《基于推特数据以 24 小时为周期的情绪变化研究》

基于我自己的情绪变化，我到了晚上就会出现情绪低沉，情绪消极的情况，所以试图探究出人的情绪是否与时间节点存在周期性关系，以及一些相关性。调查出某种规律性的联系。

本研究的研究对象为 2012 年 2 月 15 日到 2 月 29 日的英国伦敦地区的全天的推特数据，原数据形式为 CSV 文件。

1. 数据基本量的研究：

1.1 数据总量：

经过统计数据总量为 1017495 组，数据数量级为百万级。可以得到一些相对有规律性的总结。

1.2 每日的数据总量：

每天的数据总量分别为[66516, 70431, 67916, 69683, 72053, 73625, 76229, 71925, 77067, 79918, 73691, 69905, 68683]，并作出直方图，如表 1 所示。可以发现每天的推特数据量几乎一致，猜测推特数据量具有着周期，为之后的情绪变化研究提供了规律总结的可能性，周期性和重复性将会在之后的研究过程中体现地更加明显。

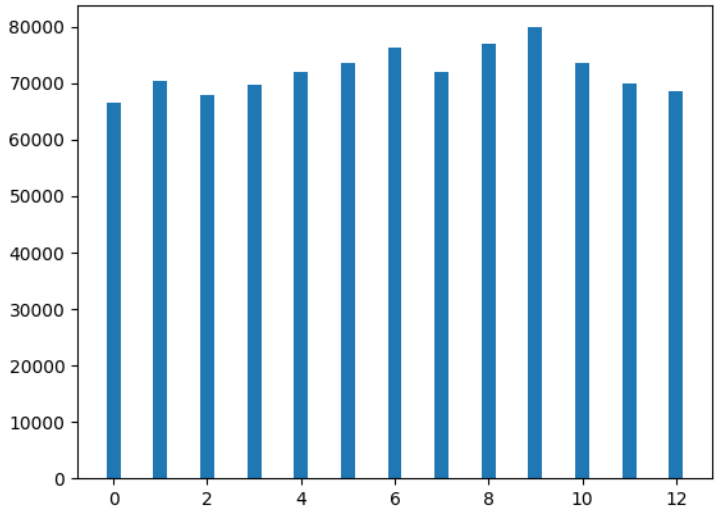
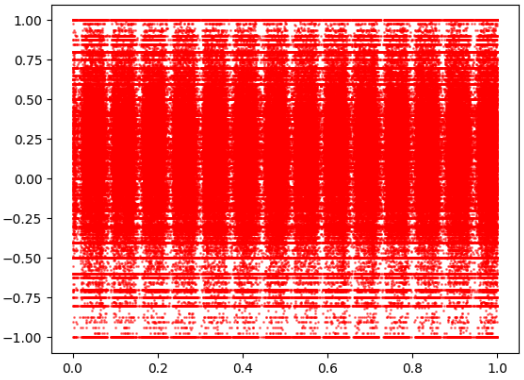


表 1 每日的推特数据总量的直方图

2. 情绪变化研究过程：

本次研究的研究对象 CSV 文件中的推特数据主要有思维分别是时间，推特内容，经度以及纬度。本次研究主要侧重于时间和推特内容的关系。

推特内容的呈现方式是以字符串形式存在，是一种自然语言的形式存在，如果对自然语言进行情绪分析，需要使用到 NLP(Natural Language Processing)，NLP 是人工智能领域的一个分支，能对自然语言进行分析，我使用的 python 里的 Textblob 库，Textblob 库能对一段自然语言进行情绪分析，并返回主观性以及极性（sentiment.polarity），sentiment.polarity 的返回值越接近+1 情绪越积极，返回值越接近-1 情绪越消极。



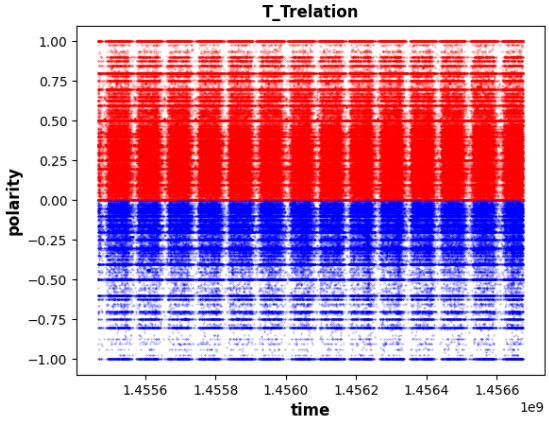


表 2 情绪极性与时间的关系

2.1 情绪周期性：

现对散点图对时间和情绪进行简单研究，先遍历 CSV 文件，在遍历的同时对时间和推特内容进行相应的数据处理。注意要在遍历的同时，对数据结果进行操作，否则处理时长会翻倍。同理在做时间数字归一化的时候也要在遍历的同时，做一些处理，可以有效地减少处理时间。在上述图可以发现，推特数据上发布内容的情绪极值和发布时间具有着十分明显的周期性。与时间有着很明显的关联性，但由于数据量过于庞大，散点图不能很明显地对比出情绪的消极性和积极性与时间的关系。现对数据量进行简单统计。

