数据收集

本报告的数据来源于3个不同的维度
1）项目提供的twitter_archive_enhanced.csv，其中包含了待分析的基本数据
2）项目提供的tweet_json.txt，其中包含了每条推特信息的转发数和点赞数等额外信息
3） 项目同时提供了一个图像预测文件image-predictions.tsv的下载链接

其中twitter_archive_enhanced.csv和tweet_json.txt直接拷贝到了项目目录下
而image-predictions.tsv会使用RequestAPI的方式从网络中获取

最终想分析的是tweet推广中获取一些思路（从一个新账号，做到头部账号的方式），基于这样的目的进行数据收集及清洗

评估数据

在收集到数据之后，我们会通过编程和肉眼分析的方式，来初步定为数据的整洁度和数据质量问题

数据质量

数据质量从下面几个维度考虑：
- 完整性
- 有效性
- 准确性
- 一致性

archive 表格

1）狗的名字存在一些错误的单词（such,quite,not,very,just,my,his,one,a,an）
2）需要过滤掉转发的记录，只保存原始评分的tweet记录
3）字段属性不正确（retweeted_status_id，retweeted_status_user_id）
4）timestamp的数据类型应该为datetime，而不是object
5）对于“空字段”的表示，可以统一。（NaN，None）

image 表格

1）被分析的图片的url有重复
2）数据缺失，archive中有2356条记录，而images中只有2075条记录

tweet_json 表格

数据只有2352条，与archive中的数据不一致

数据整洁度

1）狗的种类，合并成一列，将具体的类型填写其中
2）应该将3个表格的信息进行合并

清理数据

根据上面分析的质量和整洁度问题，进行数据的清理

1）首先将转发的内容删除掉，只留下原始的tweet记录
2）同时将没有图片分析的数据也删除掉
3） 将狗狗的类型进行合并，只保留一列
4） 将基础的数据整理之后，保存一个原始文件twitter_archive_handled.csv
5）为了得到转发以及图片质量的关系，生成了一个按照月统计的数据twitter_month_data.csv