数据收集
#1.WeRateDogs 的推特档案，作为 “资料来源：手头文件集#
#2.收集推特图像的预测数据#
#3.每条推特的额外附加数据，由于无法登陆推特所以直接下载 udacity 提供的数据，作为 “资料来源：手头文件集#
数据评估
#1.存在部分转发的推文#

#2.expanded_urls 存在缺失值#

#3.tweet_id 数据类型错误，将整数类型改为字符串#

#4.in_reply_to_status_id 和 in_reply_to_user_id 缺失项较多，需要删除#

#5.rating_numerator 列应该是 float 类型

#6.对于时间戳 timestamp 而言，数据应该是 datetime 类型；

#7.评分的分母不统一，有些过大#

#8.name 中存在一些提取错误的名字，比如 a, an, the 等

整洁度：
#（1）根据 tidy data 的第一条原则，变量按列来组织（即：一个变量一列），doggo, floofer, pupper, puppo 这四列应该合并为一列

#（2）三个数据集都是以 tweet_id 为观察单位，所以合并为一个单独的表格

数据清洗
质量
#1.存在部分转发的推文#
用 str.find()在 text 列找出含有“RT @”的信息（即转发数据）,之后将剩下的数据重新赋值给 twitter_archive_enhanced_clean
#2.expanded_urls 存在缺失值#
用 drop 将 expanded_urls 删除
#3.tweet_id 数据类型错误，将整数类型改为字符串#
用.astype('str')将 twitter_archive_enhanced_clean 和中的 tweet_id 数据类型改为字符串
#4. in_reply_to_status_id 和 in_reply_to_user_id 存在缺失值
用 drop 删除 in_reply_to_status_id 和 in_reply_to_user_id
#5.rating_numerator 列应该是 float 类型
用.astype(' float ')将 twitter_archive_enhanced_clean 和中的 rating_numerator 数据类型改为 float 类型
#6.对于时间戳 timestamp 而言，数据应该是 datetime 类型；
用 pd.to_datetime 将时间戳 改为 datetime 类型

#7.评分的分母不统一，有些过大。但是过大的分母，并不一定的错误的，分母不等于10 的 text 原文中有一些没有被约分的分数，比如说 99/90，其实是指 9 只 11/10 的狗狗评分，这样的评分可以通过将分子除以分母获得一个 rating 列来解决
#8.name 中存在一些提取错误的名字，比如 a, an, the 等
用 replace 错误的名字删除

整洁度
#1. 用.str.join 将 doggo, floofer, pupper, puppo 这四列应该合并为一列#
#2. 三个数据集都是以 tweet_id 为观察单位，所以用 merge 合并为一个单独的表格

数据可视化：导入 WorldCloud 库