2.2 情绪数据量的对比：

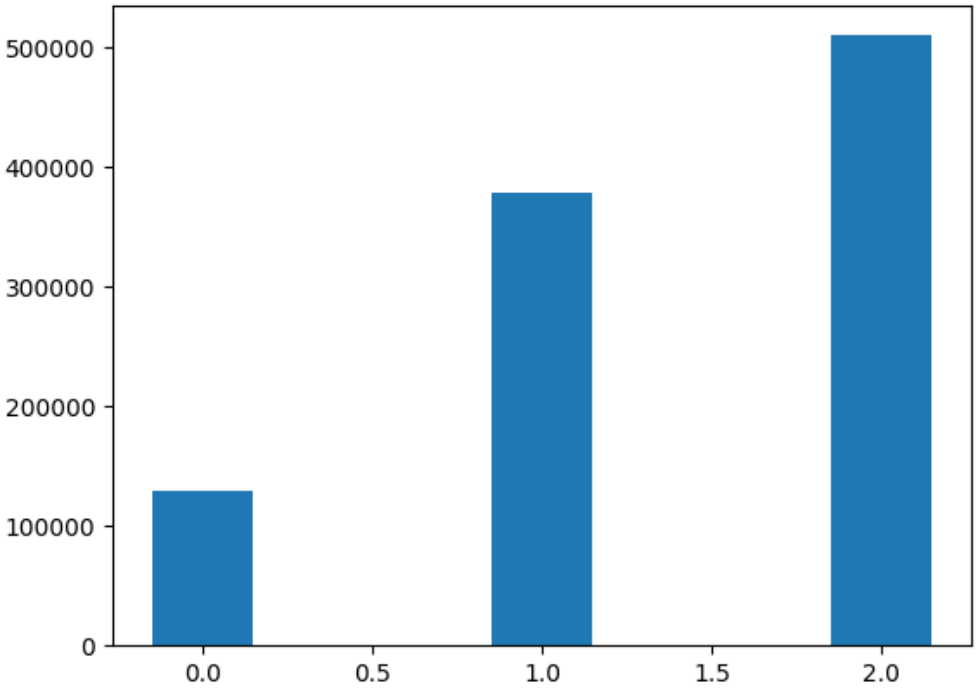
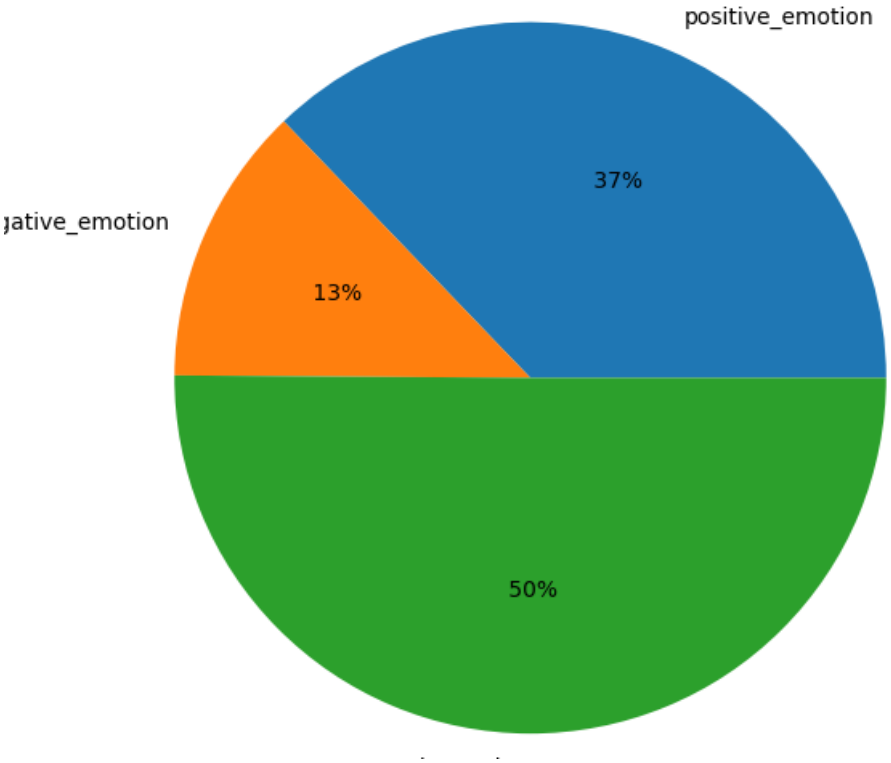


表 3 消极，积极，中性数据量的柱状图



简单地用 count 计数并且遍历，可以计算出消极情绪有 129377 个，积极情绪有 378157 个，中性情绪有 509961 个。用柱状统计图和扇形统计图可视化之后，如图。在图中可以发现在推特中，消极情绪数量小于积极情绪数量，积极情绪数量小于中性情绪数量。

表 4 积极，**消极**，**中性情**绪的扇形统计图

2.3 情绪分析细化

之前对情绪的种类和数目进行了统计和可视化，是一个总体分析，现对利用数据，对每天乃至每个小时进行分析进行分析和总结，以求得某种固定且对现实生活具有指导意义的数据规律。

2.3.1 处理方法：

利用 datatime 中的 striptime,将字符串的时间转化为 datetime 类型，再利用 timestamp 函数获取时间戳，得到可以进行比较的数字。再利用字符串格式化的方式将 csv 文件中的每天乃至每个小时的数据提取出来。

