* 简介

微博分类是对微博账号赋予一个类别标签。例如美食、影视等。因为博主关注的兴趣点不尽相同。有人喜欢发关于电影的微博、有人喜欢发关于财经的微博等。有些博主不是个人而是一个机构组织，这样的类别区分更明显。

微博分类主要目的是进行用户定向，也就是给博主打上标签。在给用户推荐内容的时候可以使用打上的标签，也可以统计关注不同类别内容的比例。

类别体系采用里的类别体系。

所有的类别大致如下：



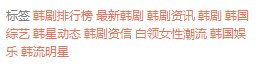
对一个微博账号进行分类的依据是该博主的简介、标签、meta信息

简介即博主自己写的简单介绍。例如：



从简介可以看出该博主大致喜欢发笑话的微博。

标签即博主自己制定的一些定向关键词例如：



通过这些标签可以看出博主大致喜欢韩流方面的微博

Meta信息：即微博的title、keywords、description。

例如上面两位博主的meta信息分别为：

第一位：

<title>冷笑话精选的微博\_微博</title>

<meta content="冷笑话精选，冷笑话精选的微博，微博，新浪微博，weibo" name="keywords" />

<meta content="冷笑话精选，搞笑大奇葩，吐槽老专家，认真讲笑话，冷哥帅掉渣！。冷笑话精选的微博主页、个人资料、相册，寻找设计、程序、编辑。新浪微博，随时随地分享身边的新鲜事儿。" name="description" />

第二位：

<title>韩剧情报网的微博\_微博</title>

<meta content="韩剧情报网，韩剧情报网的微博，微博，新浪微博，weibo" name="keywords" />

<meta content="韩剧情报网，wei\_xin账号：hjqbw\_2012；合作QQ1102153120 专注韩剧，及韩剧相关最新最快信息。韩剧情报网的微博主页、个人资料、相册。新浪微博，随时随地分享身边的新鲜事儿。" name="description" />

从meta信息里也能看出两位博主的兴趣爱好。

* 输入输出
* 输入:

微博账号

简介

标签

Meta信息

* 输出:

该微博帐号的类别预测

* 样本采集
* 使用Google Chrome浏览器。登陆微博，点击发现，进入微博分类浏览页面。点击其中的一个分类，例如星座。

工具->JavaScript控制台：

输入下面的脚本：

var items=document.getElementsByClassName('WB\_info');

for(var j=0;j<items.length;j++) {

var suba=items[j].getElementsByTagName("a");

var suba\_href=suba[0].getAttribute("href");

var suba\_nick=suba[0].getAttribute("nick-name");

var suba\_card=suba[0].getAttribute("usercard");

console.log(suba\_card+"\t"+suba\_nick+"\t"+suba\_href);

}

var items=document.getElementsByClassName('WB\_info');

console.log(items.length);

for(var j=0;j<items.length;j++) {

console.log(items[j]);

}

获得星座分类下的微博帐号，例如输出结果如下：

id=2606843472 白羊座-心理学 <http://weibo.com/u/2606843472>

id=2484832905 精选星座微博 <http://weibo.com/xingzuoer>

id=2259504050 教你易经风水术官网 <http://weibo.com/jnyjfss>

id=1731285993 塔羅星座艾菲爾老師 <http://weibo.com/iefeel>

id=2461933493 星座风水命理 <http://weibo.com/u/2461933493>

id=2288577367 每天学点风水学 <http://weibo.com/u/2288577367>

id=2606923882 双鱼座-心理学 <http://weibo.com/u/2606923882>

id=2606919782 水瓶座-心理学 <http://weibo.com/u/2606919782>

id=1733779624 星座宝典 <http://weibo.com/startips>

id=2689629361 水瓶座心情专区 <http://weibo.com/u/2689629361>

id=2689634741 双子座心情专区 <http://weibo.com/u/2689634741>

id=1858907320 天蝎座秘语 <http://weibo.com/txzmy>

id=2606880752 巨蟹座-心理学 <http://weibo.com/u/2606880752>

id=2606875360 双子座-心理学 <http://weibo.com/u/2606875360>

id=2122237241 双鱼座蜜语 <http://weibo.com/aishipin>

点击下一页，获得尽量多的该分类下的账号。

* 在Google Chrome浏览器随机输入一个上面的url。例如<http://weibo.com/u/2606923882>

执行下面的脚本获得meta信息：

var info="";

var title=document.title;

info=info+"title:"+title+"\t";

var metas=document.head.getElementsByTagName("meta");

for(var j=0;j<metas.length;j++)

