数据科学导论 —— 文本挖掘

王小宁

中国传媒大学数据科学与智能媒体学院

2021 年 6 月 16 日

中国传媒大学
COMMUNICATION UNIVERSITY OF CHINA

目录

文本挖掘概述

问题的提出

文本挖掘基本流程

文本挖掘概述

文本挖掘

- 在网络普及的时代，Web 页面、新闻消息以及电子邮件等都包含了大量的文本信息，这些文本信息构成了粗糙的非结构化文本数据。
- 文本挖掘（text mining）是以文本作为挖掘的对象，寻找信息的结构、模型、模式等隐含的具有潜在价值的知识的过程。
- 文本挖掘成长为数据科学领域备受关注的领域之一。

问题的提出

例子

- 考虑 corpus.JSS.papers 包中的 JSS_papers 数据集。repos = "http://datacube.wu.ac.at/"
- 该数据集提供了 Journal of Statistics Software (JSS) 期刊的摘要信息，包含标题（title)、创建者（creator)、主题（subject)、描述(description)、出版商（publisher）等 15 列 1046 行。该数据及是语料库（corpus）结构的数据集。那么，如何总结该杂志 2010-08-05 到 2014-05-06 所刊文章的基本内容，提炼出该杂志 1996 年到 2020 年 24 年间文章主题?

```
> tail(JSS_papers)
        title
[1041,] "spBayesSurv: Fitting Bayesian Spatial Survival Models Using R"
[1042,] "bridgesampling: An R Package for Estimating Normalizing Constants"
[1043,] "spBayesSurv: Fitting Bayesian Spatial Survival Models Using R"
[1044,] "Algebraic Analysis of Multiple Social Networks with multiplex"
[1045,] "Fitting Prediction Rule Ensembles with R Package pre"
[1046,] "Fitting Prediction Rule Ensembles with R Package pre"
        creator                       subject
[1041,] character,3                   character,2
[1042,] character,3                   character,2
[1043,] character,3                   character,2
[1044,] "Rivero Ostoic, J. Antonio"   character,2
[1045,] "Fokkema, Marjolein"          character,2
[1046,] "Fokkema, Marjolein"          character,2
        description
[1041,] "Spatial survival analysis has received a great deal of attention over the last 20 years due to the " [truncated]
[1042,] "Statistical procedures such as Bayes factor model selection and Bayesian model averaging require th" [truncated]
[1043,] "Spatial survival analysis has received a great deal of attention over the last 20 years due to the " [truncated]
[1044,] "multiplex is a computer program that provides algebraic tools for the analysis of multiple network " [truncated]
[1045,] "Prediction rule ensembles (PREs) are sparse collections of rules, offering highly interpretable reg" [truncated]
[1046,] "Prediction rule ensembles (PREs) are sparse collections of rules, offering highly interpretable reg" [truncated]
        publisher                                 contributor date         type        format
[1041,] "Foundation for Open Access Statistics"   ""          "2020-02-27" character,3 "PB"
```

文本挖掘基本流程

文本数据获取

- 文本数据的获取有多种途径。常见途径之一为网页文本的抓取，其逻辑为计算机模拟访问 URL 地址，获得其 HTML 源码或 Json 格式的字符串。
- 文本数据的抓取往往利用专门的爬虫工具，获取数据。
- 网络公开数据集。

文本特征表示

- 文本挖掘的首要任务是寻找合适的文本表示方式。
- 这种表示方式既要包含足够的信息以达到识别文本内容的目的，又要将文本信息这一非结构化数据转化为结构化数据以方便计算机处理。
- 这就涉及文本特征的抽取和选择。
- 文本特征指的是关于文本的元数据，可以分为描述特征，如文本的名称、日期、大小、类型；语义特征，如文本的作者、标题、机构、内容。

文本特征表示

- 当文本内容被简单地看成由它所包含的基本语言单位（字、词、词组或短语等）组成的集合时，这些基本的语言单位被称为词项(term)。
- 如果用出现在文本中的词项表示文本，即在词袋假设——文本所含内容只与其包含的词项以及词项出现情况有关而与词项出现的顺序无关，那么这些词项就是文本的特征。

文本特征表示

- 通过某种方式量化文本中抽取的特征词项，并用这些特征项以结构化的方式来表示文本的过程，称为文本表示。
- 对于文本内容的特征表示主要有布尔型、向量空间模型、概率模型和基于知识的表示模型。
- 布尔型和向量空间模型易于理解且计算复杂度较低，所以成为了文本表示的主要工具。

