基于Flink的文本流事件监测

小组成员及分工

小组组成员

姓名	学号
陈炳生	21210240007
罗翔	21210240030
施渝斌	21210240035
赵一峰	21210240441

小组分工

- 陈炳生：复现算法Phase 1，集群部署
- 罗翔：　复现算法Phase 2，数据清洗，优化算法
- 施渝斌：复现算法Phase 3，可视化分析
- 赵一峰：数据接入，文档撰写

项目代码仓库：GitHub - SunflowerAries/Text-Event-Detection

系统目标及功能

项目选题为基于 Storm/Flink 的文本流事件监测，其实现主要参考论文 *Parameter Free Bursty Events Detection in Text Streams*。本系统基于Flink框架、Kaggle上新闻集完成论文中提出的算法，并对其进行改进，实现对文本流的事件监测系统。

本系统的功能主要为在文本流中实现事件监测，包括三个流程：

1. 首先通过输入文本流识别出突发特征；
2. 接着对特征进行聚类，将特征聚类成一个个事件；
3. 最后判断这些事件的热点时间。

环境搭建与数据获取

环境搭建

在本项目中，选择使用 Flink 来进行文本流事件监测，Flink 是一款分布式的计算引擎，Flink 将计算过程建模为数据流上的有状态的计算（Stateful Computations Over Streams），认为有界数据集上的批处理是无界数据流的一种特例，实现了流批一体。基于 Flink 框架，我们首先进行了环境的搭建，在老师提供的五台服务器上构建 Flink 集群，主要包括：选择主节点并对其进行相关的配置，把配置好的 Flink 文件打包后分发给其他四台机器统一配置。

数据获取

本项目为文本流事件监测系统，需要数据来构造文本流，文本流是带有日期标签的一系列文本。在 kaggle 上找到了 New York Times Articles & Comments (2020) 新闻数据集，其中包含2020全年的纽约时报16K+的文章和相关评论，其主要包含内容标题、摘要、关键词、时间等。

A headline	A abstract	A keywords	# word_count	🗓 pub_date	# n_comments	A uniqueID
Protect Veterans From Fraud	Congress could do much more to protect Americans who have served their country from predatory for-pr...	['Veterans', 'For-Profit Schools', 'Financial Aid (Education)', 'Frauds and Swindling', 'Colleges an...	680	2020-01-01 00:18:54+00:00	186	nyt://article/69a7090b-9f36-569e-b5ab-b0ba5bb3ccbd
'It's Green and Slimy'	Christina Iverson and Jeff Chen ring in the New Year.	['Crossword Puzzles']	931	2020-01-01 03:00:10+00:00	257	nyt://article/9edddb54-0aa3-5835-a833-d311a76f1e7c
Meteor Showers in 2020 That Will Light Up Night Skies	All year long, Earth passes through streams of cosmic debris. Here's a list of major meteor showers ...	['Meteors and Meteorites', 'Space and Astronomy', 'Earth', 'Solar System']	1057	2020-01-01 05:00:08+00:00	6	nyt://article/04bc90f0-b20b-511c-b5bb-3ce13194163f
Sync your calendar with the solar system	Never miss an eclipse, a meteor shower, a rocket launch or any other astronomical and space event th...	['Space and Astronomy', 'Moon', 'Eclipses', 'Seasons and Months', 'Solar System', 'Meteors and Meteo...	0	2020-01-01 05:00:12+00:00	2	nyt://interactive/5b58d876-9351-50af-9b41-a312490d2728
Rocket Launches, Trips to Mars and More 2020 Space and Astronomy Events	A year full of highs and lows in space just ended, and the 12 months to come will be full of new hi...	['Space and Astronomy', 'Private Spaceflight', 'Rocket Science and Propulsion', 'National	1156	2020-01-01 05:02:38+00:00	25	nyt://article/bd8647b3-8ec6-50aa-95cf-2b81ed12d2dd

