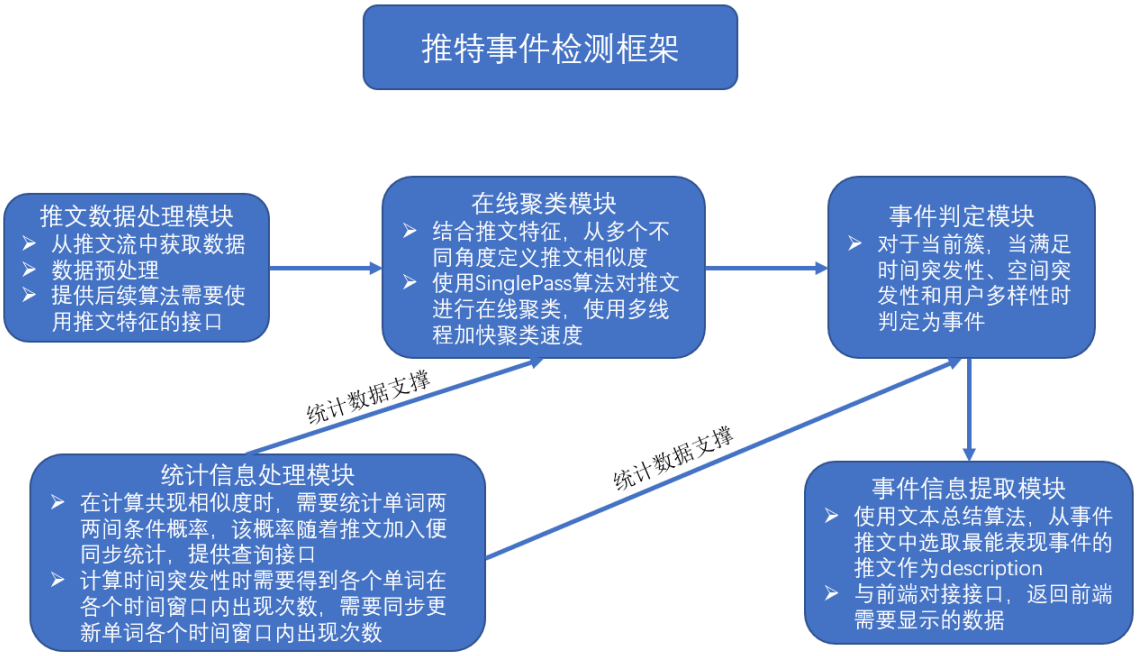
事件检测

分为推文数据提取、在线聚类、事件判定和事件信息提取四个模块，事件提取出来的信息需要与前端对接。除此之外，还需要维护一些全局变量，例如两两单词间条件概率，每个单词的word2vec向量，每个单词在不同时间段出现次数，进而可以看出这些单词在哪些时间段具有突发性。



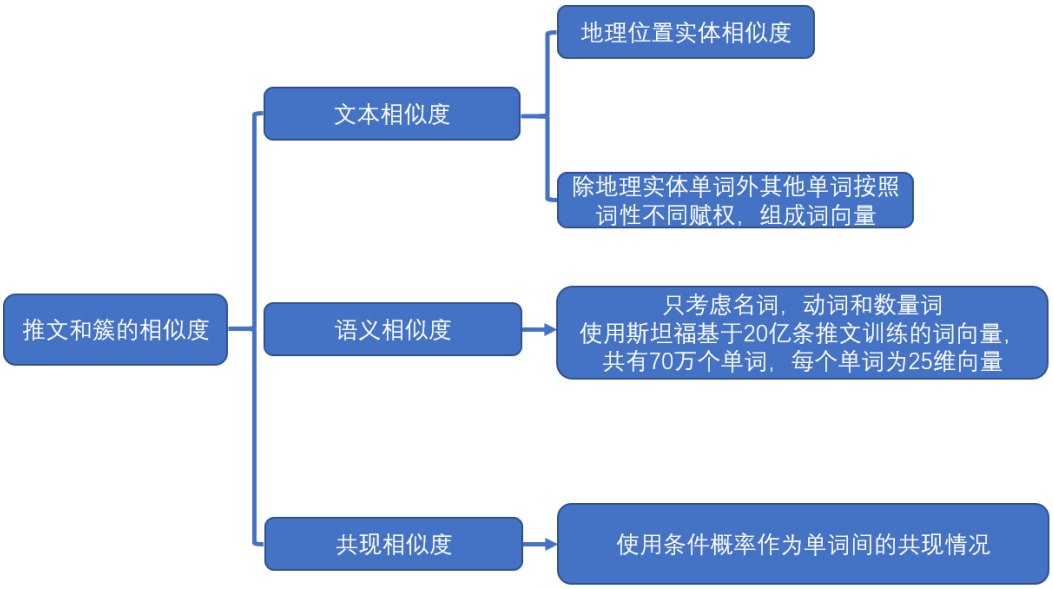
一、推文数据提取模块

- `text`：推文文本
- `words`：推文文本进行分词、词性检测和单词是否是[人名、地名、机构名]
- `event`：根据模型判断该条推文是否是事件推文，只处理event=true的推文
- `coordinates`：推文所属经纬度坐标
- `lang`：推文语言，只处理英文推文，即需要满足lang=en
- `created_at_ts`：推文发送的时间戳(单位:s)
- `user`：发送者名称(screen_name)

二、在线聚类模块

- 使用SinglePass在线聚类方法进行聚类，主要是考虑到实时性

- 使用聚类方法就需要定义推文间相似度，短文本相似度不好处理，需要从不同的角度定义相似度



- 簇模块需要有以下字段：
  - `textList`：推文文本组成的列表
  - `timeList`：推文时间组成的列表
  - `userList`：推文发送者组成的列表
  - `locDict`：地理实体单词及其出现次数组成的字典
  - `ner`：三种实体组成的字典
  - `wordsWeight`：每个单词及其权重
  - `semanticVector`：由每个单词的word2vec向量得到簇的语义向量
- 斯坦福的推文词向量网址为斯坦福推文词向量下载

三、事件判定模块

- 事件定义：特定时间和空间内发生的事，还需要有大量人参与讨论
- 时间突发性：根据簇内推文的单词在该时间段是否具有突发性来判断(z-分数判断是否具有突发性)
- 空间突发性：根据簇内地理实体及其出现次数，利用信息熵来判断事件是否具有空间突发性
- 用户多样性：根据推文发送者数量情况判断，如果一个簇内推文是由少数几个人发布的，很可能是水军或者广告
- 经过上述三个过滤器后簇被认作是事件

四、事件信息提取模块

经过事件判定的簇

- `description`：选择一条具有代表性的推文
- `location`：地理位置，ex: [{latitude: 48, longitude: -2.92, location: "brittany"}]
- `time`：事件时间
- `ner`：实体，ex：{org: [], per: [["a", 2], ["b", 1]], loc: []}
- `picture_path`：返回`description`对应推文的图片，如果没有则为None