{

var name=metas[j].getAttribute("name");

if(name=='keywords')

{

var keywords=metas[j].getAttribute("content");

info=info+"keywords:"+keywords+"\t";

}

if(name=='description')

{

var description=metas[j].getAttribute("content");

info=info+"description:"+description+"\t";

}

}

该页面的输出结果是：

"title:双鱼座-心理学的微博\_微博 keywords:双鱼座-心理学，双鱼座-心理学的微博，微博，新浪微博，weibo description:双鱼座-心理学，解析双鱼座的心理学！~~~~~~~~~ 喜欢就 【✚加关注】｡◕‿◕｡ @双鱼座-心理学。双鱼座-心理学的微博主页、个人资料、相册。新浪微博，随时随地分享身边的新鲜事儿。

* 执行下面的脚本获得简介和标签信息：

var info="";

var items=document.getElementsByClassName('item\_text W\_fl');

for(var j=0;j<items.length;j++) {//(2/3,无标签)

if(items[j].innerHTML.indexOf("简介")>=0)

{

// info=info+items[j].innerText.replace(/(^\s\*)|(\s\*$)|(\n)/g, "")+"\t";

info=info+items[j].innerText+"\t";

//console.log(items[j].innerText);

}

if(items[j].innerHTML.indexOf("标签")>=0)

{

info=info+"标签:\t";

var suba=items[j].getElementsByTagName("a");

for(var k=0;k<suba.length;k++)

{

info=info+suba[k].innerText+"\t";

//console.log(suba[k].innerText);

}

// console.log(items[j]);

}

}

console.log(info);

该页面的输出结果是：

简介： 解析双鱼座的心理学！~~~~~~~~~ 喜欢就 【✚加关注】｡◕‿◕｡ @双鱼座-心理学 标签: 双鱼座 双鱼座心理学 双鱼座运势 双鱼座专区 星座 心理学 十二星座

这样获得每个类别下的样本信息。

最后获得样本格式如下：

http://weibo.com/renminfayuanbao 爆料 人民法院报 人民法院报，人民法院报 人民法院报，《人民法院报》。人民法院报。

http://weibo.com/u/3508612897 爆料 京法网事 京法网事，京法网事 京法网事，北京法院网。京法网事。

http://weibo.com/qindashao 财经 电商大战 电商大战，电商大战 电商大战。电商大战,武汉理工大学，深圳市腾讯计算机系统有限公司。

http://weibo.com/zuoluogufeiyang 财经 佐罗投资札记 佐罗投资札记，佐罗投资札记 佐罗投资札记，厚恩投资执行董事、总经理。佐罗投资札记，北京厚恩投资管理有限公司。 投资修行者。

http://weibo.com/qsgt 财经 有限次重复博弈 有限次重复博弈，有限次重复博弈 有限次重复博弈，专注小盘成长。有限次重复博弈，阿里酱油加工职业专科学校。

http://weibo.com/jinrongbaguanv 财经 金融八卦女 金融八卦女，金融八卦女 金融八卦女，三千线人。【微信账号jinrongbaguanv】八卦从来不是目的, 而是接近真相的手段之一。。金融八卦女,Columbia University，某基金公司。

