1.题目：

社交媒体与情感表达

咎志浩 41414222 会计学（注册会计师方向）

2.研究问题：

[“美国人民在谈论特朗普这一\“话题总统\”时，所表现的情感极性”：与原本猜想不同，在各个社交媒体中，对特朗普表现的情感极性总体为积极，”这种情感极性是否因社交媒体而异？”：不同社交媒体中，人们表现的情感极性大致相同。但极性的分布有所不同]

3.研究动机和背景：

在发现澳大利亚的黑天鹅之前，欧洲人认为天鹅都是白色的。毕竟他们所见到的天鹅都是白色的。 但随着第一只黑天鹅在澳洲的发现，这个不可动摇的信念崩溃了[1]。一只黑天鹅的出现直接颠覆了一直被坚信的常识，也由此衍生出一个概念：黑天鹅事件（Black swan event）。黑天鹅事件是指非常难以预测，且影响重大的事件，通常会引致某个或者多个领域发生颠覆的事件。[2]

自 2016 年以来，世界政治格局可以说一直被黑天鹅事件的阴影笼罩着。从英国“脱欧”，到特朗普当选为美国第 45 任总统，以及前不久结束的英国提前大选。这些政治“黑天鹅”事件的结果都让人出乎意料，甚至影响了世界宏观经济的波动。同时，在这些意外事件发生后，人们也在反思、归纳，试图找出这些事件背后的规律。

而用以解释这些事件的一个概念便是：沉默的大多数(Silent Majority)[3]。沉默的大多数原指的是不愿公开发表意见的人。而就美国大选而言，沉默的大多数更多地表现为“被沉默”。

大众媒体是信息传递的高速公路。它能有效、快速地传递来自世界各地的信息。但从大众媒体的本质上来看，大众媒体的背后都是人。只要是人，就会有预设的立场。因此，所有大众媒体都无法避开报道偏向性。下图便是一个典型的案例。

How the Media can manipulate our viewpoint

在美国大选过程中，由于媒体本身的立场以及特朗普本人不拘一格的说话风格，使得媒

---

[1] Puhvel, Jaan (Summer 1984). "The Origin of Etruscan tusna ("Swan")". The American Journal of Philology. Johns Hopkins University Press. 105 (2): 209–212.

[2] Taleb, Nassim Nicholas (22 April 2007). "The Black Swan: Chapter 1: The Impact of the Highly Improbable". The New York Times. Retrieved 20 January 2016.

[3] Silent majority" Cambridge Advanced Learner's Dictionary (1995), accessed 22/2/2011

体报道往往更偏向于夸大特朗普言语的荒谬性。从而不断地向媒体受众灌输一种认知：“这样荒唐的人必然无法胜任总统”。但媒体也因此忽略了他所代表的民众利益以及站在他背后的选民的声音。也由此，大众媒体在一定程度上“人为”地产生了沉默的大多数。

由此，本文希望绕过大众媒体这一媒体中介，通过对 twitter 以及 reddit 上关于特朗普的评论进行情感分析，直接地了解美国人民自己发出的声音。同时，由于 twitter 和 reddit 虽然是用户自己发布信息，但本质上也是一种信息媒体。本文也希望能够由此探究这样情感是否也会因平台而异。

4. 数据：

本文主要使用了来自 twitter 与 reddit[4]的数据。

Twitter 是一个面向大众的 Microblog 平台。本文通过 twitter 自身提供的 api，获取的数据可以被认为是在固定时间随机获取的。取得了 3000 条左右涉及到关键词 trump 的评论。

Reddit 是一个用户自发上传新闻的网站，由于其本身网站设定：每一个子板块都会有一个特定的话题，任何一个帖子如果被 downvote 过多后，便会被折叠。因此每个子版块都会有一个固定的情感氛围。所以，本文分别抓取了/Donald Trump 和 Impeach_Trump 两个子板块下等量话题下的评论。总计 2000 余条。