图1. 数据中包含主要内容

数据清洗及预处理

在数据集中，包含本项目中用不到的无关项，首先对数据进行进行清洗，仅保留数据中的时间及正文部分，删除其他如标题、作者、分类等项。同时，基于项目的需要，对新闻正文进行一些简单的预处理工作，包括：对新闻正文进行分词、大写字符替换为小写字符、复数转为单数、去掉停用词等。至此，本系统需要的输入数据已准备完成。

算法实现

算法实现是本项目的核心工作，基于 *Parameter Free Bursty Events Detection in Text Streams* 中所提出的算法，进行算法的改进与实现。

代码结构

```
1  Text-Event-Detection
2  |    BurstyAggregate.java
```

```
 3 |      BurstyProcess.java
 4 |      Document2Feature.java
 5 |      Feature2Event.java
 6 |
 7 |----lib
 8 |         Binomial.java
 9 |         BurstyProb.java
10 |         UnionFind.java
11 |
12 |----module
13 |         Document.java
14 |         Event.java
15 |         Feature.java
16 |         FeatureOccurrence.java
17 |         FeatureWithTimeStamp.java
18 |         HotPeriod.java
19 |         PerDayInfo.java
```
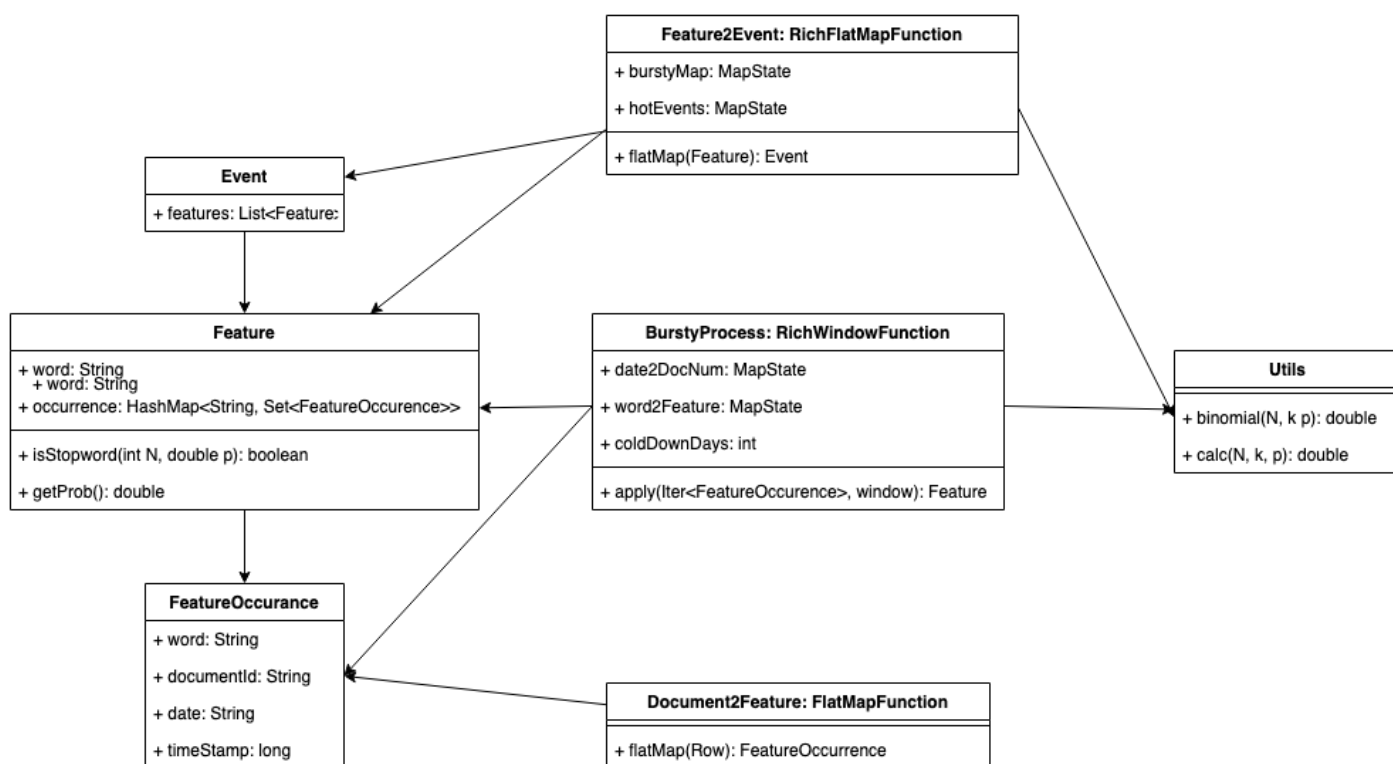
UML类图

