所示，主要包括以下三个模块：①文本数据。文本数据采集是文本挖掘的核心工作和首要任务，爱文本系统网页端(www.aitexts.com)提供了覆盖各个来源的海量财经领域文本的千亿条文本挖掘信息。②文本分析。这一模块提供了文本研究领域常见指标的自动统计，包括基本信息统计、自定义词频统计、自定义情感分析、可读性统计等。③文本挖掘。这一模块首先将文本转化为在计算机中可以处理的中间形式，接下来集成了文本挖掘中的常见模型，包括相似度统计、文本分类、文本聚类和主题模型。用户无需编程，即可在这一模块挖掘得到文本中的模式和知识，并且可以进行模式和知识的评价、展示等。

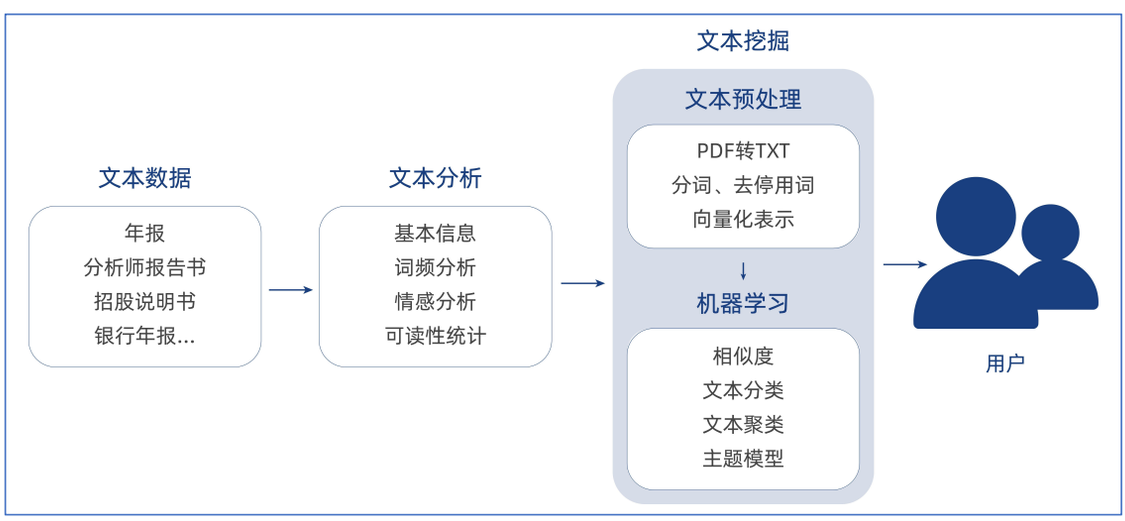


图 1 智能文本分析系统工作流程

1. 文本数据

爱文本系统网页版(www.aitexts.com)提供了研究和投资中常用的各类重要文本及其分析结果，文本类型包括公司公告、新闻报道、研究报告、公司研究和综合文本等。这些文本获取难度较高、数量巨大、处理过程复杂而在学术研究和投资决策中又经常用到，为此，为了让广大用户实现“文本自由”，本系统从文本获取、数据清洗、数据核查和结果呈现等方面对这些重要文本进行了全方位处理，力图科学、简约和多维地为使用者提供各类文本分析结果。

这一模块的文本数据包括公司公告、新闻报道、研究报告、公司研究、综合文本等五个

模块；文本挖掘信息包括上述模块中文本的基本信息、词频统计、可读性、相似度等。

1.1. 公司公告

主要囊括了来自上海证券交易所和深圳证券交易所的 A 股上市公司公告文本，不仅包括年报、管理层讨论与分析、招股说明书、企业社会责任报告、澄清公告等重要公告，还提供并购重组、增发预案、独董意见等多种常见公司公告。其中，大部分公告均提供了长时间序列的多年度文本，并提供基本信息、词频统计、语调分析和相似度计算等重要字段的计算结果。用户可以在该模块实现任意词搜索词频，提供基于 LM 词典、NTUSD 词典和爱文本词典的情感分析，多维度的可读性衡量指标等重要分析结果。

1.2. 新闻报道

主要包括上市公司新闻报道以及股吧评论文本数据，其中，公司新闻文本数据来源涵盖 400 多家网络媒体和 600 多家报纸刊物，以这些海量数据为基础，通过数据清洗、专家判断等手段，结合人工智能算法为用户提供上市公司财经新闻基本信息、财经新闻量化统计信息以及新闻相似度等重要文本字段；股吧评论则根据中国最大的股吧论坛网民对上市公司股吧帖子进行文本分析、数量统计等，提供帖子总量、正面帖子量、负面帖子量、阅读数、评论数等文本字段。

1.3. 研究报告

主要提供券商和金融分析师发布的研究报告文本数据，包括公司研究、晨会报告、宏观策略、行业研究、债券研究和期货研究等研究报告文本，其中，公司研究提供了 2007 年以来的 40 多万份上市公司报告，为学术研究提供了重要数据支持。本模块数据提供基本信息、词频统计、语调分析、可读性和相似度等多个文本分析维度。用户同样可以在该模块实现任意词搜索词频，提供基于 LM 词典、NTUSD 词典和爱文本词典的情感分析，多维度的可读性衡量指标等重要分析结果。

1.4. 公司研究

本模块提供了上市公司经营过程中产生的重要文本信息分析结果。这些文本包括上市公

司监管问询、业绩说明会、IPO 路演、关键审计事项和投资者问答等，其中，监管问询是监管机构对上市公司经营情况提出的问询问题以及上市公司的回复情况文本分析；业绩说明会是上市公司每年召开业绩说明会的投资者问答文本分析；IPO 路演是拟上市公司在进行路演时的投资者问答文本分析；关键审计事项是审计师在审计中对上市公司年报提出的关键事项说明文本分析；投资者问答则提供了来自投资者平台上投资者与公司董秘的问答互动文本分析。本系统对以上这些文本提供了词频统计、语调分析和可读性等重要文本分析结果。

1.5. 综合文本

综合文本主要展示在学术研究和投资实践中可能用到的其他重要文本分析结果，目前该模块包含了银行年报和券商年报等重要文本。通过手工收集了 200 多家国内银行和 100 多家券商的各年度年报，并对这些年报进行了深度清洗与分析，提供基本信息、词频统计、语调分析和可读性等重要文本分析结果。

1.6. 文本信息

1.6.1. 基本信息

该模块涵盖了海量文本的一些基本信息，例如，文本 PDF 大小、页数和字数等。具体文本挖掘信息包括：证券代码、证券简称、统计日期、会计年度、文本类型、报告大小、总页数、总字数和总词数等。

