# TopicSketch使用文档

## 1 模块介绍

TopicSketch是设计用来实时检测大量微博流上的突发话题。因为现有的topic modelling方法很难处理大规模数据流，我们提出一种基于sketch技术的topic modelling： TopicSketch。这种方法的优点在于通过降维技术它大大降低了数据的复杂性，因此它能实时地处理大规模的数据。仅用一台机器，TopicSketch每天能够处理上千万条微博。TopicSketch的输入是微博流，输出是检测到的话题的关键词已经该话题的简单摘要。

## 2 模块功能

Figure 1 显示了TopicSketch 模块的主要流程，分为以下五个部分: (I) 微博数据流(tweet stream), (II) 数据摘要(sketch), (III) 数据监控(monitor), (IV) 话题检测(estimator) 以及 (V) 话题摘要(reporter). 以下是整个流程的分步解析:

1. 每来一条新的微博，数据摘要会高效地更新。
2. 相应的数据监控模块会监控更新的数据摘要。
3. 一旦数据监控模块检测出异常信号，检测突发话题的模块会被通知。
4. 一旦被通知话题检测模块会从数据摘要中分析出突发话题。
5. 最后话题摘要模块会根据话题关键词获取相关微博来对该话题进行简单的摘要。

## 3 模块部署

### 3.1 准备

1. 首先应注意: 要保证输入的微博数据流按时间顺序 !!! 并实现文件topic_sketch/stream.py中ItemStream数据流接口。experiment/tweet_stream.py中TweetStreamFromDB是一个例子。

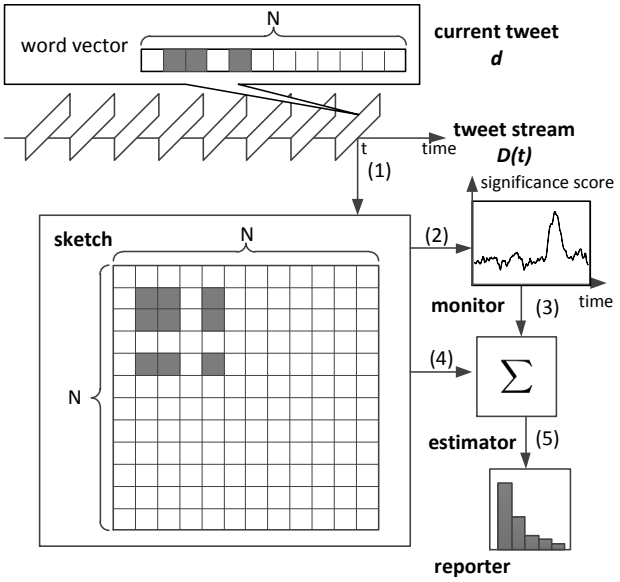


Figure 1: TopicSketch framework overview

2. 为了增加数据处理能力，有些组件是用C以及C++写的，因此需要安装 gcc 和 cython (pip install cython)。不要忘了copy这些库文件到主文件夹下。

3. 编译文件夹c下面的hash.h：

   (a) For mac or linux:

   ```
   gcc -shared -I/usr/include/python2.7/ -lpython2.7 -o mlh.so  hash.c
   or gcc -fPIC -shared -o mlh.so  hash.c
   ```

   (b) For windows:

   ```
   gcc -c  hash.c -IC:/Python27/include -o mlh.o
   gcc -shared mlh.o -LC:/Python27/libs -lpython27 -o mlh.dll
   ```

4. 编译文件夹cython下面的fast_signi和fast_smoother

   (a) For mac or linux:

   ```
   python setup.py build_ext --inplace
   ```

   (b) For windows:

   ```
   python setup.py build_ext --inplace --compiler=mingw32
   ```

5. stop words 文件twitter-stopwords.txt : 提供预设的stop words。

## 3.2 参数设置

请在twitter.cnf中设置参数。像很多系统一样，大多数参数是不需要特别设置的，用默认的就好。 以下介绍几个比较重要的。

1. 数据源, 如果使用experiment/tweet_stream.py 中TweetStreamFromDB需要设置如下参数

```
[database]
host = ...
user = ...
db = ...
charset = utf8
```

2. detection 参数
   (a) window_size = 24*60 （时间窗口，默认一天24小时）
   (b) cycle = 7*24*60 (衰减周期，默认一周7天)
   (c) average = 10.0/(24*60) （请使用默认值）
   (d) threshold_for_cleaning = 0.1 （请使用默认值）
   (e) capacity_for_cleaning = 10000000 （请使用默认值）
   (f) signi_threshold = 15.0 （检测阀值）

除此之外，检测事件的输出在experiment下面的event_output.py里面设置，当前代码将输出写入redis数据库。这里是例子

```
_KEY_EVENT_CHANNEL =  ...
_KEY_EVENT_PREFIX = _KEY_EVENT_CHANNEL + ':'
_KEY_EVENT_LATEST_EVENT_ID = _KEY_EVENT_PREFIX + 'nextId'
_KEY_EVENT_LATEST_TOPIC_ID = _KEY_EVENT_PREFIX + 'next_topic_Id'
_KEY_EVENT_IDS = _KEY_EVENT_PREFIX + 'ids'
_KEY_EVENT_TIMESTAMPS = _KEY_EVENT_PREFIX + 'timestamps'
_KEY_EVENT_KEYWORDS = _KEY_EVENT_PREFIX + 'keywords'

_HOST = X.X.X.X
```

## 3.3 运行

请在有足够内存（大于等于16G）的机器上运行TopicSketch。在做好准备工作和设置参数后，run python main.py。

## 3.4 功能扩展

除此之外，TopicSketch还实现了如下扩展功能。

1. 关键词追踪。
2. 个性化监听以及预警。