事件抽取项目第二版技术方案

一、 项目背景

目前在微博头条的相关业务中缺少事件抽取这一基础模块，开发事件抽取模块可以实现同一事件文章的聚合。

二、 需求分析

1、 问题与现状

- 现状：

  目前依靠文章排查模块，减少相似文章重复展现，并实现热点事件发现。

- 问题：
  - ✓ 文章主题相同，但标题及内容有差异，会分在不同的组中，会出现较多相似内容同时展现问题。
  - ✓ 同一事件的文章会分在不同的组中，不利于热点发现与召回。

2、 目前的需求

- 减少重复展现及优化展现内容
  - ✓ 同一事件的文章聚合，减少同一事件内容同时展现
  - ✓ 制定对应时间轴规则，展现事件开始、发酵、结果、深度剖析等不同阶段内容。
- 热点聚合和发现

✓ 聚合热点事件相关文章，提高热点召回的覆盖率与准确率。

- ✓ 扑捉实时热点，提高热点文章的点击率。

三、 项目评价标准

- ➢ 初版评价标准

  随机抽取一批热点事件，分别对比排重模块和事件抽取模块的热点事件内容聚合效果，主要是内容聚合的准确率和覆盖率。

- ➢ 第二版版评价标准

  随机抽取一批热点事件，观察事件抽取模块的热点事件内容聚合效果，主要是内容聚合的准确率和覆盖率。

四、 初版技术方案

- ➢ 初版功能说明
  - ✧ 输入一篇文章，输出该文章的事件 id
  - ✧ 以热点聚合与发现为目标，非热点事件不给于过多关注。
- ➢ 初版具体过程与方案

方案一目的在快速输出，作为项目的 baseline,期望对热点聚合的准确率和覆盖率有一定提升；方案二为项目优化做数据积累。两个方案应该同时启动。

1. 方案一：

✧ 方案内容：

- ✓ 技术负责模块开发：
  - ◆ 根据 gsize，top 等特征，筛选出热点事件

- ◆ 以热点事件为聚类中心完成聚类

✓ 产品和运营同学负责效果评估

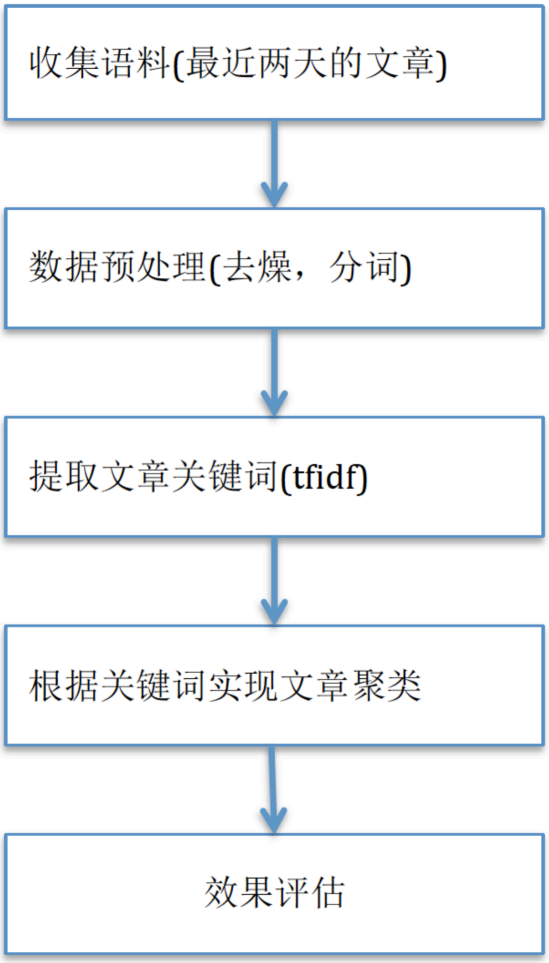
✧ 架构设计

注：文章表示成向量后，提特征的过程可以有多种算法：

1、 根据 tfidf 提几个关键词

2、 autoencode

3、 doc2vec



2. 方案二：

- ✧ 方案内容：
- ✓ 数据标注：
  - ◆ 技术提供简洁明确的标注样例
  - ◆ 运营同学在日常评估过程中帮助积累数据
- ✓ 模块开发、效果评估、架构设计暂时不关注
- ➢ 初版聚类技术方案的问题

1、 要解决的问题：

- ✓ 针对热点文章的事件聚合：
  - ◆ 第一个要解决的问题：针对挑选的热点文章，找到其中属于同一事件的文章
  - ◆ 第二个要解决的问题：找到与挑选的热点文章属于同一事件的文章

