使用朴素贝叶斯进行文本挖掘

黄勇 201521210022

1. 研究目的

朴素贝叶斯和决策树算法一样，都是一类基本的算法总称，基于朴素贝叶斯定理，可以实现多种分类和聚类方法。贝叶斯分类是通过简单的贝叶斯假设，假定特征条件相互独立的分类方法，从而实现有监督的学习方法，随着社交网络、移动互联网技术的兴起，当数据量不断的增大时，有监督的学习成本也不断的增大，而且海量数据进行学习训练时，需要考虑很多时间成本因素。此时，诸如支持向量机、随机森林、贝叶斯网络等算法复杂度都超过了 $O(n^2)$，不适应当前对于海量文本的计算，而同时，朴素贝叶斯由于对于条件独立的假设，对于当前很多现实 d 额数据集来说，都具有非常合理的假设，并且在这种假设下也能获得非常好的结果，从而能在很多实际商业应用获得青睐。

本文从朴素贝叶斯出发，研究朴素贝叶斯在文本分类中的应用，由于互联网和搜索引擎的发展，朴素贝叶斯在文本的分类、文档聚类、甚至情感分类、舆情监测等方向都有巨大的发展，在垃圾邮件检测等方面已经获得了巨大的成就。因此，本文采用朴素贝叶斯方法对于希拉里邮件进行分类、聚类研究。主要目的是为了能熟练使用朴素贝叶斯在文本挖掘中的各种应用，并通过文本挖掘发现和希拉里相关的邮件情感变化、以及邮件涉及的相关主题、在这些主题下的关键词语。除此之外，我们还希望能发现希拉里邮件中的高频词，通过这些高频词来刻画希拉里的关心内容的重点。总之，借用 7000 篇希拉里邮件是能够实现全面有效的分析希拉里邮件中的各个方面，挖掘潜藏语义情景。

2. 实验材料：被试信息、扫描参数、实验设计

本次实验选用了希拉里邮件[1]这一份公共数据集。这份数据来源于顶级数据挖掘竞赛网站 Kaggle 上，由于希拉里被怀疑滥用政府的公共邮箱，使用公共邮箱处理自己的私人事务，从而受到联邦调查局的调查，并因此公开了超过 7000 分邮件数据，这些邮件数据被存储在 SQLite 数据库中，总共包括 4 个基本表：Emails.csv、Persons.csv、Aliases.csv、EmailReceivers.csv。这四个表中涉及到希拉里发送和回复的邮件内容、邮件的头信息、并包括很多涉及到的人（部分人被匿名）这些主要信息，我们期望能从这些信息中进行挖掘，找出有效信息。下表列出了相关邮件信息的数据量。

条目	总数
Emails	7945
Persons	513
Alias	850

Email Receivers	9306

3. 数据处理方法(原理，处理流程，软件等)

数据分析方法主要采用 R 语言和部分 SQL 查询语言，通过将 SQL 嵌入到 R 语言中进行统计分析，从而获得分析结果。R 是用于统计分析、绘图的语言和操作环境。R 是属于 GNU 系统的一个自由、免费、源代码开放的软件，它是一个用于统计计算和统计制图的优秀工具。可以认为 R 是 S 语言的一种实现。而 S 语言是由 AT&T 贝尔实验室开发的一种用来进行数据探索、统计分析和作图的解释型语言。最初 S 语言的实现版本主要是 S-PLUS。S-PLUS 是一个商业软件，它基于 S 语言，并由 MathSoft 公司的统计科学部进一步完善。后来 Auckland 大学的 Robert Gentleman 和 Ross Ihaka 及其他志愿人员开发了一个 R 系统。而 R 是一套由数据操作、计算和图形展示功能整合而成的套件。包括：有效的数据存储和处理功能，一套完整的数组（特别是矩阵）计算操作符，拥有完整体系的数据分析工具，为数据分析和显示提供的强大图形功能，一套（源自 S 语言）完善、简单、有效的编程语言（包括条件、循环、自定义函数、输入输出功能）。

当前随着开源软件的兴起，R 语言在开源社区中获得巨大的欢迎。不仅能获得 oracle 和微软等公司的大力赞助，推进 R 语言的规范化和标准化，而且开源软件 Spark 和 Hadoop 等开源软件都对 R 语言进行大力支持，开发了官方的接口。

