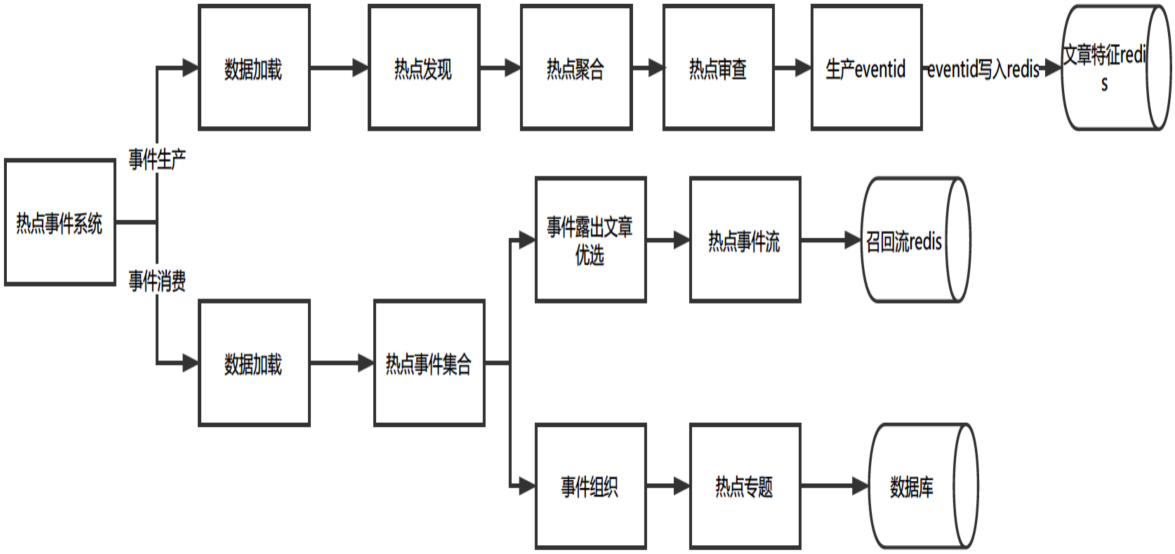
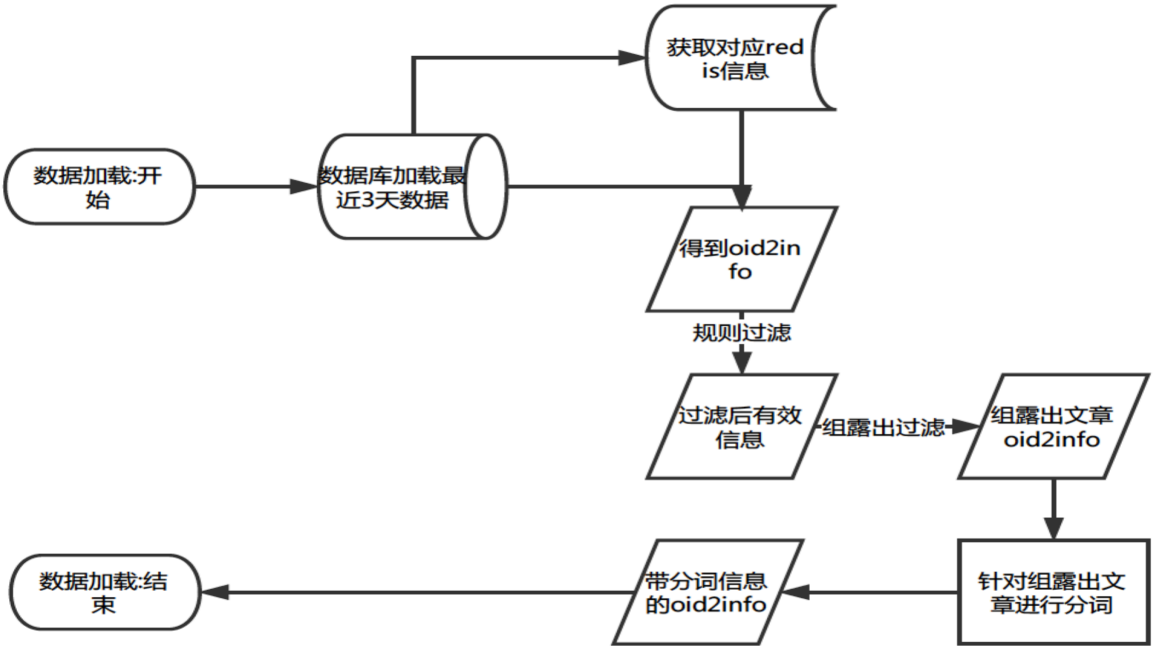
热点项目任务分解