绘制UML类图，展示一个Java类的属性和功能。

图2. 算法UML类图

算法中存在问题及改进

1. 批处理转化为流处理

原文中的算法均基于批处理，每个步骤都需要收集全体的数据才能进行下一步，这样不具有检出突发事件任务的实时性；

在我们的实现过程中，基于Flink流处理框架对算法进行部分改进，将批处理转化为流处理。

2. 冷启动

在流处理环境下，原文中算法在实现过程中存在冷启动问题：在数据较少时，将所有feature判定为bursty，与真实情况存在较大偏差，造成算法失效；

在我们的实现过程中，设置cooldowndays阈值，在第一步中收集一部分初始数据后再流入下一步。

3. 算法复杂度

在算法第2步"From Bursty Features To Bursty Events"过程中，在计算时需要搜索所有 E_k 算法复杂度极高；

在我们的实现过程中，改用贪心算法来实现相关过程：

a. 初始化 $E = \emptyset$。

b. 取当前未被聚类的feature中出现次数最多的记为 f_{max} ，对E中所有事件 E_k ，求 $E_k + \{f_{max}\}$ 中最小的 $cost_1$ ，再当前未被聚类的feature中与 f_{max} 不同的所有 f ，取所有 $\{f_{max}, f\}$ 中的最小 $cost_2$ ，比较 $cost_1$ 、 $cost_2$ ，若 $cost_1$ 更小，且其比原 E_k 的cost更小，则将 f_{max} 加入 E_k ，否则将其丢弃，若 $cost_2$ 更小，则将 $\{f, f_{max}\}$ 加入E。