除了 R 语言的开源、方便计算、成熟的可视化体系。我们采用了朴素贝叶斯这一系列算法对于我们的文本数据进行描述和整理，从而获得有价值的文本信息，将文本的数据能可视化、有效地展示给大家。

朴素贝叶斯的基本原理可以由以下公式进行解释，除了最基本的朴素贝叶斯条件相互独立的假设之外，朴素贝叶斯随着不同的分类模型，对于条件的分布有着自己的假设，大体上可以分为基于高斯模型的朴素贝叶斯、基于伯努利二项式模型的朴素贝叶斯、基于多项式模型的朴素贝叶斯假设。

P(Category | Document) = P (Document | Category) * P(Category) / P(Document)

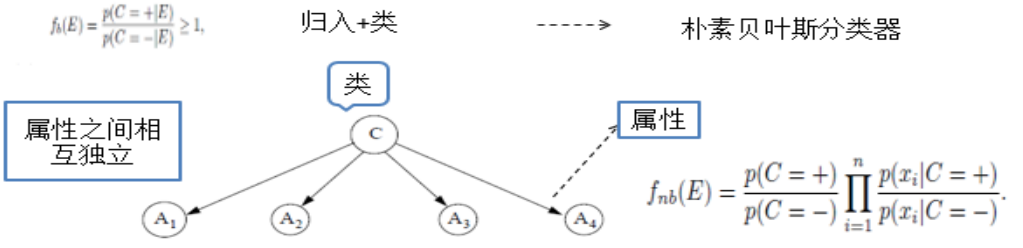


Figure 1: An example of naive Bayes

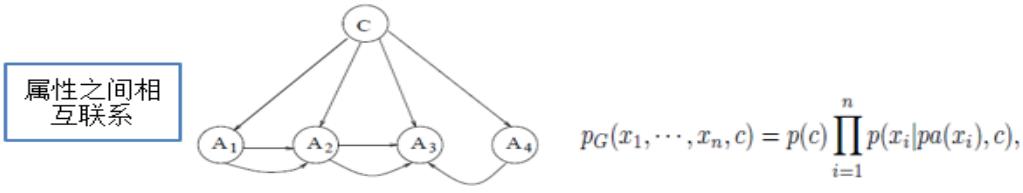


Figure 2: An example of ANB

在我们之前的论文研读中发现[2]，朴素贝叶斯分类模型建立分两个步骤：第一步,建立一个模型，描述预先的数据集或概念集。通过分析由属性描述的样本（或实例，对象等）来构

造模型。假定每一个样本都有一个预先定义的类，由一个被称为类标签的属性确定。为建立模型而被分析的数据元组形成训练数据集，该步也称作有监督的学习。

除了研究朴素贝叶斯的分类算法外，我们也关注了朴素贝叶斯的分类模型假设，同样由于很多文本分析的模型试验表明，基于伯努利二项式模型相比于多项式模型，在小文本下比较有优势，而文本量较大时可能会比较差，从而我们选择了基于多项式模型的朴素贝叶斯模型[3]。对于朴素贝叶斯模型条件独立的假设，我们特意查找论文，研读发现，在大多数情况下，朴素贝叶斯能快速处理大规模数据集合，对于模型之间有相互关系的变量时，由于分类过程中的属性相互抵消，模型通常可以看作条件相互独立，这样对于大部分分类、尤其对于文本分类来说有着非常重要的作用[4]。

由于属性之间的相互依赖、但是到了计算模型和模型计算两个属性之间属于不同概率大小的时候，概率和概率之间的比值通常在 1 附近，这样的比值结果在对于文本分类这一概率分析时，不会影响模型的分类准确率。从而对于我们分析希拉里邮件的情感态度时，有着非常重要的意义。下图展示了使用朴素贝叶斯和考虑属性之间相互联系的贝叶斯网络时，模型计算所得的概率之间的差别。作者用严谨的数学知识证明了朴素贝叶斯在一般文本分析中的合理性和重要性。

$$dd_G^+(x|pa(x)) = \frac{p(x|pa(x),+)}{p(x|+)}$$
$$dd_G^-(x|pa(x)) = \frac{p(x|pa(x),-)}{p(x|-)}.$$
$$ddr_G(x) = \frac{dd_G^+(x|pa(x))}{dd_G^-(x|pa(x))}.$$

