有关二〇二三年四月至二〇二三年九月推特数据爬虫相关工作的说明

说明人：蔡明方、胡骋

1. 爬虫工作概述

自2023.4到2023.9这个时间段内，共先后调研数十种爬虫方法，最终选出两种可行方法，并先后使用爬取了以“Uyghur”、“Tibet”、“nuclear”以及“One Belt One Road”在内的超十万条数据，并在爬取后对数据进行了去重等处理。

二、爬虫方法

1、stweet工具包

经过调研，发现GitHub上开源的工具包Stweet可以自由爬取任意时间段任意关键词和任意数据以及任意推文id号的相关信息（包括但不限于作者姓名，简介，IP地址，账号创建时间，推文内容，发布时间等）相关代码见附件一。

2、twarc工具包

Twarc工具包为推特官方的爬取工具包，通过在Google开发者平台上注册账号并申请或开通爬虫服务，然后通过命令行的方式进行爬取。具体方法如下：

先使用pip命令安装twarc和twarc\_csv:

Pip install twarc

Pip install twarc\_csv

然后再使用如下命令进行爬取：

Twarc2 search “关键词” 关键词.jsonl

然后可以通过如下命令将jsonl文件转换为csv文件：

Twarc2 csv 关键词.jsonl 关键词.csv

更多的操作命令见如下链接：

<https://twarc-project.readthedocs.io/en/latest/twarc2_en_us/#search_1>

三、数据处理工作：

由于爬取下来的数据样式种类繁杂，所以对爬取下来的数据进行了处理，所做的处理包括但不限于按推文去重，去掉url以及修改@用户等。具体代码见附件二。

四、已知的问题

1、由于在2023年6月份推特进行了全面更新，导致原先的工具包Stweet暂时无法使用，可以等待其更新后再使用。

2、twarc工具包所需要的Google开发者平台的账号是需要花钱开通爬虫服务的，FREE版本每月可以爬取1000条，basic版本每月可以爬取5000条，pro版本每月可以爬取100000条。如果申请学术api可以每月爬取200000万条，之前两个月是有学术api的，但由于种种未知的原因导致现在学术api不可用，可再次申请通过后可再次使用。

3、由于twarc工具包将数据转换为csv格式后的解码方式为python语言无法识别的，故在解析时会出现部分乱码，导致在去重时会去掉这部分无法解析的推文使得推文数量有一定的减少，但根据之前的经验，影响可以忽略不计。