1.6.2. 词频统计

这一模块收录了上亿条词汇在不同文本中的词频信息。系统默认只可以输入中文关键字，并且关键词只针对两个及两个以上字符的词汇(纯数字除外)。具体文本挖掘信息包括：证券代码、证券简称、统计日期、会计年度、文本类型、关键字、词频。

用户可以通过在搜索框中输入关键词获得不同文本中的关键词出现次数。如果有较为少见的关键词未收录在该模块中，用户可以在机器学习模块中通过自定义文本和关键词的方式进行分析。

1.6.3. 可读性

这一模块结合中文文本可读性的已有研究，使用一些较常用的指标对可读性进行衡量。具体文本挖掘信息包括：证券代码、证券简称、统计日期、会计年度、文本类型、总页数、总字数、总词数、专业词密度、句子数、分句中的平均字数、副词和连词的比例。

2. 自由搜索

在一份给定的文件里，词频(Term frequency，TF)指的是某一个给定的词语在该文件中出现的次数。词频可以用来评估一个词对于一个文件或者一个语料库中的一个领域文件集的重复程度，词频统计为学术研究提供了新的方法和视野。

在这一模块用户可以通过在搜索框中输入关键词获得不同文本中的关键词出现次数。如果有较为少见的关键词未收录在该模块中，用户可以在机器学习模块中通过自定义文本和关键词的方式进行分析。用户可以下载系统收录的不同类型的文本，并统计词频信息。

系统默认只可以输入中文关键字，并且关键词只针对两个及两个以上字符的词汇(纯数字除外)。具体文本挖掘信息包括：证券代码、证券简称、统计日期、会计年度、文本类型、关键字、词频。用户也可以输入关键词获得自定义文本中的关键词出现次数。

图 2 智能文本分析系统工作流程

3. 文本分析

文本分析模块提供了文本研究领域常见指标的自动统计，包括基本信息统计、自定义词频统计、自定义情感分析、可读性统计等。

3.1. 基本信息

该模块自动计算了用户自定义文本的一些基本信息，例如，文本 PDF 大小、页数和字数等。

3.2. 情感分析

文本情感分析作为自然语言处理领域和计算语言学的基本任务，近几年在工业界和学术界受到越来越多的关注。在文本情感分析问题中，基于词典的方法是一种较为常见的经典方法。词典法使用一个由正面和负面两类情感词组成的情感词典，通过统计文本中正负面情感词的数量来确定文本的情感倾向[3]。

在爱文本智能文本分析系统的情感分析模块，用户可以单独或批量上传用户自定义的文件，根据爱文本系统提供的情感词典或者用户自定义情感词典，计算积极/消极词汇数、TONE 等情感分析结果。

3.3. 可读性

在财经领域，财务报告可读性能够在一定程度上反映公司的信息披露质量，可读性差的财务报告代表了糟糕的信息环境，其对业绩持续性、分析师行为 、自愿性信息披露等均有较为重要的影响。对财经领域文本的可读性分析，有助于分析师或管理层通常有动机提供额外信息来帮助投资者理解公司的经营行为[2]。

在爱文本智能文本分析系统的可读性统计模块，用过可以选择自定义的文本或文本集合，系统自动统计一些较常用的可读性指标，并提供统计结果的 Excel 文档。

4. 文本挖掘

4.1. 文本处理

在进行文本分析的过程中，首要任务就是将文本转化为文本挖掘算法在计算机中可以处理的中间形式。爱文本智能文本分析系统提供了菜单式的文本处理模块，包括 PDF 转 TXT、分词、去停用词、特征词库构建、文本向量化表示等。

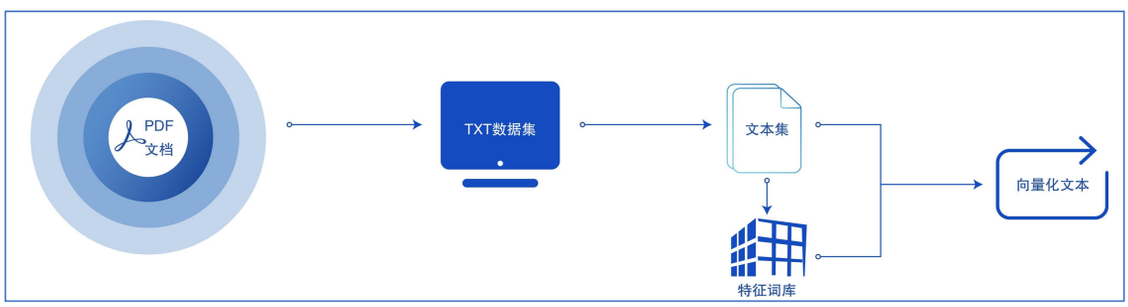


图 3 文本处理流程

在爱文本系统中，自定义文本表示的工作步骤主要包括以下三步(图 2)：①对文本集合中的文本分词、去停用词；②在第一步得到的文本集合中建立特征词库；③将文本集合和特征词库作为输入项，得到特征权重表示的文本集合。

4.1.1. 数据读取

在财经领域，常见的文本包括年报、分析师报告、招股说明书、银行年报等，这些报告通常存储为 PDF、DOC、DOCX 等格式。爱文本智能文本分析系统提供了将 PDF、DOC、DOCX 直接转换为 TXT 的功能，用户可以通过输入文件或文件夹，直接读取 PDF、DOC、DOCX 的内容，转换成方便计算机处理的 TXT 格式。

4.1.2. 分词

中文分词是中文文本处理的一个基础步骤，也是爱文本智能文本分析系统的基础模块。不同于英文的是，中文句子中没有词的界限，因此在进行中文自然语言处理时，通常需要先进行分词，分词效果将直接影响文本分类、文本聚类等模块的效果。在系统中，单独或批量上传用户自定义的文件，可以得到全模式、精准模式和搜索引擎模式等多种分词结果。

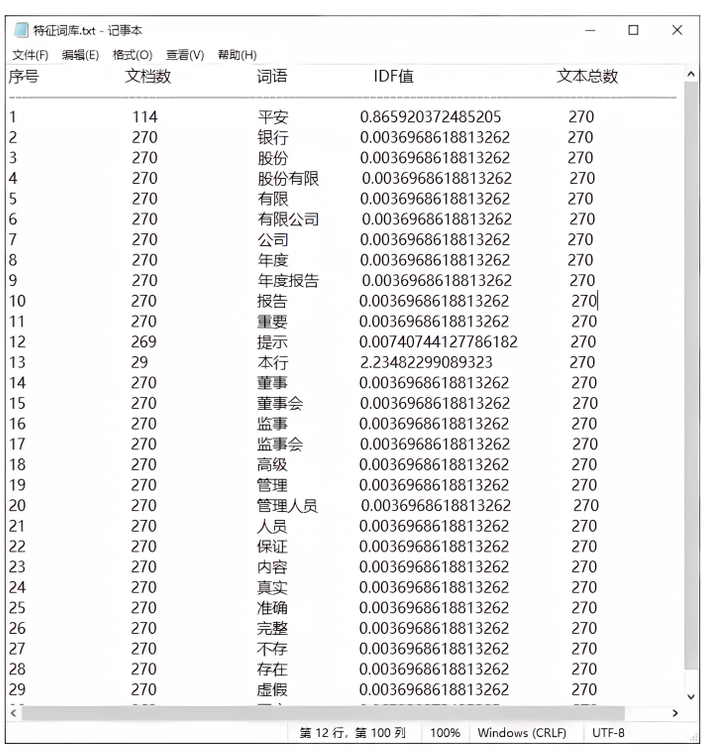
4.1.3. 去停用词

自然语言处理过程中，原始语料库存在许许多多的停用词。停用词(Stopwords)是指在文中出现频率较高但自身并无明确意义的词，比如，中文常见的“的、得、了、例如、吗、 大约、这、地”等词，这些词主要是副词、介词、连接词、助词等。