2.3.2 一天中的情绪分布：

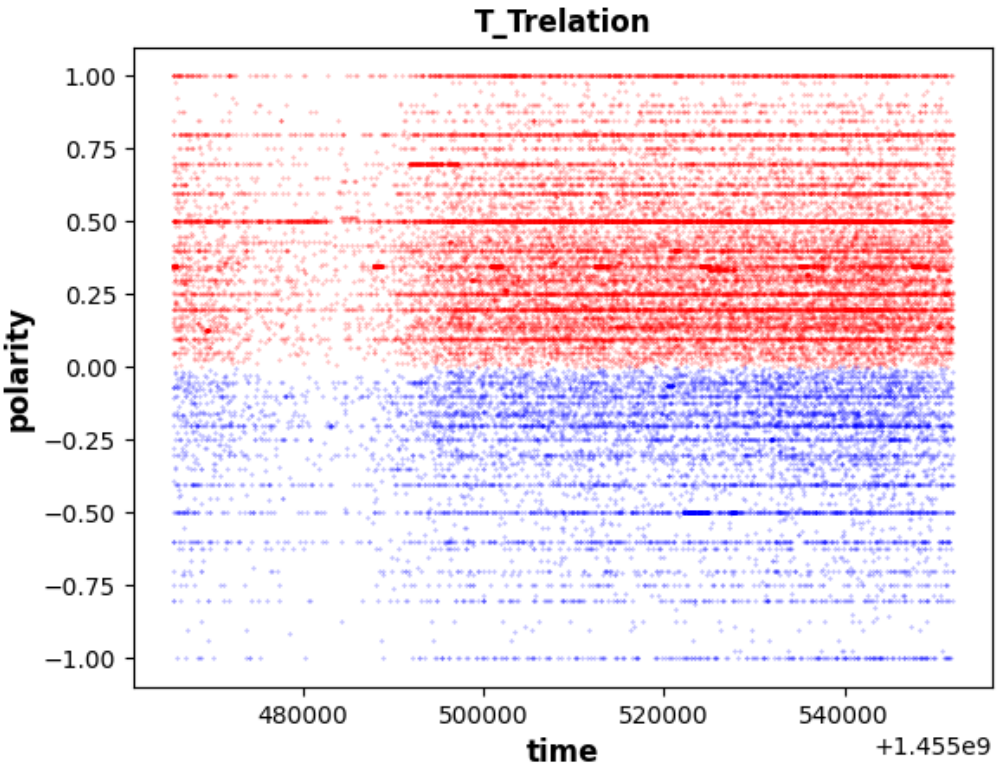


表 5 2016 年 2 月 15 日的情绪分布散点图

在图中我将极性为中性的数据剔除，保留了消极情绪和积极情绪的分布，可以明显地看出，在一天的时间里情绪分布有着很明显的规律性。

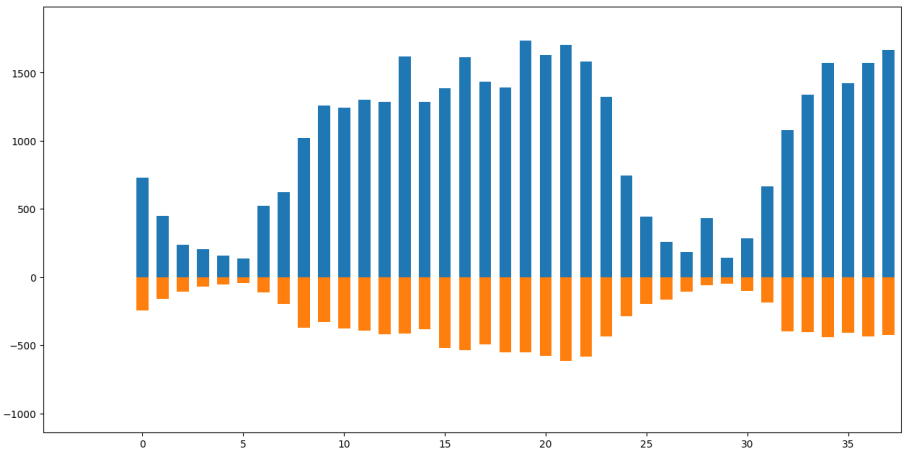
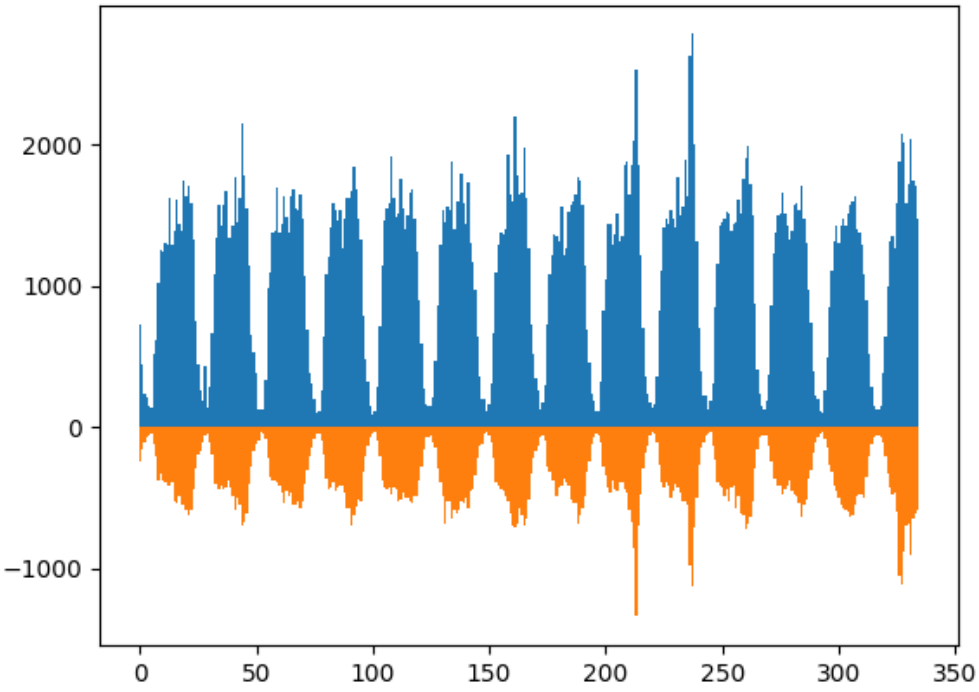