5. 分析方法：

情感分析方法：

一开始自己也纠结于是自己做词典与分析方法还是使用第三方库的接口直接做情感分析。但后面尝试了 TextBlob 的情感分析接口之后，发现它的词典甚至能分析出:(这样一个单纯表情文本。充分认识到自己甚至不能做到它 1/10 的效果。果断地决定采用 TextBlob 的情感分析接口。 TextBlob 是一个基于 Pattern Library 和 NLTK 的第三方包。主要提供了基于字典的情感分析和基于电影评论语料库的朴素贝叶斯模型情感分析。因为对于后者来说，情感分析的准确度完全取决于语料库的好坏，而自己在看了几天的文章之后，仍然无法弄清楚如何自己训练语料库，并以此做情感分析。所以最后迫不得已采用了 TextBlob 基于 Pattern Library 的词典分析方法。自己在人为比对之后发现其实效果也不错

数据处理：

1.抓取了 Twitter 与 Reddit 的评论文本。但存在以下问题：1）时间序列过于集中于某一个时间点。2）Twitter 数据未能抓取到点赞数。因为数据爬取本身花了太长时间（2 周），自己又花了 3 天尝试了许多改进也没有作用。同时时间序列分析本身与题目相关性不大，所以最后放弃了做时间序列分析。

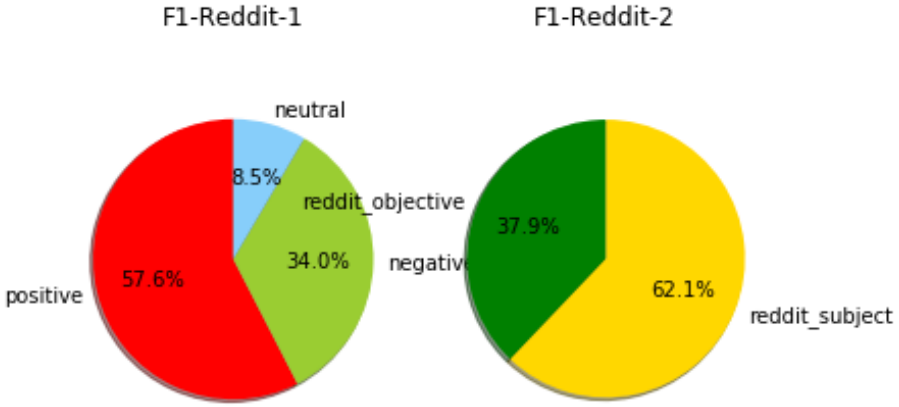
2.存在一些文本无法被 TextBlob 分析，典型形式为： Sentiment = 0 & Subjectivity = 0 。因此将这些数据筛掉了。

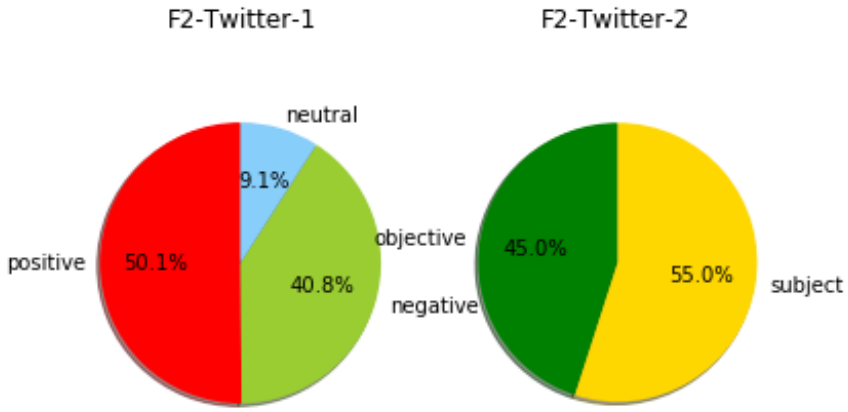
3.将 Reddit 按照 Score 分数加权，即：若评论 A 的 Score 分数为 B（B>0），Sentiment 分数为 C，则将其看作有 B+1 个人的 Sentiment 分数为 C 。若 B < 0 ,则取其 Sentiment 分数的负值。