c. 重复步骤b，直到所有feature均被聚类。

系统部署及结果展示

部署环境

操作系统：

    CentOS

软件环境：

    Java8、Flink-1.14.4

系统部署及运行

系统整体Pipeline如下图所示，两种记号分别代表数据流和处理过程。

图3. 算法部署Pipeline

系统部署在 Flink 集群中，基于 Flink 的可视化管理界面，可以对系统运行状态、运行结果的进行可视化的展示，部分系统运行时可视化结果如下图。
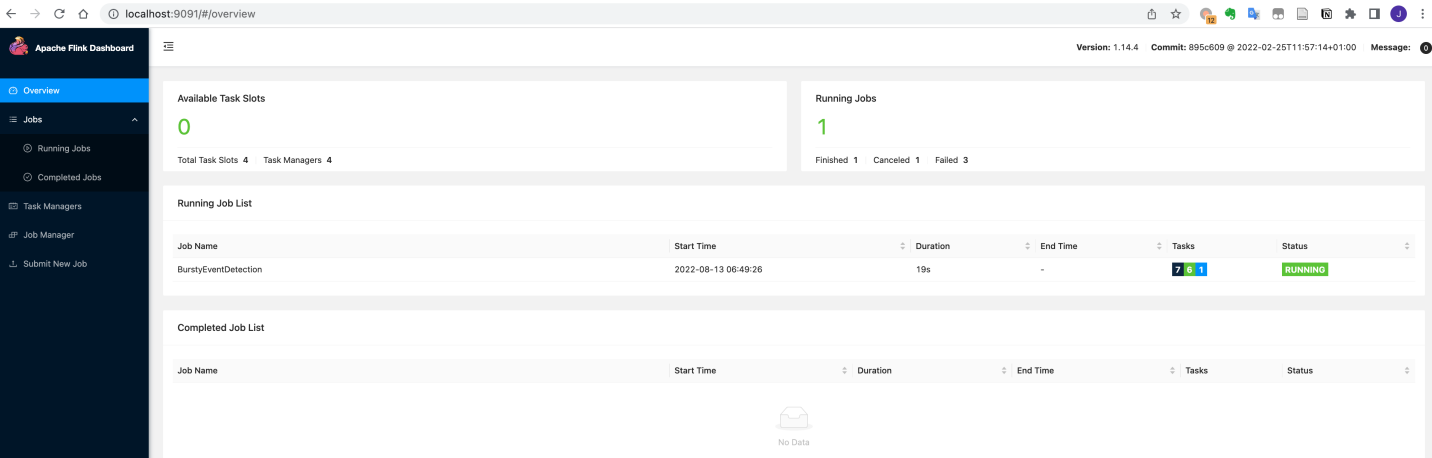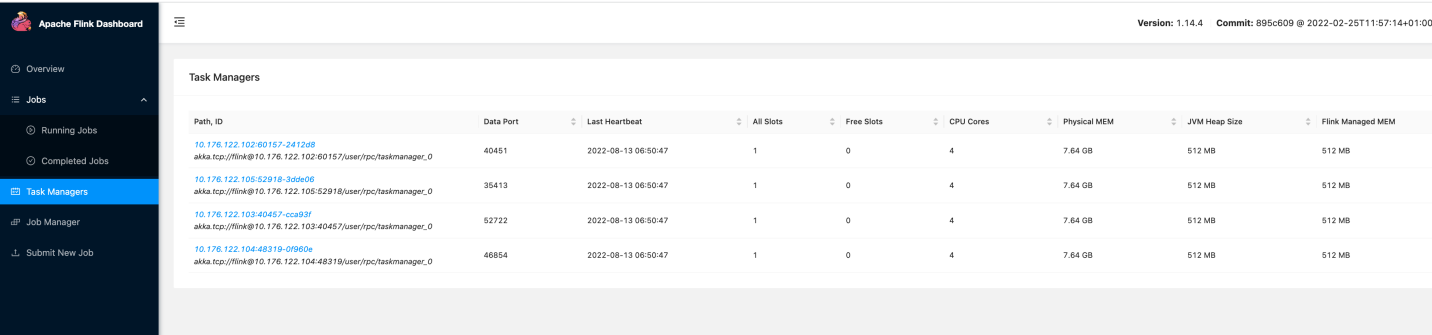
图4. Flink 可视化管理主界面
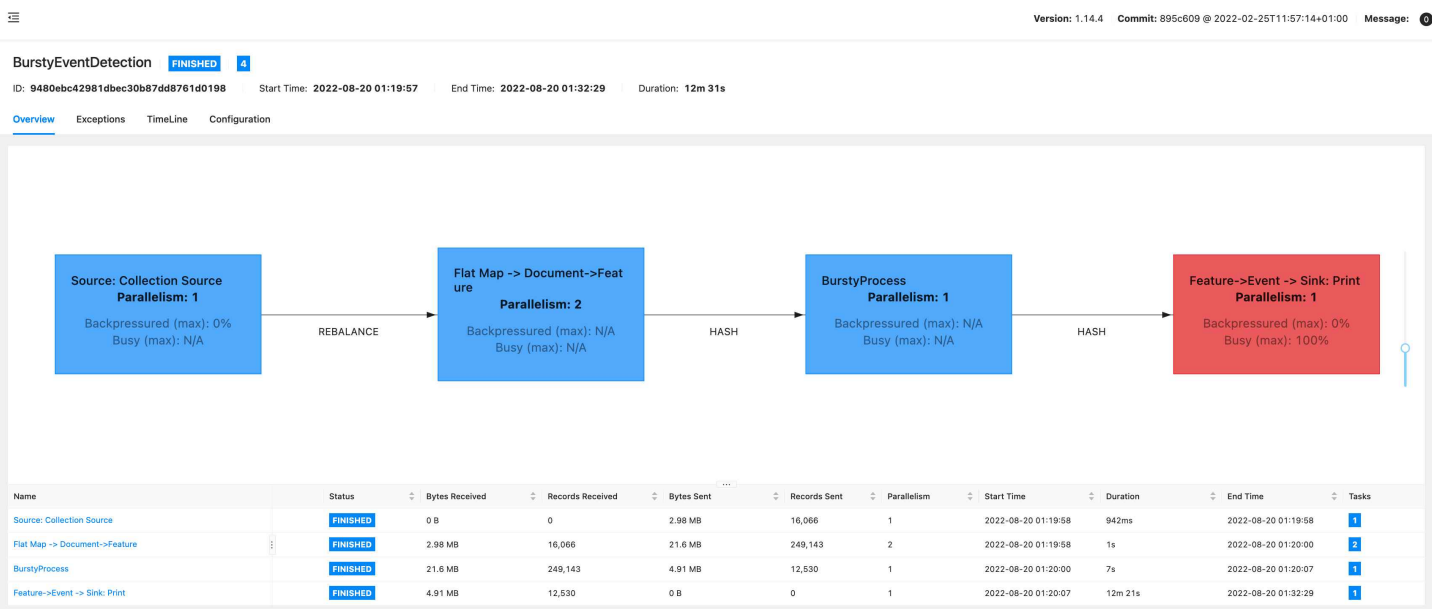