描述属性x的局部依赖性将E分到+类或-类的**倾向程度**。

$$f_b(x_1, x_2, ..., x_n) = f_{nb}(x_1, x_2, ..., x_n) \prod_{i=1}^{n} ddr_G(x_i),$$

找到属性之间有联系的分类器和朴素贝叶斯之间的联系

除了利用朴素贝叶斯进行情感分类之外，我们还考虑到朴素贝叶斯最近很多年的发展，很多作者基于朴素贝叶斯提出了多层次的模型，其中最重要的一种是主题模型（topic model）。LDA 是一个 Bayes hierarchy model，假设文档由主题组成，主题由词构成。利用贝叶斯推断相关参数。文档到主题服从 Dirichlet 分布，主题到词服从多项式分布。利用 Bayes 推断、Gibbs sampling 等做分布参数的推断。

一篇文章有若干主题，围绕这几个主题遣词造句，表达成文。LDA 根据给定的一篇文档，推测其主题分布。

对于数据的处理主要分为三个步骤：(1) 将获得的文件存入数据库中，并按照 SQLite 数据库的规范进行数据预处理，并编写后续的数据库查询语言，除此之外，我们还要确定数据库中重要的查询字段，确保查询的准确性和有效性。(2) 对于基本的查询字段进行数据统计分析，除了获取查询相关的任务的 raw Text 信息，我们还要将查询的基本数据进行去除停用词分析，除了将停用词去除、我们还假设所有的英文单词属于单个单词，即认为 Beijing Normal University 是属于三个单词，这种分词方法属于 unigram。(3) 对于进行主题分析（topic model）的模型来说，为了考虑模型的有效性，我们将英语中的停用词最小化，减少很多必要的模型的数据预处理流程。

除了必要的数据处理之外，我们将使用了很多 R 语言的开源包，这些开源社区的 R 语言统计分析包可以帮助我们快速有效的进行情感分析[5]和主题模型分析[6]。情感分析的这些模型使用了朴素贝叶斯模型，同时朴素贝叶斯模型进行情感分析也能帮助我们的数据提供规范的数据训练模型、训练模型需要大量的统计分析、标注工作。利用这样的数据集能很好的帮

助我们很好预测情感类型。

4. 结果图表等

我们首先将数据进行了数据预处理，并通过 R 语言编程实现了模型的基本数据统计。首先，我们筛选出希拉里邮件的部分停用词和不恰当的词语。除了将希拉里的接受邮件和希拉里的发送邮件分别进行查询处理外，我们也统计了这些词语的词频，通过绘制词语的词云图，我们可以有效发现词语中的高频重要词语，这些词语可以有效帮我们分析希拉里日常的工作和日常情况。

接受邮件 1

发送邮件 1

通过对于高频词的统计分析、绘制可视化图表我们可以看出：希拉里接受的邮件大多和 Obama 等政治消息有关、而希拉里发送的邮件却非常的简短、很少涉及政治消息，这些邮件很模糊的提到 work、pls 等和工作相关的语言。

我们的第二部是对希拉里接受和发送的邮件进行情感分类。情感分类可以划分为：积极和消极的两种分类方式、也可以化为喜怒哀乐等多种分类的情感分类方式，这两种不同的分

类方式只是由于类标号的不同导致模型的结果不同，而模型的本质算法都是基于朴素贝叶斯。下图展示了模型在使用 R 语言的 syuzhet 包进行识别模型中的相关情感词语。

anger	anticipation	disgust	fear	joy	sadness	surprise	trust
1	0	0	2	1	1	1	3
15	16	3	21	13	10	10	41
1	0	1	3	0	3	1	1
1	2	1	2	0	2	0	2
15	15	4	20	13	10	10	40
2	1	1	2	0	2	1	1
9	11	8	12	5	7	4	19
15	16	3	21	13	10	10	41
22	21	5	27	21	16	11	44
19	22	8	26	7	9	10	30
0	5	2	3	4	3	2	15
19	22	8	26	7	9	10	30
0	1	0	1	0	1	0	2

我们将模型的情感分类首先分为两类，正向的情感分类词语和非正向词语。非正向的词语包括中立的情感词语和负向的、消极的词语。下图展示了我们的模型对 7900 多封邮件的分类结果。除了采用两类分类方法外，我们也采用了多种情感的分类方法。这种多情感分类方法也有不同的效果，同时，和以前的分类一样，我们也考虑了接收邮件和发送邮件的两种不同内容包括，包括发送邮件的情感和接受邮件的情感分类。