表 6 一天 24 小时，每个小时的消极情绪积极情绪的对比图（宽度为 0.6）

除了情绪的变化之外，其实也可以看出活跃度的变化，在报告的最后我会补



充相关的研究。

2.3.2 十三天每个小时的情绪（消极和积极）数量对比：

表 7 每个小时的消极情绪和积极情绪的对比图（宽度为 1）

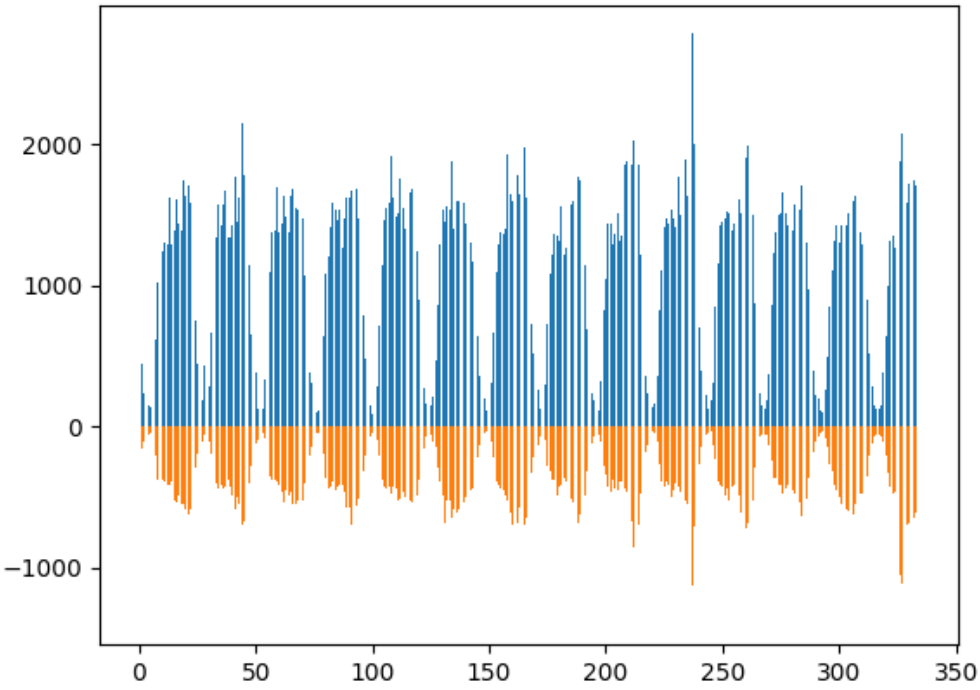


表 8 每个小时积极情绪和消极情绪的数量对比（宽度为 0.6）

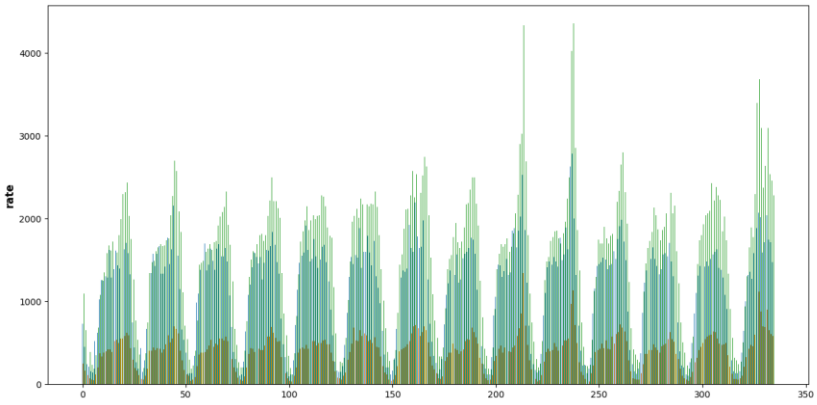


表 9 十三天每个小时，**消极**，积极，**中性情**绪的数量统计的柱状图

从表 7，8，9 可以看出虽然情绪数量的周期性具有很明显的特征，但是对于我们要研究的课题相违背，积极情绪的数量在每一天的每个小时都多于消极情绪数量。现做折线图更加直观地研究