6. 结果

[4] 抓取 reddit 的代码借鉴并使用了来自 github 的开源项目 https://github.com/asabine/reddit_crawler

0)在话题 “Trump”下所表现的情感极性”：



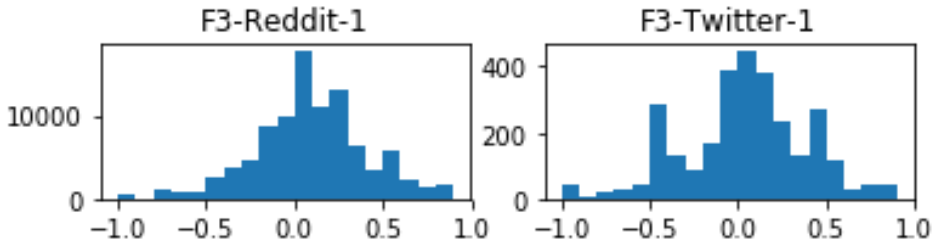


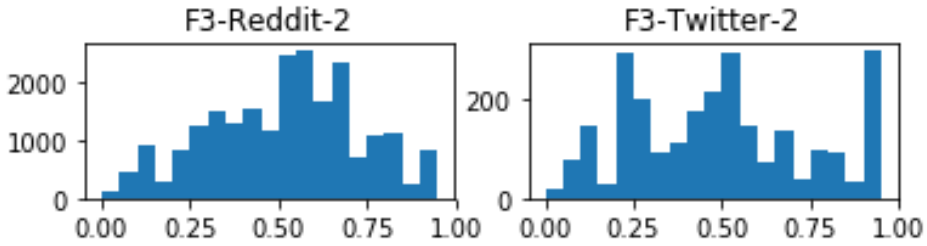
分析结果如图所示上图所示：即无论是在 Reddit 或是 Twitter 上，关于特朗普的评论都是积极大于消极。但两者之间差距不大，持中立立场的评论数占 9% 左右。也就是说 50% 左右的人群是倾向于支持特朗普的。这也与大选时的选票情况基本相符。

同时这些评论的主观性明显大于客观性。这与本文预期基本相符。因为社交媒体本身具有社交性，用户发出的评论客观性大于主观性几乎是必然的。这也能从某种程度上验证文本分析工具的稳健性。

1)社交媒体之间差异

F1 与 F2 我们可以明显看出，相比于 Twitter， Reddit 中对特朗普这一话题的评论积极的评论的比例要多 7.5%。同时，中性评论的比例差别不大。差异主要体现在积极/消极的情感差异。





F3 是根据情感分数频率得出的频率分布图。通过两两比较，我们可以直观地看出：就情感极性而言，Reddit 的情感分布更偏向于正向。Twitter 的情感分布则更均匀对称，偏向于两极。主/客观程度分布也类似于此：Reddit 的评论更偏向于主观，而 Twitter 则是两极对称。

最后，本文对情感分数与主客观程度分别进行了一次均值计算，得出：Reddit 平均情感分数为 0.094，平均主观程度为 0.558。Twitter 平均情感分数为 0.0352。平均主观程度为 0.557。说明情感分数总体来说仍是中性偏向于积极，分数因社交媒体的差异而略有不同，但差异不大。

2)总结

以上数据与图表首先在一定意义上确认了沉默的大多数的存在。他们可能无法在大众媒体上发声，但能在社交媒体上发表自己的观点。其次，也说明了特朗普并非像大众媒体所描诉的那样不受支持。从整体上来说，偏向于支持他的人要多一些，但优势不大。可以用“支持的人不多，反对的人不少。”来概括。最后，说明了各个社交媒体中用户评论的情感分数有一定差异，但差异不在情感方向，而在于情感分布。

7. 代码运行说明

0).运行代码需要安装的库

- 数据抓取
  - tweepy: $ pip install tweppy
  - scrapy: $ pip install Scrapy
- 数据清理以及情感分析
  - re: $ pip install re

- textblob
  - $ **pip install -U textblob**
  - $ **python -m textblob.download_corpora**