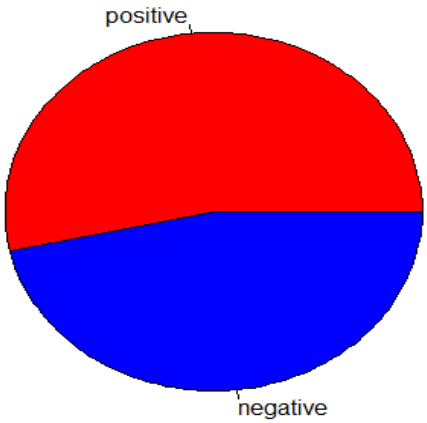


Figure 1

由上图可以看出，希拉里发送和接收到的邮件中，多数邮件是积极的情感倾向，少数是中立和负向的情感倾向。同时，从下面的表二和表三可以看出，希拉里的邮件中多数表达了信任（trust）和 anticipation 这两种主要的情感。

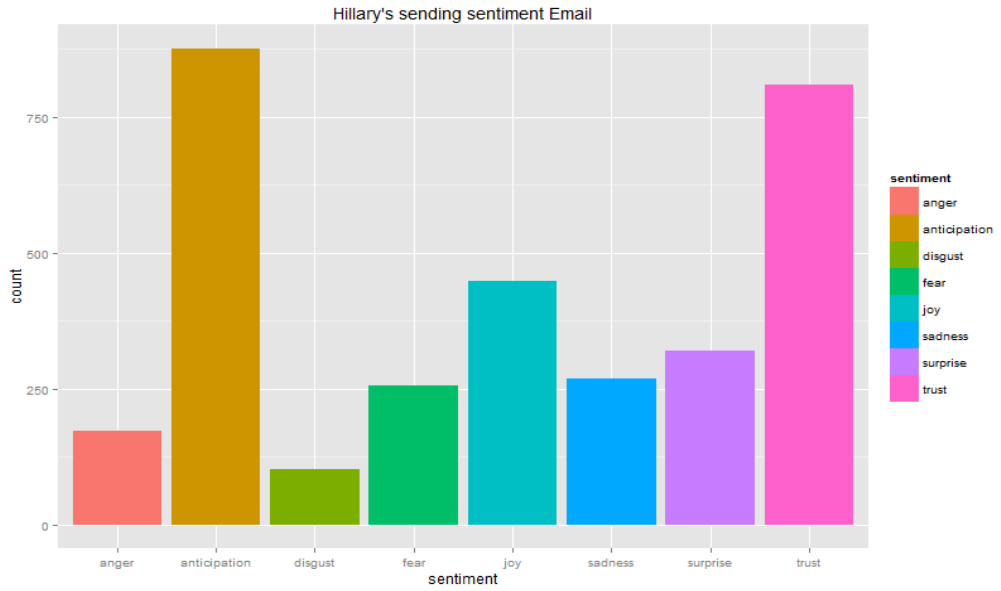
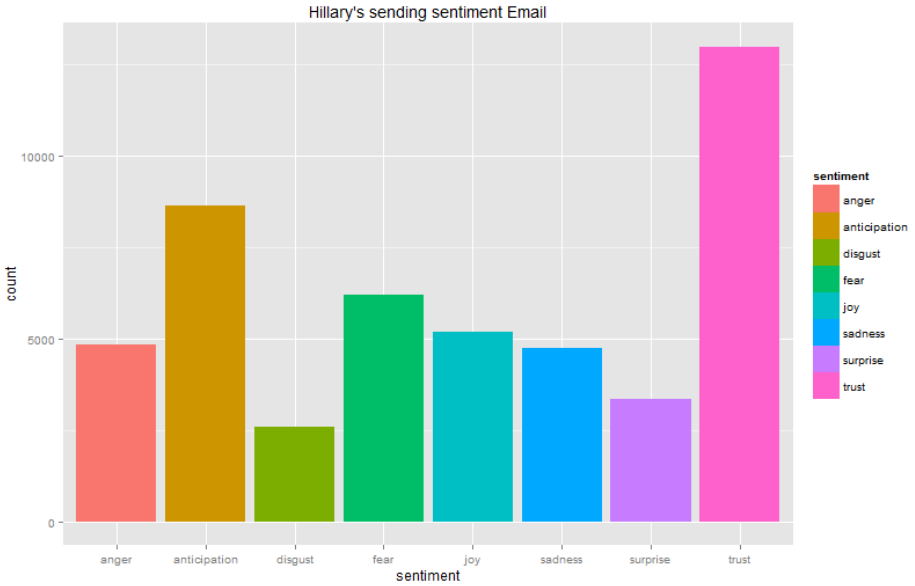