去停用词不但可以节省计算机的存储空间、提高执行效率，而且能在一定程度上使文本关键词更加集中、突出，文本语义表达更加明确，进一步提高正在执行任务的效果。在爱文本智能文本分析系统中，用户可以使用系统提供的停用词表去停用词，也可以通过自定义词表的方式去停用词。

4.1.4. 特征词库

在分词、去停用词文本集合中建立特征词库是文本向量化的基础工作。这部分工作内容包括从文本中提取关键词列表和统计每个词的文档频率两个部分。图 3 为爱文本构建的特征词库示例，包括 5 个字段，分别是：①序号，特征词的编号；②文档数，包括特征词的文档数目；③词语，特征词；④IDF 值，特征词的逆向文档频(公式 3-1)；⑤文本总数，文本集合中的文档总数。


特征词库.txt - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

序号	文档数	词语	IDF值	文本总数
1	114	平安	0.865920372485205	270
2	270	银行	0.0036968618813262	270
3	270	股份	0.0036968618813262	270
4	270	股份有限	0.0036968618813262	270
5	270	有限	0.0036968618813262	270
6	270	有限公司	0.0036968618813262	270
7	270	公司	0.0036968618813262	270
8	270	年度	0.0036968618813262	270
9	270	年度报告	0.0036968618813262	270
10	270	报告	0.0036968618813262	270
11	270	重要	0.0036968618813262	270
12	269	提示	0.00740744127786182	270
13	29	本行	2.23482299089323	270
14	270	董事	0.0036968618813262	270
15	270	董事会	0.0036968618813262	270
16	270	监事	0.0036968618813262	270
17	270	监事会	0.0036968618813262	270
18	270	高级	0.0036968618813262	270
19	270	管理	0.0036968618813262	270
20	270	管理人员	0.0036968618813262	270
21	270	人员	0.0036968618813262	270
22	270	保证	0.0036968618813262	270
23	270	内容	0.0036968618813262	270
24	270	真实	0.0036968618813262	270
25	270	准确	0.0036968618813262	270
26	270	完整	0.0036968618813262	270
27	270	不存	0.0036968618813262	270
28	270	存在	0.0036968618813262	270
29	270	虚假	0.0036968618813262	270

第 12 行，第 100 列 100% Windows (CRLF) UTF-8

图 4 特征词库示例

4.1.5. 文本向量化

文本在计算机中存储和表示的方式，也会对文本挖掘任务产生较大影响。爱文本智能文本分析系统提供了文本挖掘中常见的几种特征权重计算模块，分别是 TF-IDF 权重(TF-IDF Weighting)、TFC 权重(TFC Weighting)、LTC 权重(LTC Weighting)和熵权重(Entropy Weighting)。

1) TF-IDF 权重

TF-IDF 权重是一种使用非常广泛的权重表示方式，它考虑了词的文档频率信息，最初在信息检索中使用。TF-IDF(term Frequency-Inverse document Frequency) 权重，以词的逆向文档频数(Inverse Document Frequency: IDF)对词的词频作加权处理。其基本思想是：词在文档中出现的次数越多就越重要；同时也认为词的文档频数越大，该词的重要性就越低。

$$\mathrm{a}_{ij} = tf_{ij} \times \log(N / n_j) \tag{3-1}$$

在上述公式中，当$\mathrm{N} = n_j$，权重为 0。在小数据集上经常会发生这种情况，为了防止这种情况发生，一般要做平滑处理，如下式表示：

$$\mathrm{a}_{\mathrm{ij}} = (tf_{ij} + 1) \times \log((N + 1) / n_j) \tag{3-2}$$

2) TFC 权重

TF-IDF 权重没有考虑文本长度的不同对词权重的影响。为消除文本长度对权重的影响，TFC 权重对 TF-IDF 权重进行了“归一化”处理，可以使每个文本的特征向量都变成长度为 1 的单位向量。

$$\mathrm{a}_{ij} = \frac{(tf_{ij} + 1) \times \log((N + 1) / n_j)}{\sqrt{\sum_{p=1}^{M} [(tf_{pj} + 1) \times \log((N + 1) / n_p)]^2}} \tag{3-3}$$

3) LTC 权重

LTC 权重是 TF-IDF 权重的一种变形形式，可看作是对式 2-2 进行了归一化处理。

$$a_{ij}=\frac{\left(tf_{ij}+1\right)\times\log\left(N/n_j\right)}{\sqrt{\sum_{p=1}^{M}\left[\left(tf_{pj}+1\right)\times\log\left(N/n_p\right)\right]^2}} \tag{3-4}$$

或

$$a_{ij}=\frac{\left(tf_{ij}+1\right)\times\log\left((N+1)/n_j\right)}{\sqrt{\sum_{p=1}^{M}\left[\left(tf_{pj}+1\right)\times\log\left((N+1)/n_p\right)\right]^2}} \tag{3-5}$$

4) 熵权重

熵权重又称熵加权法，是基于信息论的加权算法，相对较为复杂。权重表示如下：

$$a_{ij}=\log\left(tf_{ij}+1\right)\left(1+\frac{1}{\log\left(N\right)}\sum_{p=1}^{N}\left[\frac{tf_{pj}}{n_p}\log\left(\frac{tf_{pj}}{n_p}\right)\right]\right) \tag{3-6}$$

其中 $\frac{1}{\log(N)}\sum_{p=1}^{N}\left[\frac{tf_{pj}}{n_p}\log\left(\frac{tf_{pj}}{n_p}\right)\right]$ 表示词 i 的平均不确定或熵。

4.2. 机器学习

当完成文本处理后，文本就被表示成可以被相关方法处理的中间形式了，此时就可以利用数据挖掘、机器学习、模式识别、自然语言处理等领域中的方法挖掘面向特定应用目标的知识或模式。文本挖掘是指为了发现知识，从文本数据中抽取隐含的、以前未知的、潜在有用的知识的过程。它是一个分析文本数据，提取文本信息，进而发现文本知识的过程[1]。