图5. 对子节点进行任务管理

图6. 系统实时运行状态

```
 1    HotPeriod{features=[brooklyn, bill], date=2020-10-07}
 2    HotPeriod{features=[pence, vice, kamala], date=2020-10-07}
 3    HotPeriod{features=[pence, vice, kamala, president], date=2020-10-07}
 4    HotPeriod{features=[pence, vice, kamala, mike, president], date=2020-10-07}
 5    HotPeriod{features=[fashion, france], date=2020-10-07}
 6    HotPeriod{features=[stimulu, economy, aid, federal, affordable, economic, act, security, relief, department], date=2020-10-07}
 7    HotPeriod{features=[television, medium], date=2020-10-07}
 8    HotPeriod{features=[pence, vice, kamala, political, joseph, biden, president], date=2020-10-07}
 9    HotPeriod{features=[college, university], date=2020-10-07}
10    HotPeriod{features=[presidency, pence, vice, kamala, political, joseph, biden, president], date=2020-10-07}
11    HotPeriod{features=[presidency, pence, vice, kamala, political, debate, joseph, biden, president], date=2020-10-07}
12    HotPeriod{features=[presidency, pence, vice, harri, kamala, political, debate, joseph, biden, president], date=2020-10-07}
13    HotPeriod{features=[presidency, pence, vice, harri, kamala, political, debate, joseph, mike, biden, president], date=2020-10-07}
14    HotPeriod{features=[new, york], date=2020-10-07}
15    HotPeriod{features=[brooklyn, bill, cuomo, queen, blasio, andrew, nyc], date=2020-10-07}
16    HotPeriod{features=[aid, affordable, economic, act, security, relief, department], date=2020-10-07}
17    HotPeriod{features=[cuomo, queen, blasio, andrew, nyc], date=2020-10-07}
18    HotPeriod{features=[cuomo, queen, blasio, andrew, reopening, nyc], date=2020-10-07}
19    HotPeriod{features=[cuomo, queen, blasio, andrew, reopening, nyc], date=2020-10-08}
20    HotPeriod{features=[television, medium, content], date=2020-10-07}
```
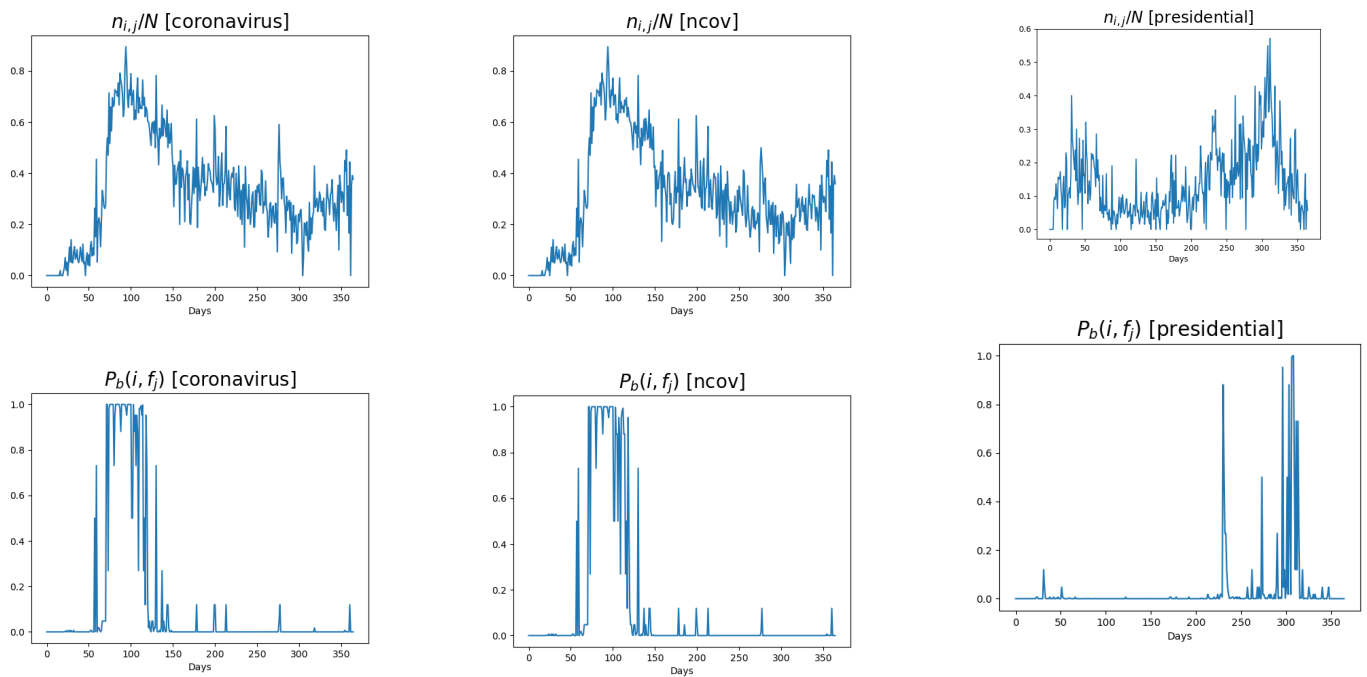
图7. 系统输出结果