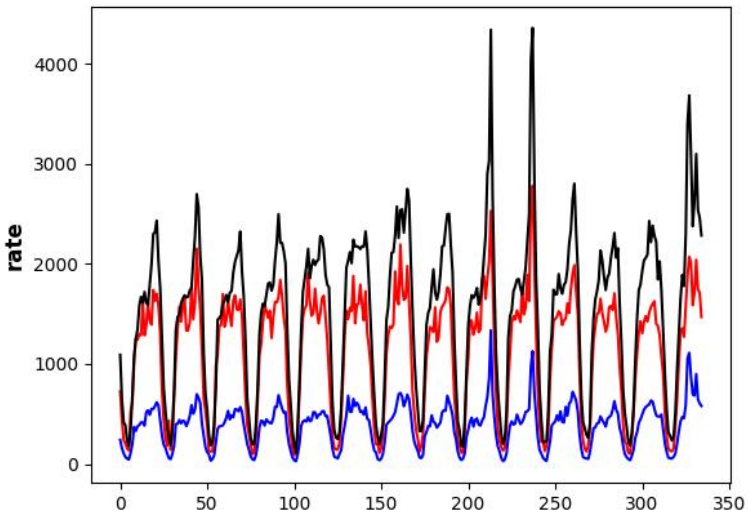


表 10 十三天每个小时，**消极**，积极，**中性情**绪的数量统计的折线图

从以上这些图，可以明显地看出来，积极情绪在每天的每一个小时的数量都是多于消极情绪的，甚至在白天上午 10 点到下午 3 点左右的时间，是远远多余消极情绪的。但也可以发现在夜晚时分除了推特数量的下降之外，消极情绪数量和积极情绪数量接近，也就是比值发生了明显的变化，先对消极情绪和积极情绪

的数量的比值进行研究。

2.3.3 十三天相同时段的情绪（消极和积极）数量对比

在之前的研究可以发现，情绪变化有着很明显的周期性，所以对十三天的同一时段的情绪数量取均值，进行 24 小时的情绪研究。

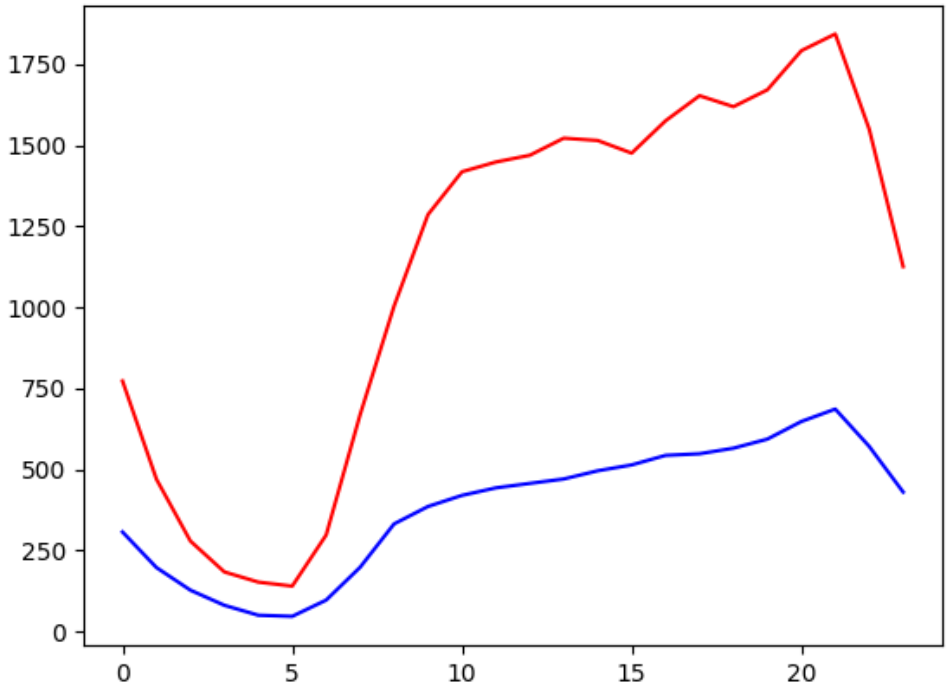


表 11 十三日同时段的消极情绪数量和积极情绪数量

在这一张图可以清楚地发现消极情绪和积极情绪数量关系，在凌晨 3 点左右，消极情绪和积极情绪同步进入一天中的最低点，但两者数量相差不多。现做出消极情绪数量和积极情绪数量的比值变化。