一、工作流



二、事件生产基础架构：

1、信息加载

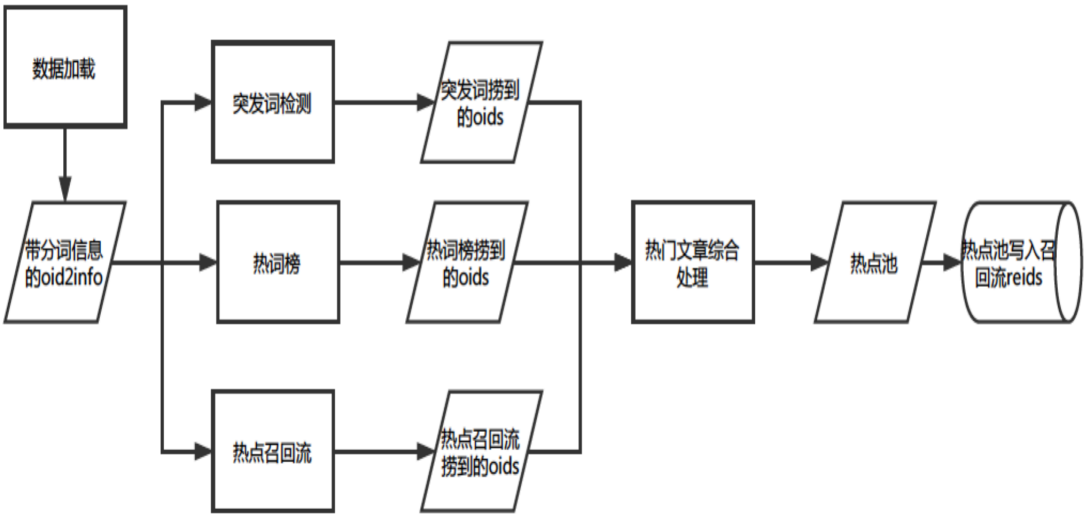


2、热点发现

该模块包括的子模块有：突发词检测模块、热词榜、热门文章召回流(媒体热门、微博热门、全局热门、各领域垂直热门)