图 5 文本挖掘流程

爱文本智能文本分析系统集成了常见的文本挖掘模型(图 4)，包括相似度计算、文本分类、文本聚类、主题模型等模块。这些模块通过菜单式的操作方式，避免了繁琐的编程工作，供广大文本挖掘兴趣爱好者研究和学习。

4.2.1. 相似度

在自然语言处理中，相似度是一种非常有用的工具，可以帮助研究者解决很多问题。爱文本智能文本分析系统提供了文本相似度计算模块，这一模块实现了三种常见的文本相似度计算方式，分别是：欧式距离、余弦距离和 Jacard 距离。

文本相似度的计算是建立在文本向量化表示的基础上，爱文本智能文本分析系统提供了文本挖掘中常见的几种特征权重计算模块，分别是 TF-IDF 权重、TFC 权重、LTC 权重和熵权重。

1) 欧氏距离

欧几里得度量(Euclidean Metric)也称欧氏距离，是一个通常采用的距离定义，指在 n 维空间中两个点之间的真实距离，或者向量的自然长度(即该点到原点的距离)。在二维和三维空间中的欧式距离就是两点之间的实际距离。

设特征空间 χ 是 n 维实数向量空间 R^n，$x_i, x_j \in \chi$，$x_i = \left(x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(n)}\right)^T$，$x_j = \left(x_j^{(1)}, x_j^{(2)}, \cdots, x_j^{(n)}\right)^T$，$x_i, x_j$ 的 L_p 距离定义为：

$$L_p\left(x_i, x_j\right) = \left(\sum_{l=1}^{n} \left|x_i^{(l)} - x_j^{(l)}\right|^p\right)^{\frac{1}{p}} \tag{3-7}$$

这里 $p \geq 1$,当 $p = 2$ 时，称为欧式距离(Euclidean distance)，即

$$L_p\left(x_i, x_j\right) = \left(\sum_{l=1}^{n} \left|x_i^{(l)} - x_j^{(l)}\right|^2\right)^{\frac{1}{2}} \tag{3-8}$$

2) 余弦距离

余弦相似度，又称为余弦相似性，是通过计算两个向量的夹角余弦值来评估它们的相似度。它等于两个向量的点积(向量积)除以两个向量长度(或大小)的乘积。

设 n 维空间有两个向量 $\vec{a} = \left(x_1, x_2, \cdots, x_n\right)$ 和 $\vec{b} = \left(y_1, y_2, \cdots, y_n\right)$，$\left\|\vec{a}\right\|$ 和 $\left\|\vec{b}\right\|$ 表示向量 a 和 b 的大小，它们的夹角为 $\theta\left(0 \leq \theta \leq \pi\right)$，则余弦相似度定义为：

$$\text{Similarity}=\cos(\theta)=\frac{\vec{a}\bullet\vec{b}}{\|\vec{a}\|\times\|\vec{b}\|}=\frac{\sum_{i=1}^{n}x_i\times y_i}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\sqrt{\sum_{i=1}^{n}(y_i)^2}} \tag{3-9}$$

夹角余弦值得范围从-1 到 1：-1 意味着两个向量指向的方向正好截然相反；越接近于 1，表示两个向量越相似；越接近于 0，表示两个向量不相似。

3) Jacard 距离

Jaccard 距离是用来衡量两个集合差异性的一种指标，它是 Jaccard 相似系数的补集，被定义为 1 减去 Jaccard 相似系数。而 Jaccard 相似系数(Jaccard similarty coefficient)，也称 Jaccard 指数(Jaccard Index)，是用来衡量两个集合相似度的一种指标。设 $x=(x_1,x_2,\cdots,x_n)$ 与 $y=(y_1,y_2,\cdots,y_n)$ $(x_i,y_i\geq 0)$ 是两个实向量，则它们之间的 Jaccard 相似系数定义为：

$$J(x,y)=\frac{\sum_{i=1}^{n}\min(x_i,y_i)}{\sum_{i=1}^{n}\max(x_i,y_i)} \tag{3-10}$$

Jaccard 距离定义为：

$$d(x,y)=1-J(x,y) \tag{3-11}$$

4.2.2. 文本分类

文本分类是爱文本智能文本分析系统的核心模块。这一模块的功能包括两个阶段(参见图 5)：①设计阶段。批量上传自定义训练集，在数据集上构造文本分类模型。②应用阶段。使用模型预测测试集文件和预测集文件的类别，计算精确率、召回率、PR 曲线与 F1-Score 等多个分类结果。

在爱文本智能文本分析系统的文本分类模块，会将数据分为三大部分，分别是训练集、测试集、预测集。其中，训练集用于模型构建；测试集用于评估模型的准确率；构建好的模型在预测集中预测每个文本的类别。

在进行文本分类操作之前，需要对文本进行向量化表示，爱文本智能文本分析系统提供了文本挖掘中常见的文本向量化计算模块，分别是 TF-IDF 权重、TFC 权重、LTC 权重和熵权

重。

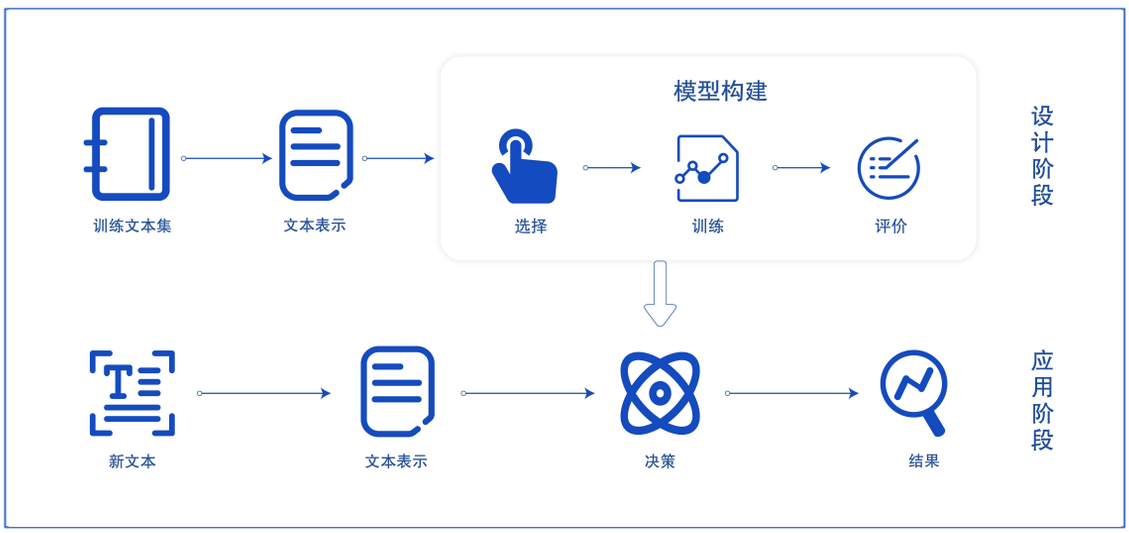


图 6 文本分类模块工作流程

目前，大量的基于统计学习、模式识别、机器学习的算法在文本分类中得到了广泛的研究和应用，这些方法已经成为当前文本分类技术的主流技术。爱文本智能文本分析系统文本分类模块实现了常用的文本分类算法：支持向量机、Rocchio 算法、决策树、K 近邻、朴素贝叶斯等。下面简单介绍这几种分类算法。