系统结果展示

Bursty Features

图8. Bursty Features阶段输出结果展示

上图展示了系统中Bursty Feautres阶段的算法结果，包括[coronavirus, ncov, presidential]的结果。

Bursty Events

```
800    2020-01-20,[cipollone, pat],0.004583
801    2020-01-20,[associates, ties, russian, interference],0.002473
802    2020-01-20,[eco, tourism],0.002473
803    2020-01-20,[toxic, substances, hazardous],0.006693
804    2020-01-20,[markle, queen, britain, royal, duke, meghan, sussex, harry],0.000618
805    2020-01-21,[thunberg, greta],0.000911
806    2020-01-21,[renewal, planning],0.000623
807    2020-01-22,[official, misconduct],0.000229
808    2020-01-22,[wuhan, coronavirus, epidemics],0.000911
809    2020-01-22,[retail, shopping],0.000911
810    2020-01-22,[hackers, cyberattacks],0.000911
811    2020-01-22,[wuhan, ncov, coronavirus, epidemics],0.000911
812    2020-01-22,[shooting, saeed, guantanamo, air, computer, bay, cuba, william, works, salman, alshamrani, barr, naval, privacy, apple, pensacola,
813    2020-01-22,[wuhan, ncov, coronavirus, epidemics, viruses],0.000911
814    2020-01-22,[ncov, coronavirus, epidemics, viruses],0.000911
815    2020-01-22,[bezos, jeffrey],0.000623
816    2020-01-23,[eli, manning],0.002473
817    2020-01-23,[santos, bic],0.002473
818    2020-01-23,[post, text],0.002473
819    2020-01-23,[santos, bic, banco],0.002473
820    2020-01-23,[severe, sars],0.002473
821    2020-01-23,[santos, bic, banco, dos],0.002473
822    2020-01-23,[severe, sars, syndrome],0.002473
823    2020-01-23,[decisions, verdicts],0.002473
824    2020-01-23,[hall, derek],0.002473
825    2020-01-23,[santos, bic, banco, dos, angola],0.002473
826    2020-01-23,[santos, bic, banco, dos, angola, sonangol],0.002473
827    2020-01-23,[santos, bic, banco, dos, angola, sonangol, eurobic],0.002473
828    2020-01-23,[wuhan, severe, sars, syndrome, ncov, coronavirus, epidemics, respiratory, viruses],0.002219
```