http://weibo.com/renzhiqiang 财经 任志强 任志强，任志强 任志强，知名地产商。任志强。

http://weibo.com/wangshi 财经 王石 王石，王石 王石，万科企业股份有限公司董事会主席。王石。

http://weibo.com/wangran 财经 王冉 王冉，王冉 王冉，易凯资本有限公司CEO。王冉，易凯资本有限公司。

http://weibo.com/wu2198 财经 wu2198 wu2198，wu2198 wu2198，财经名博。wu2198，农业部。

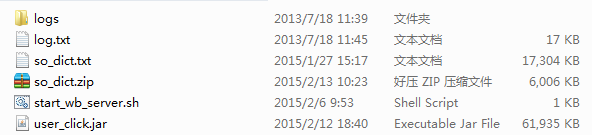
第一列是微博的url 第二列是该微博的分类 第三列是 meta、简介、标签信息 不同列之间用\t分隔

* 模型训练

过程参加模型训练文档

* 分类服务

部署微博分类服务需要的文件如下：



启动命令是：

sh start\_wb\_server.sh > log.txt 2>&1 &

start\_wb\_server.sh的内容是：

java -cp user\_click.jar cn.clickwise.clickad.classify\_pattern.ClassifyPatternServer -p 9008 -t 3 -d so\_dict.txt

可以看出该脚本调用ClassifyPatternServer提供微博分类服务。该类得使用方法如下:

usage: ClassifyPatternServer [options]

options: -h -> this help

-p auxiliary server port

-t model type //0 multiclass ,2 three level tb, 3 weibo

-d dict file

即该类支持host分类、淘宝分类、微博分类

通过http协议访问该服务：格式如下

<http://110.96.34.211:9008/cwb?s=keywords=+%09%09%09%09%E5%9F%BA%E6%9C%AC%E4%BF%A1%E6%81%AF%09%09%09%09%09%E6%98%B5%E7%A7%B0%EF%BC%9Akittiewong89330%09%E6%89%80%E5%9C%A8%E5%9C%B0%EF%BC%9A%E6%B1%9F%E8%8B%8F+%E5%8D%97%E4%BA%AC%09%E6%80%A7%E5%88%AB%EF%BC%9A%E5%A5%B3%09%E7%94%9F%E6%97%A5%EF%BC%9A1989%E5%B9%B43%E6%9C%8830%E6%97%A5%09%09%E7%AE%80%E4%BB%8B%EF%BC%9A%09%E5%96%B5%E5%96%B5%09%09%09%E6%B3%A8%E5%86%8C%E6%97%B6%E9%97%B4%EF%BC%9A%09%092009-09-23%09%09%09%09%09%09%09%09%09%09%09%09%09%09%E6%95%99%E8%82%B2%E4%BF%A1%E6%81%AF%09%09%09%09%09%E5%A4%A7%E5%AD%A6%EF%BC%9A%09%09%E5%8D%97%E4%BA%AC%E9%82%AE%E7%94%B5%E5%A4%A7%E5%AD%A6+%282007%E5%B9%B4%29%09%09%09%09%09%09%09%09%09%09%09%09%09%E6%A0%87%E7%AD%BE%E4%BF%A1%E6%81%AF%09%09%09%09%09%E6%A0%87%E7%AD%BE%EF%BC%9A%09%09%09%09%09%E4%B9%90%E8%BF%B7%09%09%09%09%E8%90%8C%E7%89%A9%E6%8E%A7%09%09%09%09%E5%BF%A0%E5%AE%9EPerfume%E7%B2%89%09%09%09%09%09%09%09%09%09&id=1252451131&title>=

s参数是输入的微博账号的各种信息：

例如：

简介： 解析双鱼座的心理学！~~~~~~~~~ 喜欢就 【✚加关注】｡◕‿◕｡ @双鱼座-心理学 标签: 双鱼座 双鱼座心理学 双鱼座运势 双鱼座专区 星座 心理学 十二星座 keyword: 双鱼座-心理学 双鱼座-心理学，双鱼座-心理学 双鱼座-心理学，解析双鱼座的心理学！~~~~~~~~~ 喜欢就 【✚加关注】｡◕‿◕｡ @双鱼座-心理学。双鱼座-心理学

要将这些信息使用urlcode编码，作为s参数的值。