1) 支持向量机

支持向量机(Support Vector Machine, SVM)已经成为倍受关注的分类技术。基于结构风险最小化原则，SVM 通过求解最优分隔超平面来得到高分类准确率的分类器(图 6)。

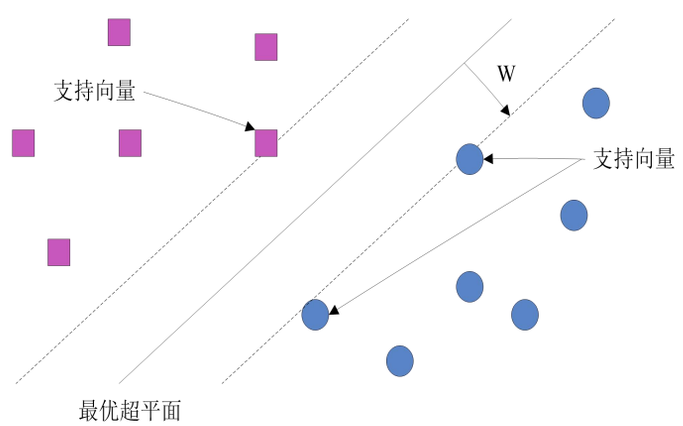


图 7 线性可分的支持向量机

考虑有数据集 $D = \{x_i, y_i\}, i = 1,2, \ldots, N$, N 为样本总数，$x_i \in R^p \subset R$, x_i是 p 维向量，$y_i \in \{-1,1\}$是二分类问题中的类标。在分类问题中，SVM 尝试找到最小化期望分类误差的分类

器 f(x)。线性分类器 f(x)是一个可以表示成 $f(x) = sgn(w^T x + b)$的超平面。找到 SVM 的最优分类器 f(x)的过程等同于优化如下公式(1)中的凸二次规划问题：

$$\max_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \tag{3-12}$$

$$Subject\ \ to \tag{3-13}$$

$$y_i(<w, x_i> + b) \geq 1 - \xi_i \quad (\xi_i \geq 0, i = 1, ..., N)$$

其中，C 是正规化参数，用于平衡分类器在数据集 D 中的时间复杂度与分类准确率。上述二次规划问题可以通过对偶函数求解。基于核方法，用核函数取代上述公式中的内积，可以将线性 SVM 转换成更为复杂的非线性 SVM。一些典型的核函数如下：

$$\text{Lin: } k(x_i, x_j) = x_i^T x_j \tag{3-14}$$

$$\text{Poly: } k(x_i, x_j) = (x_i^T x_j + 1)^d \tag{3-15}$$

$$\text{RBF: } k(x_i, x_j) = \exp\left(-\frac{||x_i - x_j||_2^2}{2\sigma^2}\right) \tag{3-16}$$

2) Rocchio 算法

Rocchio 算法[4]来源于向量空间模型理论，在向量空间法中，每个文档被看成一个词袋(也就是不考虑词项顺序关系)，然后被表示成词项权重的向量：$d_i = (w_{i,1}, w_{i,2}, \cdots, w_{i,n})$，其中 d_i 表示文档 i，n 表示词项空间的维数，$w_{i,k}$ 表示文档词项 k 在文档 i 中的权重，该权重表示该词项在文档中的重要程度，通常使用 tf-idf 方法或者其他权重表示方法。两文档的相似度可使用对应文档向量的余弦夹角计算（也可用其他方法如欧氏距离等）：$Sim(d_i, d_j) = \left(\sum_{k=1}^{n} w_{i,k} \times w_{j,k}\right) \Big/ \sqrt{\left(\sum_{k=1}^{n} w_{i,k}^2\right)\left(\sum_{k=1}^{n} w_{j,k}^2\right)}$。首先训练分类模型时，每个类使用一个中心向量(Centroid)代表,然后在分类时通过检查待分类文档和这些中心向量的相似度，把它分到最相似的中心向量所代表的类中。令 $C = \{C_i\}_{i=1}^{m}$ 表示预定义的类别集合，每个类中包含所属该类的文档集合。Rocchio 分类算法由图 7 给出。

训练过程：

对每个类C_i，通过计算该类中所有文档向量的算术平均值得到该类的中心向量$R(i)$: $R(i)=\sum_{k=1}^{\#C_i} d_{i,k} / \#C_i$，$d_{i,k}$表示类 i 中的第 k 个文档，$\#C_i$表示类 i 中的文档数。

分类过程：

对于特定的待分类文档 d

1:计算文档 d 与各类中心向量相似值$Sim(d, R(i)), i=1,\cdots,m$。

2:返回$\hat{c}(d)=\arg\max_{C_i \in C} Sim(d, R(i))$。

图 8 Rocchio 算法流程

3) 决策树

决策树(Decision Tree)[5]，是一种较早广泛应用于机器学习中的算法，它是一种基于规则的分类器，对噪声数据有很好的健壮性且能够学习析取表达式。它采用“分而治之”的策略，通过学习，自顶向下构造一棵决策树。目前存在 ID3、C4.5、C5.0 等广为应用的决策树算法。

决策树通过把实例从根结点排列到某个叶子节点来分类实例，分支表示特征到不同状态的权重，叶子节点为实例所属的分类。树上的每一个结点说明了对实例的某个属性的测试，并且该结点的每一个后继分支对应于该属性的一个可能值。分类实例的方法是从这棵树的根结点开始，测试这个结点指定的属性，然后按照给定实例的该属性值对应的树枝向下移动。然后这个过程在以新结点为根的子树上重复，直到达到叶子结点为止。

在建立决策树时，选择节点的依据是特征含有的信息量，常用的有信息增益、信息增益率、信息熵等。决策树模型通过构造树来解决分类问题。首先利用训练数据集来构造一棵决策树，一旦树建立起来，它就可为未知样本产生一个分类。在分类问题中使用决策树模型有很多的优点，决策树便于使用，而且高效； 根据决策树可以很容易地构造出规则，而规则通常易于解释和理解；决策树可很好地扩展到大型数据库中，同时它的大小独立于数据库的大小；决策树模型的另外一大优点就是可以对有多种属性的数据集构造决策树。决策树模型也有一些缺点，比如处理缺失数据时的困难，易出现过度拟合问题，因为训练集样例仅仅是所有可能实例的一部分；向决策树增加分支，可以提高其在训练集上的性能，但很可能会降低在训练集外的其他实例上的性能；另外，它还存在会忽略数据集中属性之间的相关性等问题。在决策树的学习过程中，后修剪决策树的技术对于避免决策树学习中的过度拟合是很重