1). 抓取数据代码

- 概述代码分为两部分：抓取 twitter 和 reddit。都被存放在 scrapy_data 文件夹中。抓取 twitter 的代码以 twitter.py 的形式存放，抓取 reddit 的代码以 scrapy project 的形式存放。
- twitter.py 是通过 twitter 的 API 获取流，得到数据。自己尝试了很多方法，希望能根据数据大小自动停止运行，但始终都有问题（比如sys.exit()）。**因此，数据抓取需要手动停止**。由于 authentic token 需要在 twitter 上注册 app，本身的获取过程也挺麻烦的。所以自己没有将 authentic token 删除，老师**直接运行**就好。
- reddit 数据 需要通过 Scrapy 爬取，代码参考与借鉴了 github 中的项 目。**运行需要进入project文件夹**输入 scrapy crawl reddit -a subreddit=The_Donald -a pages=30

- 因为自己在获取数据是采取了分时、分来源采集。对结果展现还是挺重要的，所以我将这两个数据集也上传了，被命名为 twitter.json, 和 reddit.json 并**且数据清理，以及情感分析读取的也是这两个文件。**

2).数据清理与情感分析代码

- 分别定义了 data_cleanr() data_cleant(), 以及 sentiment_analyze() 三个函数。作用分别是清理 reddit ， twitter 的数据。以及进行情感分析
- 为了避免出现一运行就出一堆图的情况，我把生成图的代码都注释掉了，老师如果想看，可以删掉注释符号。

8. 总结

学到了什么

这次课程作业给我带来的收获主要有 3 点。第一，使我对程序设计有了更深层次的了解：这是一门重实践、重解决问题的学科。自己在开始做项目之前觉得毫无头绪，便在网上看了一些课程。比如中国 MOOC 上嵩天老师的一系列 Python 课程，也在 Github 上找了几本关于情感分析的电子书，但实际上，这些课程在做项目的时候并没有发挥太大的作用。最后，自己硬着头皮开始做，不断地解决一个个问题，最终完成了这个 项目 。现在回顾之前的过程，发现自己更多的是在做中学 （learn by doing）。也就对程序设计有了这样一个认知。其次，自己通过这个项目对通过机器学习做情感分析有了一定的了解。虽然在最后并没有用上这个方法，但挺激发自己对这方面兴趣的。最后，提升了自己解决问题、发现问题的能力。在编写程序的过程中，总是会突如其然地出现各种问题以及程序报错。有的是知识盲点、有的是代码错误、也有的是数据处理过程中有缺陷。而自己有一个缺点就是不太愿意去麻烦别人帮自己解决问题，所以万事靠自己。最后通过思考分析＋ Google 把一个个问题解决下来。期间也有一段时间陷入绝望，调整心态之后还是继续做了下来。现在想起来还是挺有成就感的。

希望在一开始知道的事情

希望老师能讲一讲爬虫的基本原理，以及介绍一下一些基本的爬虫包。自己在爬取数据过程中花费了相当多的时间，而且效果也不尽人意。自己感觉如果能够在抓取数据之前，对实现过程能够想出一个流程图，一定能更有效地利用时间。

给类似项目同学的建议

其实我觉得也称不上是建议吧，因为自己本身基础挺差的，而且这个项目本身做的自己也觉得是尽力了但仍然不算好。有一点我觉得可能有帮助的点就是：对没有太多编程基础的同学来说，如果感觉自己在如何写程序上毫无头绪的话，不要想太多，硬着头皮直接去做。不要查一些太过于宽泛的问题，比如：如何用 Python 抓取数据、如何用 Python 进行数据分析。而是按照自己具体化的需求去查资料，比如： 如何使用 Scrapy 爬取 Reddit 。看看别人是如何具体做的。我自己一开始也觉得很迷茫，感觉自己完全不会做。花了很多时间在看关于如何爬取数据的资料上。但看完之后打开 Spyder 还是一脸茫然。最后到 Github 和 Stack Overflow 上去看别人是如何实现这些需求的，才逐步上手。

感想

经过一学期的课程学习后，自己认为当初能选上这门课程还是挺幸运的。虽然过程很艰辛，在做项目期间自己几乎每个晚上都在寝室里查资料、写代码。室友也惊讶于为什么我每天晚上都在做一个公选课作业。但最后做下来，自己还是挺开心的，即便做的结果自己也不是很满意。最后也很感谢老师一学期的教导，老师上课认真也很有趣。谢谢您~ :)