文本特征抽取

- 获得词项的基本方法为分词。
- 英文分词：英文文档 -> 停用词处理 -> 词干抽取 -> 特征词集合。
- 文档中常常包括一些使用频率极高、所含信息却非常少的词。
- 这些词所含的信息密度极低，并且它们的存在会增大词项数，增加分析维度，提高分析难度。常用这些词构造一个停用词表，在文本的特征抽取过程中删去停用词表中出现的特征词。

文本特征抽取

- 英文单词往往由于时态、词性、位置、主语等的不同而采用不同的形式。如 {connect，connected，connection，connects}，这些词要表达的都是“connect”。
- 为排除同义词不同形式造成的文本特征表达的复杂化，需对英文单词进行词干抽取（stemming）。
- 所谓的词干是由彼此互为语法变型的词组组成的非空词集的规范形式。在上例中非空词集为 V(s) ={connect，connected，connection，connects}，而词干 s = connect。
- 进行词干抽取后，将英文文档中所有的词干汇总即得到该文档的词项。

文本特征抽取

- 中文分词：中文文档 -> 停用词处理 -> 词语切分 -> 特征词集合。
- 中文文档的分词相比英文分词复杂。
- 常见的中文分词方法，包括最大匹配法、最大概率法、最少分词法、基于 HMM 的分词方法、基于互现信息的分词方法、基于字符标注的方法和基于实例的汉语分词方法等。

最大匹配法

- 最大匹配法需要一个词表，分词过程中用文中的候选词去跟此表中的词匹配。
- 如果匹配成功则认为候选词是词并予以切分；否则就认为不是词。
- 最大匹配法需设置最大词长。

例子

- 以字符串 S1 = “文本挖掘非常重要” 为例，简单介绍最大匹配法分词：

1. 设置最大词长以及初始空字符串 S2 = “ ”；
2. S2 = “ ”；S1 不为空，从 S1 左边取出候选子串；
3. W = “文本挖掘” (MaxLen = 4)；
4. 查词表，“文本挖掘” 在词表中，将 W 加入到 S2 中，S2 = “文本挖掘/”；
5. 并将 W 从 S1 中去掉，此时 S1 = “非常重要”；
6. S1 不为空，于是从 S1 左边取出候选子串 W = “非常重要”；
7. 查词表，W 不在词表中，将 W 最右边一个字去掉，得到 W = “非常重”；

8. 查词表，W 不在词表中，将 W 最右边一个字去掉，得到 W = “非常”；
9. 查词表，W 在词表中，将 W 加入到 S2 中；
10. S2 = “文本挖掘/非常/”，
11. 并将 W 从 S1 中去掉，此时 S1 = “重要”；

...

重复上述过程，直到 S1 为空，最后得到 S2 的分词结果为：“文本挖掘/非常/重要”。

最大概率法

- 假设 $Z = z_1 z_2 \cdots z_n$ 是输入的汉字串，$W = w_1 w_2 \cdots w_n$ 是与之对应的可能词串，汉语自动分词可以看做是求解使得条件概率 $P(W|Z)$ 最大的词串，即

$$W = \arg\max_W P(W|Z)$$

- 根据贝叶斯公式可得：

$$W = \arg\max_W P(W|Z) = \arg\max_W \frac{P(W)P(Z|W)}{P(Z)}$$

- 其中 $P(Z)$ 是字符串概率与 W 无关，$P(Z|W)$ 是词串到汉字串的条件概率，在已知词串的情况下，出现相应汉字串的概率为 1。因此上式可化简为：

$$W = \arg\max_W P(W)$$

文本特征表示

- 词串概率可用 n 元语法来求。如用二元语法则：

$$P(W) = \prod_{i=1}^{m} p(w_i|w_{i-1})$$

- 其中 w_i 为第 i 个词，w_0 为虚设的串首词。如果采用一元语法规则：

$$P(W) = \prod_{i=1}^{m} p(w_i)$$

- 以一元语法为例，算法的基本思想是：根据词表把输入串中所有可能的词都找出来，然后把所有可能的切分路径都找出来，并从这些路径中找出一条最佳（即概率最大的）路径作为输出结果。

文本特征表示

- 通过分词技术获得词项后，可通过向量空间模型（Vector Space Model，VSM）以及布尔模型（Bool Model，BM）将文本信息结构化。
- 向量空间模型和布尔模型都是将文档集合矩阵化的过程。布尔模型采用布尔矩阵将文档集合矩阵化，向量空间模型采用词频矩阵将文档集合矩阵化。