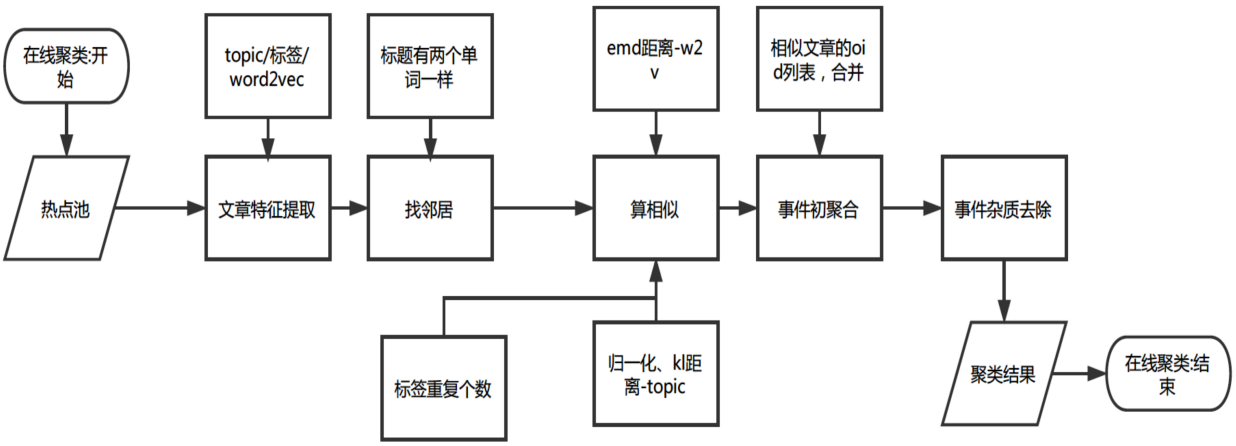
- 该模块产出结果为热点池，通过这段时间的观察，覆盖率和时效性都达到开发前的预期
- 可以优化的子模块-突发词检测模块：

  目前该子模块没有产出期望的结果，模块贡献很小。



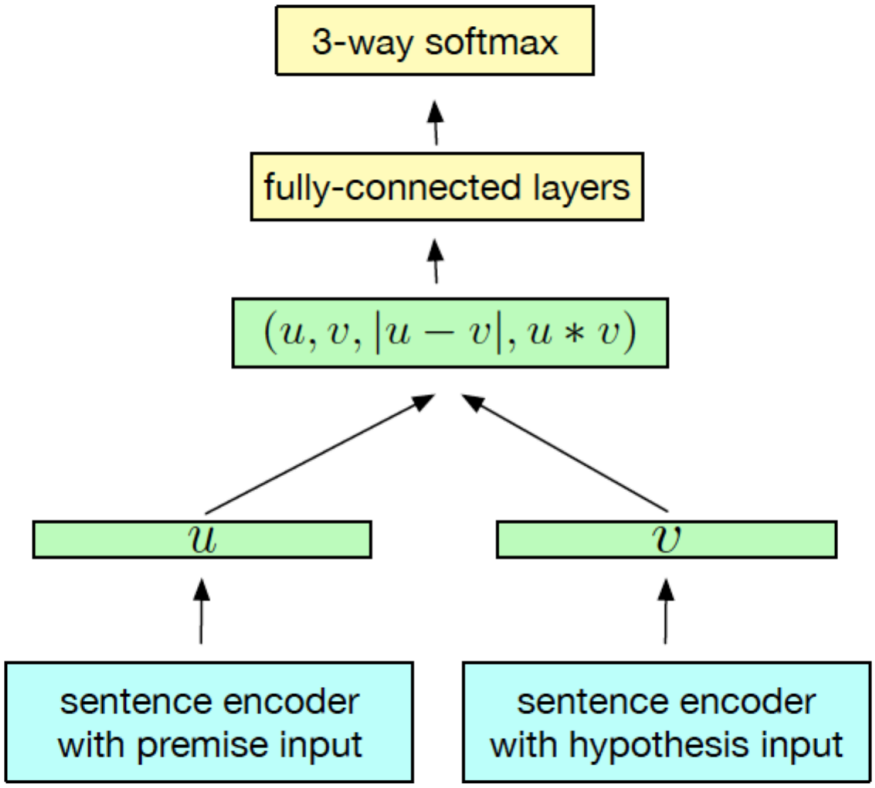
3、热点聚合

- baseline 部分利用简单的模型完成了该模块工作流开发，如下图：



- 开发神经网络语义模型

  在下图中，有监督的语义匹配模型，在训练中匹配得分就是 0 或 1，在预测中则可以用 0/1 或连续的数值。



✧ 2017/10/17~2017/10/20 数据准备(3 个工作日)

这部分工作是生成离线实验数据，数据形式为(doc1,doc2,label)。

✓ Label 的规则:

