河南洪灾微博数据收集

Part 1 简要介绍

数据分析大致思路

通过收集新浪微博平台上关于河南洪灾的信息，具体包括微博的文本内容、点赞、评论及转发信息，以及微博用户的个人信息，结合文本挖掘和图模型，进行后续分析。

Part 2 收集内容

一. 微博信息

对符合筛选条件的每一条微博信息进行爬取，共14个字段，具体为微博ID，微博用户ID，微博用户名，话题，是否原创，正文，发布渠道，发布工具，发布地点，发布时间，转发数，评论数，点赞数及微博爬取时间；以下为微博Item类字段定义：

▼ 微博Item

```
class CrawlerWbItem(Item):
    # define the fields for your item here like:
    """weibo_id, uid, user_name, tag, content, source"""
    weibo_id = Field()
    uid = Field()
    user_name = Field()
    tag = Field()
    is_origin = Field()
    content = Field()
    source = Field()

    publish_tool = Field()
    publish_place = Field()
    publish_time = Field()

    num_repost = Field()
    num_comment = Field()
    num_attitude = Field()
    time_crawl = Field()
```

【微博ID，微博文本内容，发布微博用户ID，微博发布时间，微博发布地点，微博发布渠道，是否为原创微博，微博话题标签，微博点赞数，微博评论数，微博转发数】

二. 微博评论信息

对微博的每一条评论进行爬取，共5个字段，具体字段包括该条评论所属微博的微博ID，该条评论ID，评论用户ID，评论内容，评论点赞数。以下为微博I评论tem类字段定义：

▼ 微博评论Item

```
class CrawlerWbCommentItem(Item):
    # define the fields for your item here like:
    weibo_id = Field()
    comment_id = Field()
    comment_uid = Field()
    comment_content = Field()
    num_comment_attitude = Field()
```

三. 微博点赞信息

对微博的每一条点赞进行爬取，共2个字段，具体字段包括该条点赞所属微博的微博ID，点赞用户ID。以下为微博点赞Item类字段定义：

▼ 微博点赞Item

```
class CrawlerWbAttitudeItem(Item):
    weibo_id = Field()
    attitude_uid = Field()
```

四. 微博转发信息

对微博的每一条转发信息进行爬取，共2个字段，具体包括该条转发的原微博ID，该转发微博的ID。以下为微博转发Item类字段定义：

▼ 微博转发Item

```
class CrawlerWbRepostItem(Item):
    weibo_id = Field()
    repost_weibo_id = Field()
```

五. 微博用户信息

对爬取微博用户的个人资料页进行爬取，共13个字段，具体包括用户ID，昵称，性别，地址，生日，描述，认证信息，教育，工作，标签，微博数量，关注数量，粉丝数量。以下为微博用户Item类字段定义：

▼ 微博用户Item

```
class CrawlerWbUserItem(Item):
    uid = Field()
    nickname = Field()
    gender = Field()
    location = Field()
    birthday = Field()
    description = Field()
    verification_reason = Field()
    education = Field()
    job = Field()
    tags = Field()

    num_weibo = Field()
    num_followings = Field()
    num_followers = Field()
```

Part 3 收集策略

一. 爬取流程

基于配置好的种子用户ID，

1. 通过其ID访问用户个人信息主页，对个人信息进行爬取
2. 然后访问个人微博主页，通过对用户微博进行关键词筛选，获得该用户下的可爬取微博；
3. 对满足判定条件的微博，对微博信息、点赞、评论和转发信息进行爬取和保存
    a. 在对微博信息进行爬取时，如该微博不是原创微博，会获取其原创微博的发布用户ID，把该用户的个人信息和微博也加入爬取队列；
    b. 在对微博的转发信息进行爬取时，同时会获取转发微博的用户ID，把该用户的个人信息和微博也加入爬取队列；

二. 种子用户筛选过程

【策略】

根据拟定的收集策略，种子用户的设定目的是基于种子用户的微博及微博互动信息，对研究话题展开广泛的微博信息收集，因此种子用户需满足其微博信息具有较高的曝光度。

基于此要求，先对微博话题使用关键词“河南暴雨”进行搜索，结果共计589条相关话题；
筛选阅读量大于1亿的话题有“河南暴雨互助”等共28个；
选取该28个话题的话题主持人、话题贡献排行前8位中的1个机构账号与1个个人账号作为种子账号。
基于阅读量排行筛选得到的种子用户，可满足曝光度高的要求。

▼ 阅读量超过一亿的话题

#河南暴雨互助#

#河南暴雨互助信息#

#河南暴雨救援#

#河南暴雨24小时后#

#暴雨后救命文档创建者是河南籍大学生#

#河南暴雨汛情#

#河南暴雨后又将迎战台风烟花#

#河南暴雨车险损失近10亿元#

#河南暴雨辟谣消息汇总#

#致敬河南暴雨平民英雄#

#聚焦河南极端暴雨#

#未来3天河南西北部有大暴雨#

#暴雨中的河南力量#

#150秒看河南本次极端暴雨#

#河南暴雨下的中国人#

#分享河南暴雨中可爱的平凡人#

#一张图告诉你河南暴雨有多大#

#河南会再次出现极端暴雨吗#

#国务院公布调查河南暴雨举报电话#

#河南暴雨将至高架秒变停车场#

#直接河南新一轮暴雨#

#河南本轮强降雨与720暴雨有何不同#

#河南多地暴雨致灾#

#河南暴雨直击#

#河南暴雨救援电话#

#河南暴雨救援最全电话公布#

#河南这次暴雨为什么这么强#

#航拍暴雨后的河南#

种子用户

Aa 话题	# 阅读	# 讨论	主持人	主持人属性	机构	个人
#河南暴雨互助#	170.1	2462.3	7308140398	机构	1982753487	7527950457
#河南暴雨救援#	65.3	1029.1	1644729004	机构	7546593017	3812400789
#河南多地暴雨致灾#	23.3	1875.2	2803301701	机构	1314608344	5134003829
#河南暴雨直击#	9.2	12.5	1642512402	机构	1642088277	1982753487