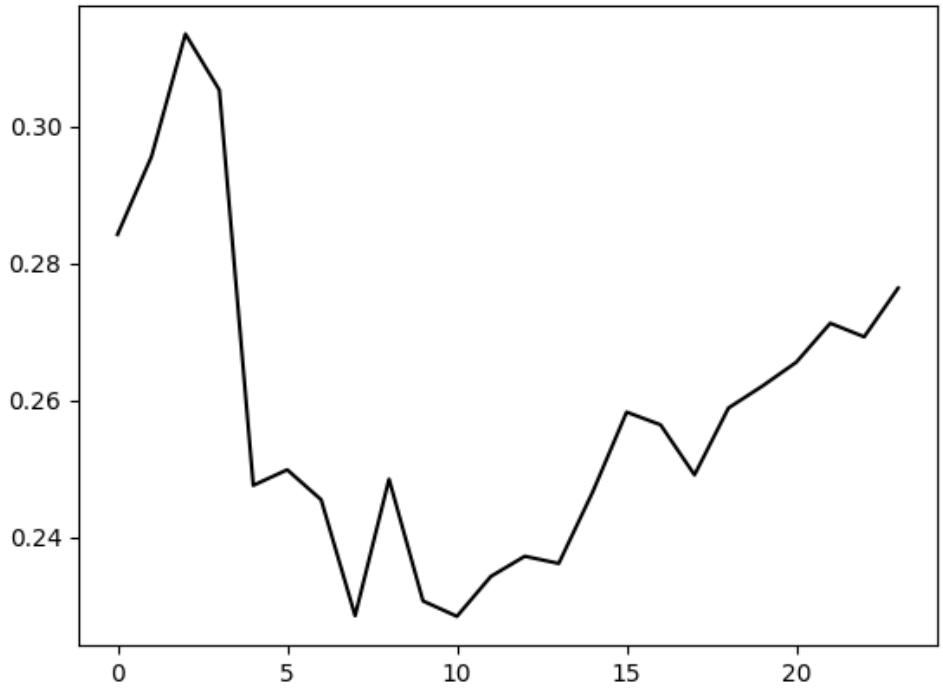


表 12 消极情绪数量与积极情绪数量比值

由图易得，在凌晨时分消极言论数量和积极言论数量的比值到达最大值。这就可以说明为什么在夜晚，社交平台上的消极言论数量感觉多于积极言论。在实际统计中可以发现，其实不是真的多于积极言论，而是相比于白天，消极言论占言论总数的比值大大地增加，所以会出现感觉消极言论数量多于积极言论的感受。

3. 情绪变化研究总结：

3. 1 数据分析的逻辑性：

从最开始的简单研究数据量，对整体的数据规模有了一定的认识和了解，到之后的加入时间戳，将时间细化，分析每个小时，每个时间段的数据数量特点，再到后面联系数据和现实生活的体验，整个数据分析的过程是循序渐进的，是具有研究逻辑和研究思维的。

3.2 数据分析的结论：

1.推特内容发布具有周期性

2.推特平台上的三种情绪的数量比较：中性情绪数量 > 积极情绪数量 > 消极情绪数量。结合生活，其实大家在社交平台上发布的内容都基本是积极乐观，会把自己好的一面展现给大家，不会在网络上发布很多的消极情绪。

3.推特平台上的积极情绪在当地时间 10 点到 15 点的时间段到达极大值。在 3 点左右达到极小值，同时也在凌晨三点左右消极情绪与积极情绪的比值到达最大。结合生活，在日常生活中，大家一到夜晚感觉社交平台上的消极言论会变得多起来，甚至感觉多于积极言论，只是消极言论的比值多于了积极言论，所以让人感觉到了压抑和忧郁的感觉。

3.2 本次数据分析的缺陷：

1.本次数据分析使用的 NLP 是 Textblob，由于是 python 的开源库，可能对于分析推特数据的针对性没有那么强，在一些语句的分析中可能出现很大的误差。

2.此外情绪分析的样本规模只有 100 万组，得出十分准确的结论还远远不够。

3..对于数据分析自然语言库的问题，我借鉴开源网站的思路，并利用 nltk 中电影评论为训练集，进行监督学习，得到一个简易的分类器。并在遍历的过程中进行数量统计，得出柱状图如图。