要的。

4) K 近邻

K 近邻分类算法(KNN)[6]是一种懒散的方法，即它没有学习过程，只是存放所有的训练例直到接到未知文本的时候才建立分类。对于每个待分类文档，首先计算其于训练集中所有文档的相似度，然后按照相似度从大到小选择前 K 个文档(k 近邻)，最后返回包含这 k 个文档中所含最多类的类标签。令 $\mathrm{C}=\{\mathrm{C_i}\}_{\mathrm{i=1}}^{\mathrm{m}}$ 表示预定义的类别集合，每个类中包含所属该类的文档集合，l(d)表示文档 d 的类标签，KNN 算法由图 8 给出。

训练过程：

对每个训练实例 $(d,l(d))$，将它加入训练实例表中。

分类过程：

对于给定的待分类文档 d：

1:通过计算文档 d 与训练实例表中各个文档的相似度得到前 k 个最相似的邻居文档 $d_1,d_2,\cdots d_k$。

2:返回 $\hat{c}(d)=\arg\max_{C_i\in C}\sum_{j=1}^{K}\delta(l(C_i),l(d_j))$，其中 $l(C_i)$ 表示 C_i 的类文档标签；

当 $a=b$， $\delta(a,b)=1$；否则 $\delta(a,b)=0$。

图 9 KNN 算法流程

5) 朴素贝叶斯

朴素贝叶斯(Naïve Bayes Classification)[7]是一种典型的基于概率统计分类算法，其数学基础是贝叶斯定理，也因为其简化和高效成为贝叶斯方法中最常用的一种。其主要思想就是计算在给定一待分类文档的条件下，其属于各个类别的条件概率,选择条件概率最高的类别为其标类别。朴素贝叶斯分类算法的一个前提假设是：在给定的文本集下，文本特征属性间是相互独立的，即构造特征向量的各个特征相互独立。其主要分类算法由图 9 给出。

训练过程：

计算特征词属于每个类别 C_j 的概率 $p(w_i|C_j)$ ，

$$p(w_i|C_j)=\frac{\sum_{k=1}^{|D|}N(w_i,d_k)+1}{\sum_{s=1}^{|V|}\sum_{k=1}^{|D|}N(w_s,d_k)+|V|}$$

其中,$|D|$ 表示类的文档数,$|V|$ 表示特征词表中的总单词数, $\sum_{k=1}^{|D|}N(w_i,d_k)$ 表示特征词 w_i 出现在类 C_i 文档 d_k 中的次数, $\sum_{s=1}^{|V|}\sum_{k=1}^{|D|}N(w_s,d_k)$ 表示 C_j 类文档中出现的所有特征词的总次数.

分类过程：

计算测试文本 d 属于类 C_i 的概率 P(Ci/ d) ,将其分到概率最大的类别中，

$$\hat{c}(d)=\arg\max_{C_i\in C}P(C_i)\prod_{j=1}^{m}P(w_j|C_i)$$

其中 $P(C_i)$ 为类 C_i 的先验概率, m 为特征项数目。

图 10 朴素贝叶斯算法流程

6) 评价指标

在文本分类器完成了训练和测试之后一个很重要的问题就是进行分类性能评估。研究分类算法的优劣就需要选择合适的评价指标对其分类结果进行评估，并且和其它算法的性能进行比较。下面简单几种介绍常用的评价指标。首先，我们做以下一些约定：

a: 正例测试文档被正确分类为属于该类的数量；

b: 负例测试文档被错误分类为属于该类的数量；

c: 正例测试文档被错误分类为不属于该类的数量；

d: 负例测试文档被正确分类为不属于该类的数量。

那么，我们可以得出以下几个指标:精确率(Precision)是分类系统分类结果与人工分类结果一致的文档在被分文档中的比率。

$$precision=a/(a+b) \tag{3-17}$$

召回率(Recall)是人工分类结果应该有的文档与分类系统一致的文档所占的比率。

$$recall=a/(a+c) \tag{3-18}$$

我们还可以定义以下几个标准：

$$fallout = b/(b+d) \tag{3-19}$$

$$accuracy = (a+b)/(a+b+c+d) \tag{3-20}$$

$$error = (b+c)/(a+b+c+d) \tag{3-21}$$

另外，常用的有 F1 测试值，它综合考虑准确率和召回率，也称为综合分类率，计算公式如下:

$$\mathrm{F1} = \frac{2*\mathrm{precision}*\mathrm{recall}}{\mathrm{precision}+\mathrm{recall}} \tag{3-22}$$

4.2.3. 文本聚类

文本聚类就是根据在文档数据集中文档表示形式及文档间关系的信息，将文档对象分组的过程。其目标就是使聚类后组内的文档之间是相似的，组间的文档是相异的。组内的相似性(同质性)越大，组间差别越大，聚类效果就越好。爱文本智能文本分析系统提供了两种常见的文本聚类算法，分别是 K-Means 聚类和 DBSCAN 聚类算法。

在进行文本聚类之前，需要对文本进行向量化表示，转化为文本聚类算法在计算机中可以处理的中间形式。爱文本智能文本分析系统提供了文本挖掘中常见的文本向量化计算模块，分别是 TF-IDF 权重、TFC 权重、LTC 权重和熵权重。

1) K-Means 聚类

K 均值[8]基本原理是首先选择 k 个文档作为初始的聚类点，然后根据簇中对象的平均值，将每个文档(重新)赋给最类似的簇,并更新簇的平均值，然后重复这一过程，直到簇的划分不再发生变化。K 均值的算法复杂度为 O (k*l*n) ,其中 l 为迭代次数，n 为文档个数，k 为类别个数(图 10)。

算法：K 均值聚类算法

1:选择 K 个点作为原始质心。

2:repeat

3:将每个点指派到最近的质心，形成 K 个簇。

4:重新计算每个簇的质心。

5:until 质心不发生变化

图 11 K 均值聚类算法

由上述算法可知，K 均值效率高，能有效处理大文本集。K 均值算法本质上是一种贪心算法。可以保证局部最小，但是很难保证全局最小。另外该方法需要预先指定 k 值和初始划分，从而容易使聚类结果受到影响，这就是它的最大的缺点。为此人们提出了一些相应的解决方法。K 中心点使用中心点定义原型，其中中心点是一组点中最有代表性的点。这种算法对于异常数据不敏感，但计算量显然要比 K 均值要大，一般只适合小数据量。