Figure 2

表三也可以看出，希拉里接受的邮件也是充满了信任感，而且这和希拉里的之前的邮件词频分析有相互照应之处，这些模型的相互照应可以帮助我们发现模型的有效性和相合性。



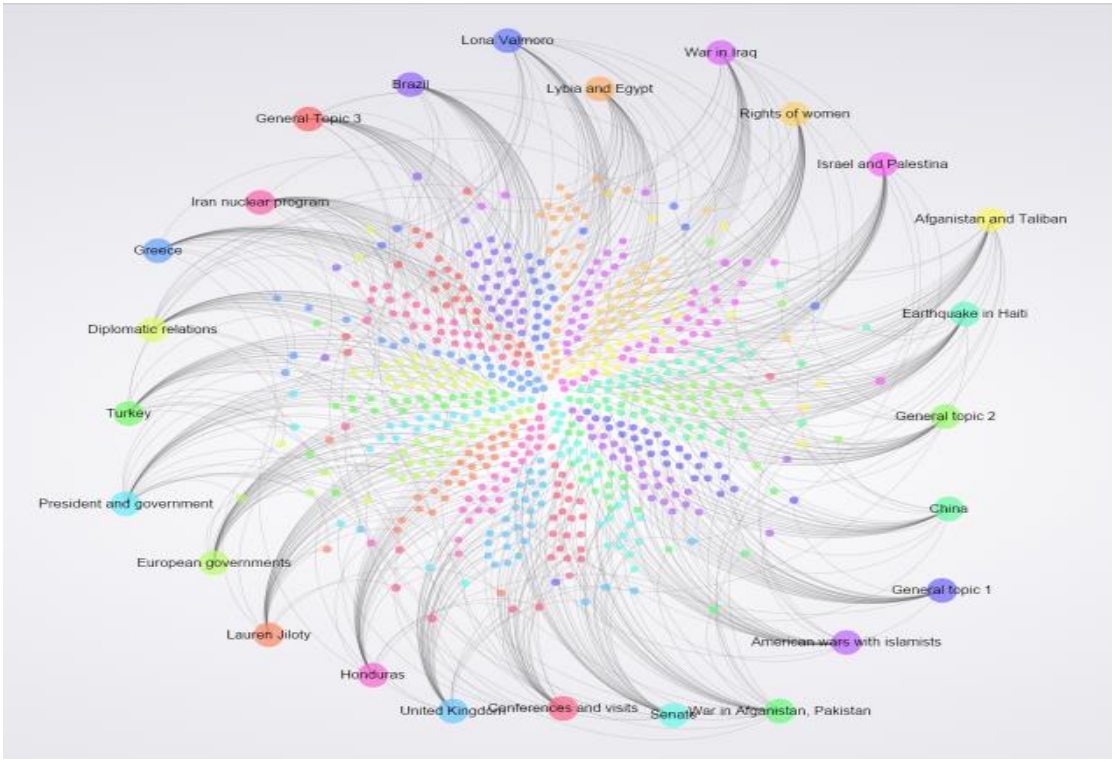
第三部我们对于模型进行了主题模型分享，主题模型分析是一种类似聚类的数据统计方法，这种聚类方法只需要设置模型的聚类总数，并且设置每次迭代的步长等较少的参数，从而可以获得模型的聚类结果。下图展示了模型的 25 个聚类主题和这些主题下的相关词语。左边属于希拉里接受邮件的 25 个主题中，最后 10 个主题，右边属于希拉里发送邮件的前 10 个主题。

