wrangle_report.pdf

获取数据

- 获取神经网络预测数据 image-predictions.tsv
  主要使用 response = requests.get(url)
  file.write(response.text)

解析 json 数据，提取转发和点赞数据

- 使用 with open 读模式打开文件，逐行使用 json.loads(line)转换为 dict 形式

for line in file.readliens():
    line = json.loads(line)

取出 line[‘retweet_count’ , ‘favorite_count’ , ‘expanded_url’ ,’ media_url_https’]

- 最后使用 pandas.DataFrame 转换为 dataframe

清理数据

- object 到 datetime64 类型转换,使用 pandas.to_datetime 函数
  pandas.to_datetime(df.column)

- 过滤回复、转发、为 NaN 的推文。主要使用 df.column.isnull() 或者 ~(df.column.isnull())取 is not null 的数据集
  twitter_archive_clean =
  twitter_archive_clean[twitter_archive_clean.in_reply_to_status_id.isnull()]

- 推文来源列包含 html 处理，主要使用正则配置出正确的内容，更新会对应索引列值

```
for idx, row in twitter_archive_clean[['source']].iterrows():
    data = re.findall(r'>(.+?)<', row['source'])
    twitter_archive_clean.loc[idx,'source'] = data[0]
```

- 分子 rating_numerator 错误值纠正。主要正则配置类型 num/num 类型数据 和 rating_numerator 值比对，不相等取出对应行索引、num/num、rating_numerator 的值到列表 wrong_numerator 中
  1. getWrongRateNumerator(twitter_archive_clean)
  2. wrong_numerator 内容：

  [{'index': 45, 'text_rates': ['3.5/10 '], 'current_rating_numerator': 5},

{'index': 340, 'text_rates': ['9.75/10 '], 'current_rating_numerator': 75},
{'index': 695, 'text_rates': ['9.75/10 '], 'current_rating_numerator': 75},
{'index': 763, 'text_rates': ['1.27/10 '], 'current_rating_numerator': 27},
{'index': 1712, 'text_rates': ['1.26/10 '], 'current_rating_numerator': 26}]

3. 更新错误的 rating_numerator 值

```
for item in wrong_numerator:
    twitter_archive_clean.loc[item["index"], "rating_numerator"] =
item["text_rates"][0].split('/')[0]
```

- 处理 name 列值，只是回补一些 name
  1. 推文中包含 named 关键字的，取 named 后面一个单词 当成 name 列值
  2. 现有的 name 列值取出到列表中（排除 a,an,O,the），取首字母大写，name 长度大于 1 的那么值逐个判断列表 name 值是否包含在每个推文中，包含，更新 name 列为列表对应列表 name 值。

- doggo / floofer / pupper / puppo 合并成一列 stage
  1. 补充数据：使用正则 re.findall(r'doggo|floofer|pupper|puppo', df.loc[idx,:].text.lower())匹配，更新对应 oggo / floofer / pupper / puppo 列值
  2. 由于这 4 列 None 值过多，不适合使用 pandas.melt 函数，所以使用 dataframe.iterrows() 循环取出 4 列非 None 值，更新到 stage 列

- 多个 dataframe 融合，使用了 pandas.merge

```
pd.merge(twitter_archive_clean, tweet_retweet_favorite_clean, on='tweet_id', how='left')
```