2) DBSCAN 聚类算法

DBSCAN 聚类算法是一种基于密度的带噪声的聚类算法，该算法可以对任意形状和大小的簇聚类，且聚类结果与数据对象的输入顺序无关，具有良好的抗干扰性[9]。算法的关键思想是对于簇中的每个数据对象，给定 Eps 的邻域内必须包含至少 MinPts 个数据对象。DBSCAN 聚类算法的相关定义如下[10]。

定义 1：(Eps 邻域)对象 p 的领域是指以 p 为圆心、Eps 为半径的圆中数据对象的集合，即 $N_{Eps}(p)=\{q \in S | dist(p,q) \le Eps\}$。

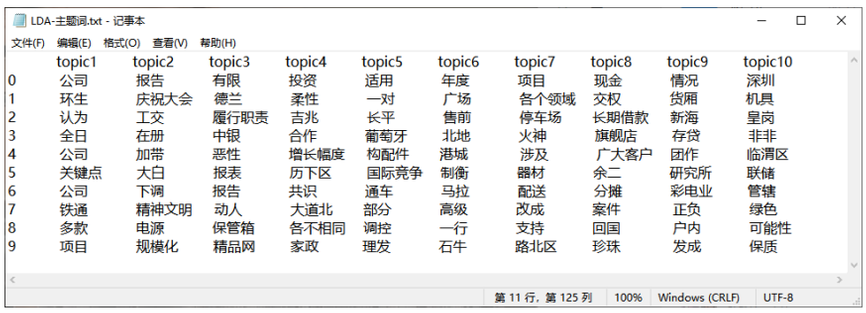
定义 2：(核心对象)如果对象 p 的密度 ρ 大于或等于密度阈值 MinPts，则 p 为核心对象，MinPts 阈值由用户指定。

定义 3：(直接密度可达)如果对象 p 在以对象 q 为核心对象的 Eps 邻域内，则对象 p 是从对象 q 直接密度可达的。

DBSCAN 算法先从数据集中任选一个未被处理过的点，如果该点 Eps 邻域内数据点的数量大于等于 MinPts，则该点被记为核心点；如果该点 Eps 邻域内的数据点的数量小于 MinPts，且所考虑的点处于一个核心点的邻域内，则该点被标记为边界点，否则被认为是噪声点。第一个核心点将形成一个簇，并将所有与核心点直接密度可达的未归类的点添加到此簇中，直到没有符合要求的点可以添加。算法随机选择未访问的数据点，聚类过程持续进行，直到所有的数据点都被访问，没有新的点可以添加到任何簇中。

4.2.4. 主题模型

爱文本智能文本分析系统实现了 LDA 主题模型，输入为文本集合，输出为不同主题的主题词(图 11)和不同文本的主题分布(图 12)。



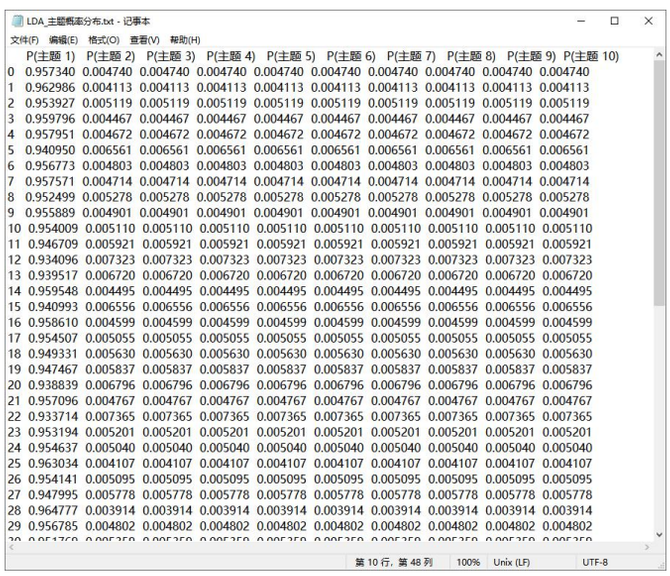
LDA-主题词.txt - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

	topic1	topic2	topic3	topic4	topic5	topic6	topic7	topic8	topic9	topic10
0	公司	报告	有限	投资	适用	年度	项目	现金	情况	深圳
1	环生	庆祝大会	德兰	柔性	一对	广场	各个领域	交权	货厢	机具
2	认为	工交	履行职责	吉兆	长平	售前	停车场	长期借款	新海	皇岗
3	全日	在册	中银	合作	葡萄牙	北地	火神	旗舰店	存贷	非非
4	公司	加带	恶性	增长幅度	构配件	港城	涉及	广大客户	团作	临渭区
5	关键点	大白	报表	历下区	国际竞争	制衡	器材	余二	研究所	联储
6	公司	下调	报告	共识	通车	马拉	配送	分摊	彩电业	管辖
7	铁通	精神文明	动人	大道北	部分	高级	改成	案件	正负	绿色
8	多款	电源	保管箱	各不相同	调控	一行	支持	回国	户内	可能性
9	项目	规模化	精品网	家政	理发	石牛	路北区	珍珠	发成	保质