项权重（term weight）

- 令 m 表示文本集中词项的数量，f_i 表示第 j 个词项，$F = \{f_1, f_2, \cdots, f_m\}$ 是项集。
- 文档 d_i 中的词项 f_i 的权重 w_{ij} 为项权重。
- 当 f_j 在 d_i 中出现时 $w_{ij} > 0$, 否则 $w_{ij} = 0$。
- 在布尔矩阵中 $w_{ij} \in \{0, 1\}$, 而词频矩阵中 w_{ij} 可用绝对词频和相对词频表示。

文本的特征表示

- 绝对词频：词项在文本中出现的次数。利用绝对词频表示文本时，文本向量的第 j 个词项的项权重为其在文中出现的次数。
- 相对词频：词项在文本中的权重。文本向量的每一项权重用词项在文本中的重要程度表示，即权重在这里是用来刻画词项在描述文本内容时的重要程度。在表示词项的重要程度时常常考虑以下两个方面的因素：

文本特征表示

- 词频 TF（Term Frequency）：词项在文本中出现的次数。
- 倒排文档频 IDF（Inversed Document Frequency）：特征词在文本集中分布情况的量化，度量特征词在文本集中出现的频繁程度。常用计算方法为

$$Log_2(\frac{N}{\{d \in D : t \in T\}})$$

- 式中 N 表示所有文档的个数,$\{d \in D : t \in T\}$ 表示包含第 t 个词的文档个数。

文本特征表示

- 相对词频可以用以下公式表示：

$$TFIDF(t,d) = \frac{TF(t,d)}{\sum_{f \in d} TF(t,d)} \times Log_2(\frac{N}{\{d \in D : t \in T\}})$$

文本特征选择

- 文本信息转化为的布尔矩阵或词频矩阵的维度非常大且矩阵非常稀疏（矩阵内很多元素为 0），严重影响了挖掘算法的性能。
- 一个有效的文本表示必须满足文本内容表达完整、较强的文本区分能力与保证尽量小的特征维度等特点。
- 需对词项进行选择即对文本的特征进行选择以达到文本向量维数压缩的目的。
- 特征选择不仅能够提高文本挖掘算法的运行速度，降低占用内存空间，而且能够去掉不相关或相关程度低的特征，提高文本挖掘的性能。

文本特征选择

- 特征选择的基本步骤为:

1. 初始情况下，特征集包括所有的原始特征;
2. 计算每个特征的评估函数值;
3. 选出评估函数值较高的前 k 个特征作为特征子集。

- 常见的特征选择方法包括文档频率选择法、信息增益选择法、交叉熵选择方法、χ^2 统计方法以及其他自定义评估函数。

文档频率选择法

- 在文本内容中，特征项的出现次数是决定特征重要性的判定依据。
- 文档频率选择法认为，特征的重要性主要由特征出现的频率次数决定，低于某个阈值的特征项一般不含或者仅有较少信息，因此选择大于制定阈值的特征项集合。

信息增益选择法

- 该方法多用于选择文本分类特征集。
- 信息增益 (Information Gain) 是信息论中的一个非常重要的概念，它实际上某一特征项在文本中出现前后的信息熵之差，用来衡量某一特征项的存在与否对类别预测的影响。计算公式为：

$$IG(t) = P(t)\sum_{c\in C} P(c|t)log_2P(c|t)+P(\bar{t})\sum_{c\in C} P(c|\bar{t})log_2P(c|\bar{t})-\sum_{c\in C} P(c)log_2P(c$$

- 其中 C 是文档的类别集合，t 为特征（词项），$\bar{t}$ 表示特征不出现，$P()$ 表示特征函数。

交叉熵选择方法

- 交叉熵与信息增益相似，二者的主要不同在于信息增益需要计算所有可能特征分值的平均，而交叉熵仅考虑一篇文档中出现的词。
- 计算公式为：

$$CE(t) = \sum_{c \in C} p(c|t) log_2 \frac{P(c|t)}{P(c)}$$

信息挖掘与主题模型

- 文本信息通过特征表示由非结构化数据转化为结构化数据，于是我们就可以采用传统的数据挖掘方法对由特征表示得到的布尔矩阵或词频矩阵进行分析得到有价值信息。
- 常见的挖掘方法包括：文本分类、文本聚类、主题模型等。