Aa 话题	# 阅读	# 讨论	主持人	主持人属性	机构	个人
#一张图告诉你河南暴雨有多大#	9.2	1856.5	2803301701	机构	1240041543	2655176573
#河南这次暴雨为什么这么强#	5.2	9.7	1828887503	机构	1672519561	3957305231
#河南暴雨救援电话#	4.5	66.8	1618051664	机构	1632060204	6093254257
#河南暴雨24小时后#	2.8	2.3	5588388976	个人	5872812692	2349255395
#国务院公布调查河南暴雨举报电话#	2.5	0.6404	1784473157	机构	1644114654	-
#暴雨后救命文档创建者是河南籍大学生#	2.1	2	2656274875	机构	2591595652	5328002537
#分享河南暴雨中可爱的平凡人#	1.9	4.3	6560653260	个人	1795548274	2106558073
#聚焦河南极端暴雨#	1.9	2.7	1644114654	机构	6463753619	5688313870
#河南暴雨辟谣消息汇总#	1.8	2.8	7373111326	个人	2518353303	7278168925
#致敬河南暴雨平民英雄#	1.8	3	1618051664	机构	-	2511776195
#河南暴雨救援最全电话公布#	1.7	21.1	2810373291	机构	1887344341	6871390978
#河南暴雨后又将迎战台风烟花#	1.6	0.87	6365990425	机构	1796087453	7440276715
#河南本轮强降雨与720暴雨有何不同#	1.6	0.4579	1688864597	个人	2616360197	7628105261
#暴雨中的河南力量#	1.5	125.4	7308140398	机构	1764222885	5120552390
#河南暴雨汛情#	1.5	1.1	1726918143	机构	5453011460	5401012250
#150秒看河南本次极端暴雨#	1.4	1.3	1644114654	机构	1807812760	6422353022
#河南暴雨互助信息#	1.4	18.8	2656274875	机构	1974576991	7064041584
#河南暴雨将至高架秒变停车场#	1.4	1.1	7453682638	个人	2309846073	5723342159
#河南会再次出现极端暴雨吗#	1.3	0.6319	3266943013	机构	1651428902	6020839748
#未来3天河南西北部有大暴雨#	1.3	2.6	1699540307	机构	1883755941	1812175903
#河南暴雨下的中国人#	1.3	1.9	2803301701	机构	6004281123	1943393985
#河南暴雨车险损失近10亿元#	1.3	0.3859	2258727970	机构	1622115061	7332032604
#直击河南新一轮暴雨#	1.2	3921	7308140398	机构	1823287210	7576657188
#航拍暴雨后的河南#	1.1	1.3	2656274875	机构	-	2700138820

三. 微博筛选条件

对个人微博主页进行逐一、逐页筛选和爬取时，遵循如下几个条件：

1. 微博发布时间区间：2021.7.10-爬取当日

   由于该任务是对2021年7月河南暴雨这一特定事件进行相关微博内容收集，河南暴雨因此限定只对2021.7.10起发布的微博内容进行收集。据维基百科“2021年7月河南水灾”词条显示，2021年7月17日以来中国河南发生特大降雨导致水灾，因此将微博收集时间点往前推一周，收集暴雨来临前一周的讨论内容。

2. 微博内容

   针对该收集目的，对微博正文或微博标签的文本内容进行判定，如包含任一指定关键词，则即认为属于相关微博，可进行后续爬取。关键词列表如下：

```
'河南',
'郑州',
"浚县",
'淇县',
'卫辉',
```

```
'新乡',
'暴雨',
'求助',
'互助',
("郑州", "暴雨"),
("河南", "暴雨"),
("卫辉", "暴雨"),
("浚县", "暴雨"),
"河南暴雨互助",
"河南暴雨",
"河南暴雨互助[超话]",
"河南暴雨求助",
"河南暴雨互助信息",
"卫辉内涝",
"卫辉求助",
"河南求助",
```

四. 部分说明

1. 个人微博主页微博顺序

   由于存在热门微博或置顶微博，在个人微博首页需设置最少爬取微博数；据观察weibo.cn端只存在微博置顶机制，其余微博按照发布时间由近到远的顺序排列，因此可设置至少爬取该用户的前3条微博，避免因置顶微博发布时间不符筛选条件导致爬取提前结束的问题。

2. 转发微博的微博信息保存是否包含原始微博

   经确认，包含。

3. 转发微博的微博信息是否能与原始微博信息做拆分

   转发微博文本内容符合 “xxx转发了xxx的微博：[原微博]，转发理由：[该用户转发时发表的内容]”的格式，因此在内容保存上转发理由与原微博信息是拆分开的。示例：

```
"Cassiopeia_Dorothy     转发了 @凤凰天使TSKS 的微博:【#凤凰天使TSKS# & @半为苍生半美人吧_沈昌珉 】◇NAVER NOW.综艺◇《 #最强昌珉的Free Hug# 》[E17.2108
转发理由:    //@半为苍生半美人吧_沈昌珉:"
```