图9. ncov-19事件输出结果展示

上图为Bursty Events阶段，2020年1月对ncov-19事件的检测结果。

```
1928    2020-06-07,[united, states, reopenings, movement, politics],0.001374
1929    2020-06-07,[police, floyd, misconduct, demonstrations, brutality, shootings, riots],0.014125
1930    2020-06-07,[police, misconduct, demonstrations, brutality, shootings, riots],0.008575
1931    2020-06-07,[police, misconduct, brutality, shootings, riots],0.008952
1932    2020-06-07,[side, upper, south, east],0.006693
1933    2020-06-07,[newspapers, liberalism],0.006693
1934    2020-06-07,[tom, cotton, news, james],0.009516
1935    2020-06-07,[tom, cotton, james],0.006693
1936    2020-06-07,[police, misconduct, brutality, shootings],0.009516
1937    2020-06-07,[rap, hip, hop],0.006693
1938    2020-06-07,[police, brutality, shootings, misconduct],0.009516
1939    2020-06-07,[side, upper],0.006693
1940    2020-06-07,[hip, hop, rap],0.006693
1941    2020-06-07,[department, new, york],0.022628
1942    2020-06-07,[police, shootings, brutality, misconduct],0.009516
1943    2020-06-07,[new, york, department],0.022628
1944    2020-06-08,[warming, global],0.002473
1945    2020-06-08,[official, ethics],0.002473
1946    2020-06-08,[bolsonaro, jair],0.002473
1947    2020-06-08,[doctors, nurses],0.002473
1948    2020-06-08,[protective, clothing],0.000911
1949    2020-06-08,[computers, internet],0.004583
1950    2020-06-08,[bolsonaro, jair, brazil],0.002473
1951    2020-06-08,[quarantine, life],0.000229
1952    2020-06-08,[psychologists, psychology],0.002473
1953    2020-06-08,[lodgings, hotels],0.002473
```

图10. George Floyd事件输出结果展示

上图为Bursty Events阶段，2020年6月对George Floyd事件的检测结果。

表1. 部分事件结果展示

Bursty Events	Bursty Features
E1（ncov-2019）	wuhan, ncov, coronavirus, epidemics, viruses, ...
E2（impeachment）	ukraine, impeachment, inquiry, complaint, ...
E3（new energy）	electronics, motors, hybird, vehicles, ...
E4（shooting）	riots, shootings, floyd, blacks, ...
E5（election）	election, presidential, trump, ...

上表展示了部分突发事件的预测结果，可以看到算法对突发事件有较好的预测能力。

分布式集群配置及运行时间

Task Managers

Path, ID	Data Port	Last Heartbeat	All Slots	Free Slots	CPU Cores	Physical MEM	JVM Heap Size	Flink Managed MEM
10.176.122.102:45675-287da1 akka.tcp://flink@10.176.122.102:45675/user/rpc/taskmanager_0	47724	2022-08-20 22:29:48	1	0	4	7.64 GB	512 MB	512 MB
10.176.122.104:40123-815fee akka.tcp://flink@10.176.122.104:40123/user/rpc/taskmanager_0	33704	2022-08-20 22:29:48	1	1	4	7.64 GB	512 MB	512 MB
10.176.122.103:42755-bce958 akka.tcp://flink@10.176.122.103:42755/user/rpc/taskmanager_0	49878	2022-08-20 22:29:48	1	1	4	7.64 GB	512 MB	512 MB
10.176.122.105:55472-2c28c0 akka.tcp://flink@10.176.122.105:55472/user/rpc/taskmanager_0	40497	2022-08-20 22:29:48	1	0	4	7.64 GB	512 MB	512 MB