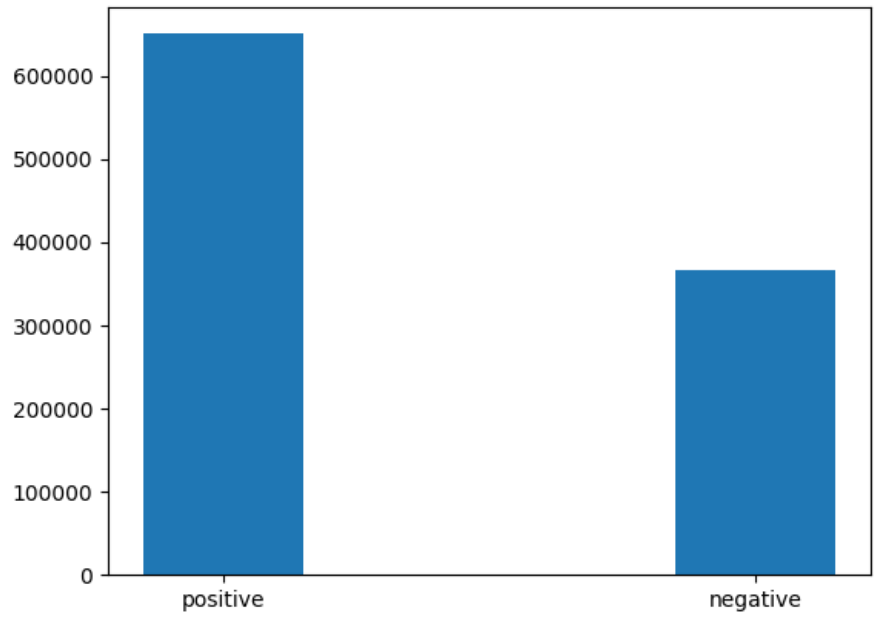


表 13 机器学习模型下的数据统计

但由于期末周的任务较多，没有将该种思路延续以及深入研究。

4. 其他方面的研究：

4.1 一天内的活跃度变化：

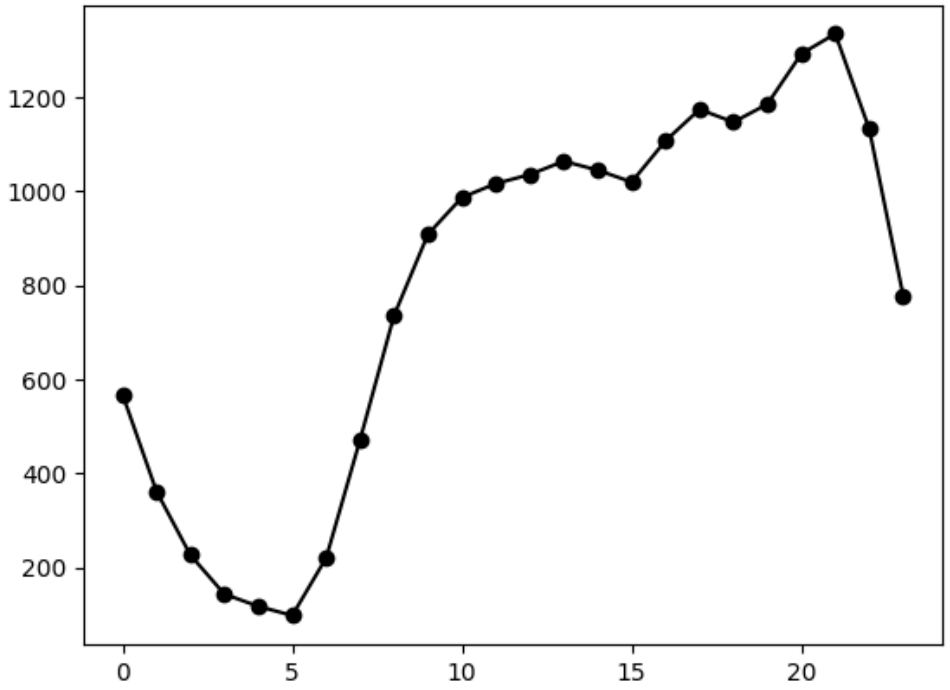
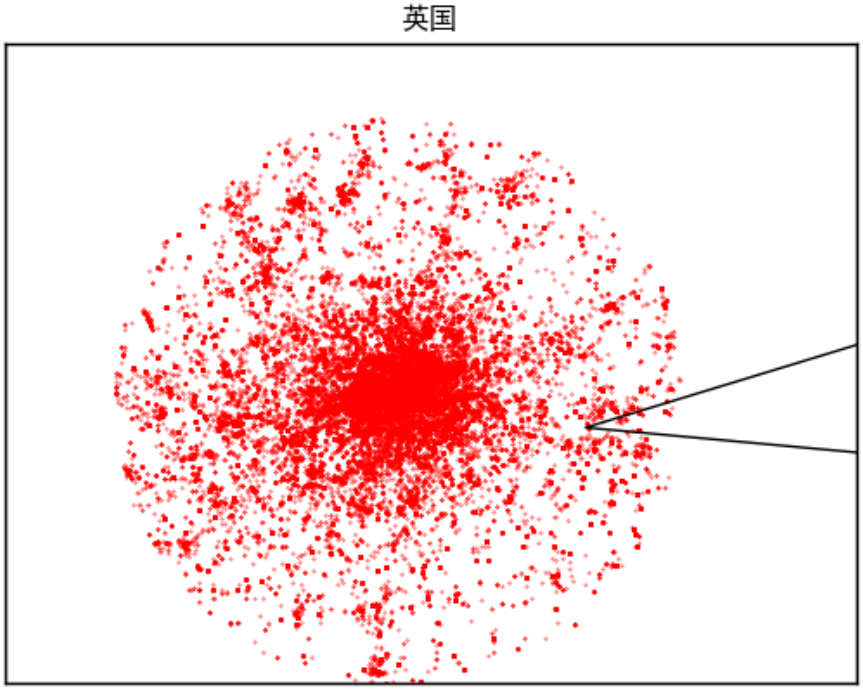


表 14 活跃度均值的变化

统计每个小时的推特发布的数量来衡量推特的活跃度，可以发现推特在 20 点左右活跃度最高，在 5 点左右活跃度最低。

4.2 特发送地点的热点图



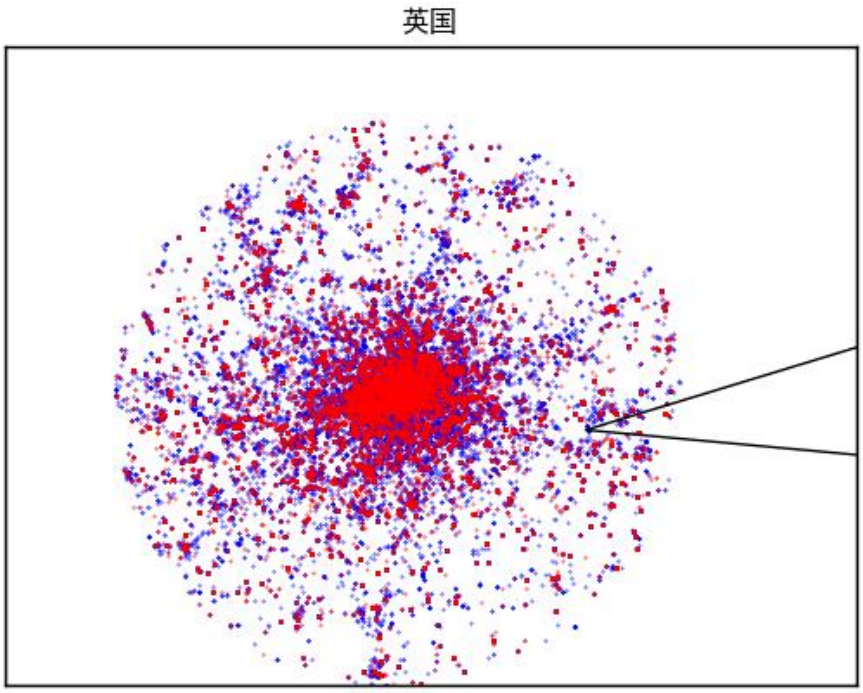


表 15 推特发送地点的热点图

可以发现在英国伦敦中心部分的推特发布较为集中，在郊区部分发布较为稀疏，这可能和人口密度也有一定关系。

4.3 好友数量的研究

4.3.1 好友数量的简单统计：

由表 16 可知，推特好友数量大多在 1000 以下，但也有很多推特用户的好友数在 1000 到 2000 之间，甚至还有推特用户的好友数达到了恐怖的 8000.个体差异较大。