2、 文章表征成向量存在问题

- ✓ tfidf 提关键词有很大问题，只能提取概念比较大的单词，关键的单词提不到
- ✓ 分词后最好只保留名词

3、 特征选择问题

- ✓ 文本特征
- ✓ topic 特征
- ✓ 大数据 12,13 标签(ttag 粒度太大)

4、 聚类算法问题解决难度大

- ✓ 针对热点文章的事件聚合，聚类算法在粒度要求上达不到标准，主要是相似度计算方法和特征表示上有很大问题。

➢ 初版聚类技术方案的问题

1、 要解决的问题：

- ✓ 针对热点文章的事件聚合：
  - ◆ 第一个要解决的问题：针对挑选的热点文章，找到其中属于同一事件的文章
  - ◆ 第二个要解决的问题：找到与挑选的热点文章属于同一事件的文章
- ✓ 刚开始只解决第一个要解决的问题

2、 方案

- ✓ 提取特征：lstm 提取关键词(1~2 天),作为文本语义特征；文章归属前 10topic 的权重；文章归属大数据 12,13 标签的权重。
- ✓ 热点文章计算相似度，设定相似度阈值，相当于简单的聚类。

3、 方案实践进展及下一步计划(2017/5/17)

- ✓ 目前利用 lstm 提取关键词的工作进展
  - ◆ 利用运营的标注取数据，得到 600+篇文章和对应的事件，分词后存储为训练数据
  - ◆ 针对训练数据，训练 word2vec 模型，得到单词对应的词向量

- ◆ lstm 模型快速应用有困难，几天内很难出结果。

✓ 下一步计划：

- ◆ 在第一版的基础上进行改动，文章标题分词后得到对应词向量作为特征，再加上 topic 和大数据 tag 特征，在热点事件中找相似事件。认定两个热点文章时一个事件的相似度阈值很重要。
- ◆ 同一事件的 eventid 是热度大的文章的 gid;没有相似文章则 eventid 为 gid
- ◆ redis 登录命令:

  redis-cli -h rs7017.mars.grid.sina.com.cn -p 7017

五、 事件抽取第二版方案

1、 事件抽取三步走：

- ✧ 事件探测，就是怎么知道这是一个(热点)事件，事件的定义是有开始、发酵、深度的新闻。
- ✧ 文章、事件的特征向量:doc2vec/topic/文章标签
- ✧ 文章间距离计算，也就怎样判定两篇文章属于一个事件。

2、 事件探测：

➢ 起步方式：

- ✧ 运营提供一批账号，比如人民网等，这些网站的特点是重大事件报道快。
- ✧ 提取这批账号的微博，提取事件，如果有多个账号发布同一事

件的文章，那么可以认定这个事件为热点事件。

3、 事件抽取

- 文章、事件的特征向量
  - ✧ 语义特征(文章表征成向量):

    标题 word2vec，各单词求和取平均，表示文章向量
  - ✧ 文章 topic 特征

    前 10topic 权重
  - ✧ 文章标签特征

    标签权重
- 文章间距离计算
  - ✧ 欧几里得距离
  - ✧ 相似度阈值

六、 事件抽取第二版方案改进

1、 目前的问题：

(1)、标题单词 worde2vec 相似度计算，不能用目前的方法，不能自己凭空堆砌公式，需要查查资料或者和人讨论。下一步的想法：

- ✓ 词的语义距离可以用 word2vec 给出的词向量之间的 Euclidean 距离/权重用 tfidf。
- ✓ 文章距离也可以用 emd 距离

- ✓ 先用标题单词重复度过滤,要有两个不同的单词重复,标题重复应该是一个单词

(2)、topic 相似度计算时两个分布的相似度计算，用 emd 距离

(3)、标签相似度计算也是两个分布的相似度计算，用 emd 距离

(4)、事件聚合后，还要判断聚合得到的是不是一个事件，依据特征：聚合结果中的 gid 组个数，gid 组内文章。事件聚合后，还要从组内找出一篇最好的文章，可以根据点击率等特征。

(5)、事件抽取完成后，存储第一次聚合结果，再有新文章进来，计算新文章同各个事件的相似度，决定新文章是不是归属一个事件。

(6)、组越来越大的时候，要有函数可以执行 delete 操作。

(7)、降低一下 vip 的门槛　vip=1 就能够进入你的事件检测

2、最新讨论结果:

A、一个文章有事件 id 后不再改变

B、事件有标准和特征,eg.标准：组个数>2,组内文章数>20;特征:指定重复的两个词，指定重复的 topic

C、第一次启动，计算 event，存到 redis。下次计算如果一个新文章符合多个事件的要求，随机选择一个事件，但原来的事件不改变。事件 id 可以随机使用事件文章的 gid.