若两篇文章的 hottag 相同，那么对应 label 为 1；其他样本为 0；这里需要想想其他维度，发散思维。

✓ 训练样本选取:

正样本: label 为 1 的文档对

负样本:同一领域 label 为 0 的文档对；负样本会不会过于扩散，发散思维。

✓ 训练文件生成:

标签，topic，word2vec 特征

✧ 2017/10/22~2017/10/24 mlp 模型(2 个工作日加周日)

这部分工作是利用标签，topic，word2vec 特征完成 mlp 模型离线训练和在线预测

✧ 2017/10/25~2017/10/31 lstm 模块(5 个工作日加周末)

这部分主要工作是完成 lstm 模块与监督学习模块的嫁接，实现整体训练。

✧ 2017/11/01~2017/11/03 整个 end2end 框架(3 个工作

日)

完成整体语义匹配模型的开发工作，产出为 auc

- ✧ 2017/11/04~2017/11/10 模型调整(5 个工作日，1 个周末)

  调整语义匹配模型，产出为更好的 AUC

- ✧ 2017/11/11~2017/11/15 新版本上线(3 个工作日，1 个周末)

  将改版后的热点事件系统上线，产出有二，一为实时更新的热点召回流；二为展示页面，输入 eventid,输出该事件的文章(可以改造组露出监控页面)。

4、热点审查

- ➢ 规则检测:聚类得到的簇中文章数目，以及簇内文章 gsize 大小

5、热点标签

- ➢ 为热点文章打标签:已经有 eventid 的不改变；无 eventid 的从簇内文章 eventid 中随机选取；整个簇内文章都无 eventid，那么随机选一篇文章的 id 作为 eventid
- ➢ eventid 存 redis 和数据库

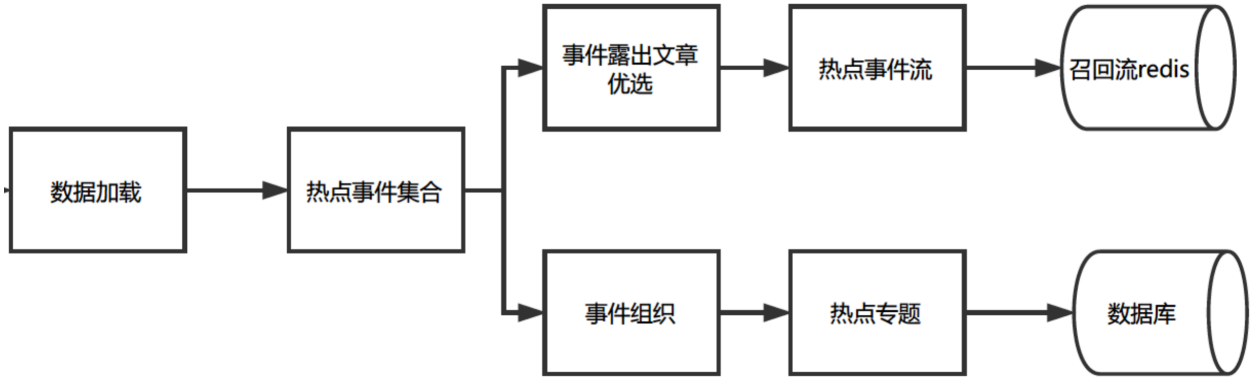
三、事件消费基础架构：

事件露出文章优选-规则选取，找热门，去除事件重复

热点事件流-产出高时效性、高点击率的热点文章

热点专题-重大热点跟踪报道

工作流如下图:



四、热点事件项目评价指标：

1、baseline 版本运营人工评审：

覆盖率：系统发现热点事件数目/人工审查热点数目

准确率:系统发现热点事件为真个数/系统发现热点事件总数

评审结果:准确率 100%，覆盖率 32%

2、有监督语义匹配模型 AUC