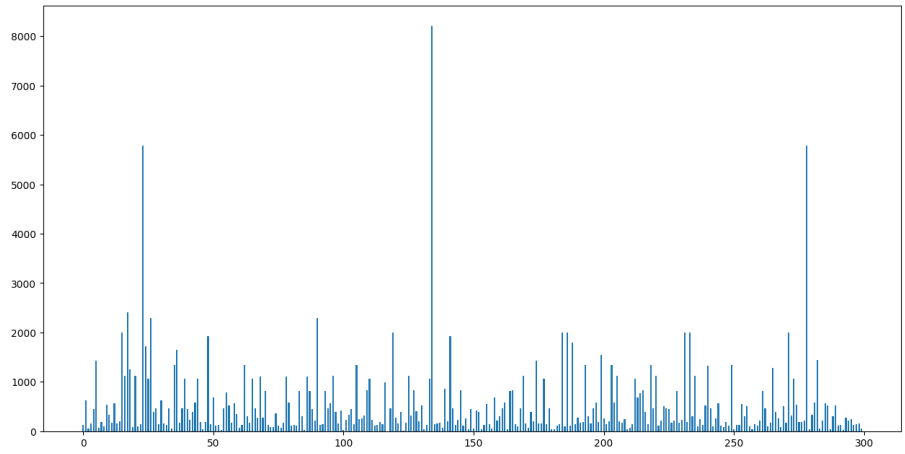


表 16 好友数量的柱状图

4.3.1 好友数量和发布内容的关系：

使用了 textblob 对推特用户的发布内容进行分析，返回情绪的极性，并用点的大小表示好友数目的多少，可以发现，好友数目越多的发布的内容更为积极和中性。

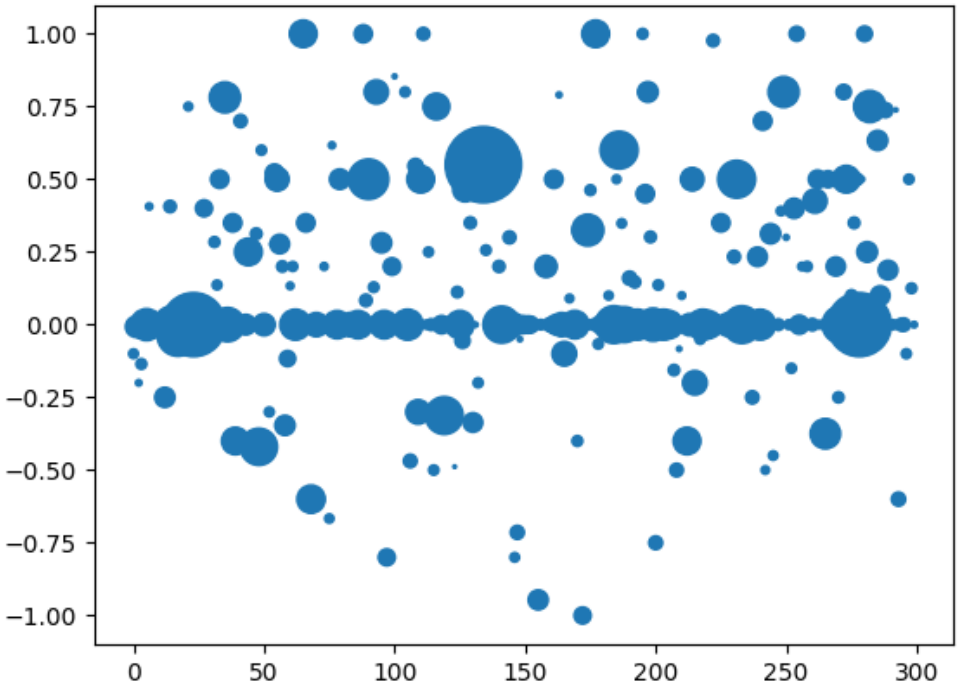


表 17 好友数量和发布内容的关系（**点越大代表好友越多**）

5. Python 数据分析的收获：

1. 数据提取，例如 csv，json 等文件的数据提取，在 json 文件中遇到 json 格式的不规范（例如空行等的出现）如何解决。
2. 利用 os 库操作不在同一个文件夹的文件
3. 掌握了基本的 NLP 库的用法，对自然语言进行简单的情绪分析，
4. 掌握了 datetime 类型转化为时间戳进行数字操作，并进行相应的可视化
5. 使用 matplotlib 对数据进行可视化，掌握了 plot，bar，scatter，basemap 等画图函数的基本应用
6. 简单应用了机器学习的理论，并用代码进行实现

在整个过程中，更多地是自学能力的提升，寻找资源能力的提升，很有收获。