文本分类

1. 获取训练文本集：训练文本集（实际上就是训练集）是由一组经过处理的文本特征向量组成，每个训练文本有一个类别标签，实际上就是类别因变量 Y 的取值；
2. 选择分类方法并训练分类模型：分类质量较好的方法包括 k-最近邻法、支持向量机、朴素贝叶斯等；
3. 用训练所得的分类模型对待分类文本进行分类预测；
4. 根据分类结果评估分类模型。

文本聚类

1. 获取结构化的文本集，结构化的文本集由一组经过处理的文本特征向量组成；
2. 执行聚类算法，获得聚类结果，可以用系统聚类、K-means 聚类等；

LDA 主题模型

- 设 $w = (w_1, w_2, \cdots, w_N)$ 为一个包含 N 个词的文档，w_k 为文档中的一个词语。
- 该文档所有词语均来一包含 V 个单词的词语集。
- 同时一个文档包含多个主题，每个主题可以采用不同的词语表达。但这些词语均来自相同的词语集。
- 例如，我们写一个关于苹果的文档，其中包含了苹果的口味和苹果的产地两个主题，但我们所用的词语均为新华词典中的单词构成词语集。

LDA 主题模型基本假设

1. 文档的主题数目是确定的；
2. 文档中出现的主题服从多项式分布 $Multi(\theta)$，其分布律受参数 θ 影响；
3. θ 的取值是随机的，满足狄利克雷分布，即 $Dirichlet(\alpha)$，其中 α 为狄利克雷分布参数；
4. 词语出现服从参数为 β 的多项式分布，其中 β 为随机变量，服从狄利克雷分布 $dirichlet(\gamma)$,γ 为狄利克雷分布参数；

LDA 主题模型

- 词语的生成过程如下：

1. 对于每个 w_k，选择主题 z_k，服从多项式分布 $Multi(\theta)$
2. 在给定主题 z_k 时，词语 w_k 条件多项式分布概率 $p(w_k|z_k, \beta)$

- LDA 主题模型是采用极大似然估计的思想来估计参数 α, β，之所以没有估计 γ 是由于现实中我们往往更关心的 β。

自然语言处理

- 人工智能就是希望计算机和人一样能够说话、理解文章内容、与人交互。自然语言理解是人工智能中很有挑战的领域，因为使用语言的能力是人所独有的，是最高智能的体现。
- 自然语言理解，有两种定义：一种是计算机能够将所说的语言映射到计算机内部表示；另一种是基于行为的，你说了一句话，计算机做出了相应行为，就认为计算机理解了自然语言。后者的定义，更广为采用。

NLP 挑战

- 自然语言有 5 个重要特点，使得计算机实现自然语言处理很困难：

1. 语言是不完全有规律的，规律是错综复杂的。有一定的规律，也有很多例外。因为语言是经过上万年的时间发明的，这一过程类似于建立维基百科。因此，一定会出现功能冗余、逻辑不一致等现象。但是语言依旧有一定的规律，若不遵循一定的规范，交流会比较困难；
2. 语言是可以组合的。语言的重要特点是能够将词语组合起来形成句子，能够组成复杂的语言表达；
3. 语言是一个开放的集合。我们可以任意地发明创造一些新的表达。比如，微信中“潜水”的表达就是一种比喻。一旦形成之后，大家都会使用，形成固定说法。语言本质的发明创造就是通过比喻扩展出来的；
4. 语言需要联系到实践知识；
5. 语言的使用要基于环境。在人与人之间的互动中被使用。如果在外语的语言环境里去学习外语，人们就会学习得非常快，理解得非常深。

NLP 分类

- NLP 主要采用统计机器学习的方法来解决。
- 分类：一个字符串 + 一个标签，这个字符串可以是一个文本，一句话或者其他的自然语言单元；
- 匹配：两个字符串，两句话或者两段文章去做一个匹配，判断这两个字符串的相关度是多少；
- 翻译：更广义的翻译或者转换，把一个字符串转换成另外一个字符串；
- 结构预测：找到字符串里面的一定结构；
- 马可夫决策过程，在处理一些事情的时候有很多状态，基于现在的状态，来决定采取什么样的行动，然后去判断下一个状态。

R 语言实现

```
install.packages(c('tm', 'XML', 'SnowballC', 'wordcloud', 'topicmodels'))
install.packages("corpus.JSS.papers", repos = "http://datacube.wu.ac.at/", type = "source")
```

本周推荐

- 参考书：《文本数据挖掘》，宗成庆等，2019，清华大学出版社
- 推荐视频：北大人工智能公开课第 8 课：李航自然语言处理的现实与挑战，爱奇艺，2017-04-23，
  https://www.iqiyi.com/w_19ru3d804d.html

谢　谢!