Online microblog stack

Repost crawler/event monitor

Note:

1. 当前index有一个问题，对于同一个mid的转发序列，无法进行重复检测，也不能够实现删除。所以就需要保证，如果一条微博的转发时间序列在被分段近似之前，要保证时间序列的值尽可能接近实际值。
2. event monitor或者repost cralwer可能会重复爬取微博，如何保证一条微博的转发次数不会被重复计算？ 所有流向

保证：

1. pubcrawler如果把数据丢给了monitor，那就不应该把对应的tweet放入到retweet中去？

采用类似于first-class citizen的方式，优先索引那些热门的微博。具体做法如下：

1. 监控public timeline，讲监控到的数据丢入刀tweet这个topic中去，同时通过判断模块，决定是否监控一条微博，对于监控的微博，将其mid发送到monitor这个topic中去
2. repost cralwer从monitor这个topic中获取新的mid，加入到监控队列中去。对于每个已经在爬取的微博，同样判断他们是否还有爬取的必要。对于每个当前在监控的mid，每批次爬去完成之后，将他们的爬取状态写回到kafka的monitor这个topic中去。
3. segmentation这个类，从monitor这个topic中读取mid，根据当前mid的状态，判断是否开始进行segmentation。
4. 对于那些没有被索引的普通微博，只能定期开启mr/spark任务，筛选出那些尚未被索引的微博，开始进行索引。