第 11 行，第 125 列 100% Windows (CRLF) UTF-8

图 12 主题关键词



LDA_主题概率分布.txt - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

	P(主题 1)	P(主题 2)	P(主题 3)	P(主题 4)	P(主题 5)	P(主题 6)	P(主题 7)	P(主题 8)	P(主题 9)	P(主题 10)
0	0.957340	0.004740	0.004740	0.004740	0.004740	0.004740	0.004740	0.004740	0.004740	0.004740
1	0.962986	0.004113	0.004113	0.004113	0.004113	0.004113	0.004113	0.004113	0.004113	0.004113
2	0.953927	0.005119	0.005119	0.005119	0.005119	0.005119	0.005119	0.005119	0.005119	0.005119
3	0.959796	0.004467	0.004467	0.004467	0.004467	0.004467	0.004467	0.004467	0.004467	0.004467
4	0.957951	0.004672	0.004672	0.004672	0.004672	0.004672	0.004672	0.004672	0.004672	0.004672
5	0.940950	0.006561	0.006561	0.006561	0.006561	0.006561	0.006561	0.006561	0.006561	0.006561
6	0.956773	0.004803	0.004803	0.004803	0.004803	0.004803	0.004803	0.004803	0.004803	0.004803
7	0.957571	0.004714	0.004714	0.004714	0.004714	0.004714	0.004714	0.004714	0.004714	0.004714
8	0.952499	0.005278	0.005278	0.005278	0.005278	0.005278	0.005278	0.005278	0.005278	0.005278
9	0.955889	0.004901	0.004901	0.004901	0.004901	0.004901	0.004901	0.004901	0.004901	0.004901
10	0.954009	0.005110	0.005110	0.005110	0.005110	0.005110	0.005110	0.005110	0.005110	0.005110
11	0.946709	0.005921	0.005921	0.005921	0.005921	0.005921	0.005921	0.005921	0.005921	0.005921
12	0.934096	0.007323	0.007323	0.007323	0.007323	0.007323	0.007323	0.007323	0.007323	0.007323
13	0.939517	0.006720	0.006720	0.006720	0.006720	0.006720	0.006720	0.006720	0.006720	0.006720
14	0.959548	0.004495	0.004495	0.004495	0.004495	0.004495	0.004495	0.004495	0.004495	0.004495
15	0.940993	0.006556	0.006556	0.006556	0.006556	0.006556	0.006556	0.006556	0.006556	0.006556
16	0.958610	0.004599	0.004599	0.004599	0.004599	0.004599	0.004599	0.004599	0.004599	0.004599
17	0.954507	0.005055	0.005055	0.005055	0.005055	0.005055	0.005055	0.005055	0.005055	0.005055
18	0.949331	0.005630	0.005630	0.005630	0.005630	0.005630	0.005630	0.005630	0.005630	0.005630
19	0.947467	0.005837	0.005837	0.005837	0.005837	0.005837	0.005837	0.005837	0.005837	0.005837
20	0.938839	0.006796	0.006796	0.006796	0.006796	0.006796	0.006796	0.006796	0.006796	0.006796
21	0.957096	0.004767	0.004767	0.004767	0.004767	0.004767	0.004767	0.004767	0.004767	0.004767
22	0.933714	0.007365	0.007365	0.007365	0.007365	0.007365	0.007365	0.007365	0.007365	0.007365
23	0.953194	0.005201	0.005201	0.005201	0.005201	0.005201	0.005201	0.005201	0.005201	0.005201
24	0.954637	0.005040	0.005040	0.005040	0.005040	0.005040	0.005040	0.005040	0.005040	0.005040
25	0.963034	0.004107	0.004107	0.004107	0.004107	0.004107	0.004107	0.004107	0.004107	0.004107
26	0.954141	0.005095	0.005095	0.005095	0.005095	0.005095	0.005095	0.005095	0.005095	0.005095
27	0.947995	0.005778	0.005778	0.005778	0.005778	0.005778	0.005778	0.005778	0.005778	0.005778
28	0.964777	0.003914	0.003914	0.003914	0.003914	0.003914	0.003914	0.003914	0.003914	0.003914
29	0.956785	0.004802	0.004802	0.004802	0.004802	0.004802	0.004802	0.004802	0.004802	0.004802

第 10 行，第 48 列 100% Unix (LF) UTF-8

图 13 文档主题分布

LDA 主题模型是由 Blei 提出的一种无监督的主题概率生成模型，可以用来识别大规模文档集中的隐藏主题信息[11]。LDA 模型分为文档(doc)、主题(topic)和单词(word)三层：一个文档是由多个主题以不同的概率生成的，每个单词也是由多个主题以不同的概率生成的，文档-主题分布和主题-单词分布都符合 Dirichlet 分布。LDA 主题模型的概率表示如公式 3-1 所示：

$$P(word|doc)=\sum_{topic}P(word|topic)*P(topic|doc) \tag{3-23}$$

$P(word|doc)$可以通过对文档分词求得，是可观测值。LDA 生成文档的过程如下：

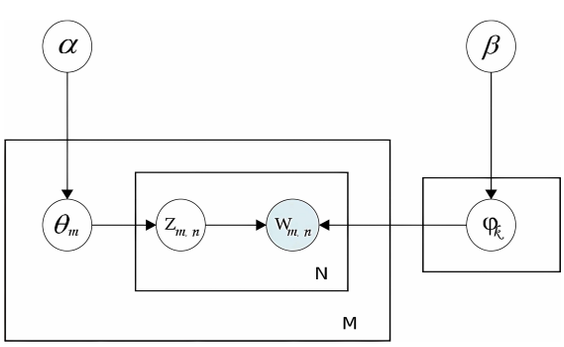


图 14 LDA 主题模型示意图

① 按照先验概率 $P(d_i)$ 选择一篇文档 d_i。

② 从超参数取值为 α 的 Dirichlet 分布中取样生成文档 d_i 的主题分布 θ_i。

③ 从主题的多项式分布 θ_i 中抽样生成文档 d_i 第 j 个词的主题 $z_{i,j}$。

④ 从超参数取值为 β 的 Dirichlet 分布中抽样生成主题 $z_{i,j}$ 相对应的词语分布 $\phi_{z_{.,j}}$。

⑤ 根据词语的多项式分布 $\phi_{z_{.,j}}$ 进行抽样，最终生成单词 $w_{i,j}$。

参考文献

1. Tan A H. Text mining: The state of the art and the challenges[C]//Proceedings of the pakdd 1999 workshop on knowledge disocovery from advanced databases. 1999, 8: 65-70.
2. 逯东, 余渡, 杨丹. 财务报告可读性,投资者实地调研与对冲策略[J]. 会计研究, 2019(10):8.
3. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. 2010. Lexicon-based Methods for Sentiment Analysis. Journal of Computational Linguists 2010.
4. Thorsten Joachims. A probabilistic analysis of the Rocchio algorithm with tf-idf for text categorization[C]//Proceedings of the Fourteenth International Conference on Machine Learning,July 1997:143-151.
5. Song Y Y, Ying L U. Decision tree methods: applications for classification and prediction[J]. Shanghai archives of psychiatry, 2015, 27(2): 130.
6. Zhang M L, Zhou Z H. ML-KNN: A lazy learning approach to multi-label learning[J]. Pattern recognition, 2007, 40(7): 2038-2048.
7. Saritas M M, Yasar A. Performance analysis of ANN and Naive Bayes classification algorithm for data classification[J]. International Journal of Intelligent Systems and Applications in Engineering, 2019, 7(2): 88-91.
8. Krishna K, Murty M N. Genetic K-means algorithm[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 1999, 29(3): 433-439.
9. Khan K, Rehman S U, Aziz K, et al. DBSCAN: Past, present and future[C]//The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014). IEEE, 2014: 232-238.
10. Al-mamory S O, Kamil I S. A New Density Based Sampling to Enhance DBSCAN Clustering Algorithm [J]. Journal of Computer Science, 2019,32(4).
11. Blei D M, Lafferty J D. Topic models[M]//Text mining. Chapman and Hall/CRC, 2009: 101-124.

爱文本·会思想的文本

分析海量文本，挖掘数据价值

联系我们

市场部

联系人：刘经斌

电话：（+86）021-66181082

手机：（+86）17821816737

邮箱：liujingbin@aitexts.com

地址：上海市移动互联网中心（华滋奔腾大厦）A座18层

扫一扫，联系我们

产品网址

https://www.aitexts.com