Topic 16	Topic 17	Topic 18	Topic 19	Topic 20	Topic 21	Topic 22	Topic 23	Topic 24	Topic 25
the	you	the	the	the	the	state	the	the	office
and	will	and	and	and	and	no	and	and	secretary's
the	for	that	for	that	the	for	for	meeting	
for	and	that	for	that	for	benghazi	that	you	state
with	for	with	with	with	was	u.s.	with	2009	department
are	call	his	not	our	with	departmer	was	percent	route
have	with	mcchrystal	his	are	boehner	case	are	enough	arrive
has	have	has	this	this	not	informatio	our	2010	depart
was	can	was	are	will	his	house	they	unfavorab	room
they	are	芒犹	?,yc	their	who	subject	from	your	private
that	have	but	have	has	doc	have	about	residence	
from	this	this	have	from	from	agreemen	has	his	conference
not	your	not	from	united	this	date:	will	haven't	•en
but	just	from	their	but	have	and	you	dki/na	time
who	from	but	they	people	house	sensitive	but	obama	daily
more	but	they	mr	was	but	produced	this	approve	and
would	get	are	will	not	they	select	foreign	heard	house
see	who	would	they	would	comm	said	disapprov	airport	
their	talk	all	has	states	are	dept	not	opinion	staff
his	tomorrow	election	about	u.s.	her	foia	who	men	the

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
the	the	the	the	2009	the	the	you	office	the
and	and	and	and	2010	and	and	the	secretary's	and
that	our	that	that	cheryl	no.	that	will	meeting	that
for	for	for	for	re:	state	for	and	arrive	for
are	with	was	was	fw:	that	she	for	state	they
with	that	have	with	mills,	benghazi	with	with	route	are
not	this	with	has	huma	u.s.	his	have	departmer	not
his	are	but	have	abedin,	house	her	this	depart	with
will	will	his	they	sullivan,	subject	was	are	room	was
has	state	not	their	<millscd@	case	has	call	private	his
this	united	from	from	jacob	informatio	obama	that	residence	but
from	their	had	are	fyi	doc	but	but	•en	this
have	have	this	who	sunday,	sensitive	have	can	conference	about
but	芒犹	?,yc	not	friday,	date:	clinton	your	time	from	northern
would	states	about	american	saturday,	select	president	from	daily	you
who	all	has	its	thursday,	produced	not	out	house	had
was	u.s.	said	this	see	departmer	from	want	airport	have
said	departmer	would	koch	wednesda	comm.	had	just	staff	would

通过对于这些主题的分析，我们可以发现，希拉里很多邮件涉及到中国、伊拉克等主题，所以我们查找了相关的网站，查看他们对于这一主题模型的使用和建立相关的数据分析结果。

下图参考一些网上资料[7]并将希拉里邮件的主题进行可视化分析结果。从而我们可以有效发现希拉里邮件中的主题和这些主题中的关键词语分布情况。



5. 讨论和结论

通过对于希腊里邮件进行统计分析和结果计算，我们可以绘制很多文本分析的结果，这些结果大都基于朴素贝叶斯进行情感分类和聚类。除了利用朴素贝叶斯进行分类和聚类之外，我们也可以看到基于贝叶斯分类在当前大数据的文本分析有很强的实践性，除此之外更重要的是结合朴素贝叶斯和很多其他的数理统计知识，这样我们可以借助数据挖掘等手段在当前的数据分析中获得更多好的结果。

6. 研究意义

本研究结合实际中的文本数据、对于希拉里邮件进行了词频统计分析、多种模型的情感分类、LDA 主题模型聚类。这些基于朴素贝叶斯的方法能有效帮助我们在文本挖掘中获得好的分类聚类结果，同时我们也看到有好的数据集、好的数据预处理方法、好的统计编程语言工具包对于一次实际数据统计分析挖掘任务的重要性。这样的数据分析有很强的实际效果，对于我们在实际的任务、比如在搜索引擎中的应用、互联网的舆情分析等方向都有很强的指导性作用，我们可以借助实用的、高性能的数据统计模型和方法，在将来的学习任务重更好的应用这些模型，从而我们可以获得更好的实际应用体验，这也算是最好的学以致用。

7. 参考文献

[1] https://www.kaggle.com/kaggle/hillary-clinton-emails

[2] McCallum, Andrew; Nigam, Kamal(1998). A comparison of event models for Naive Bayes text classification. AAAI

[3] Rennie, J.; Shih, L.; Teevan, J.; Karger, D(2003). Tackling the poor assumptions of Naive Bayes classifiers. ICML

[4] Zhang, Harry(2004). The Optimality of Naive Bayes . FLAIRS2004 conference. AAAI

[5] https://github.com/mjockers/syuzhet

[6] https://cran.r-project.org/web/packages/topicmodels/vignettes/topicmodels.pdf

[7] http://mellain.github.io./topic-term.html