图11. 分布式集群配置

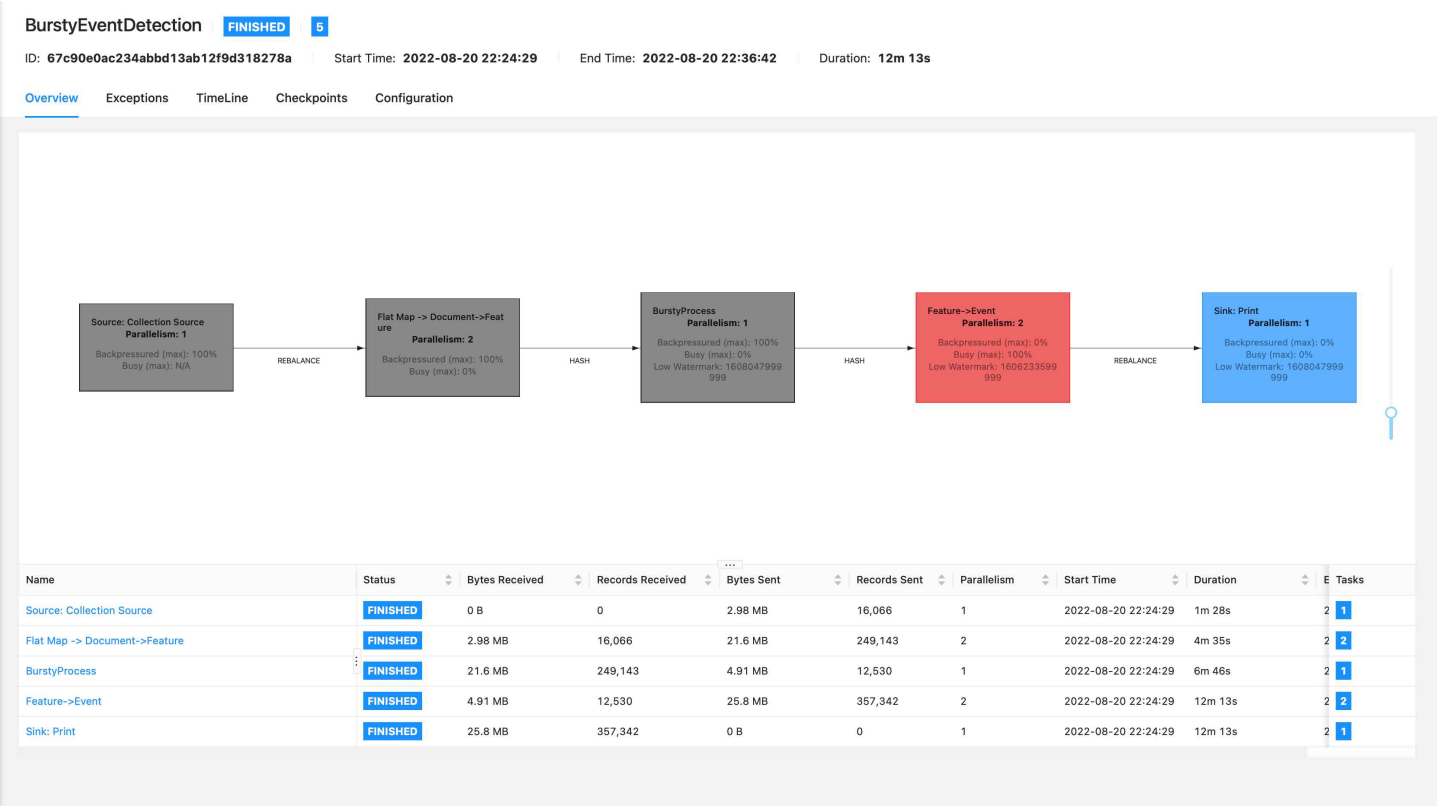分布式集群的配置信息如上图所示，每台独立机器有4核CPU及8GB的运行内存，共5台机器构成分布式集群，以10.176.122.101为主节点构建分布式集群。

图12. 分布式集群运行时间及状态

对所获取的16K+条数据进行文本流的事件监测，其运行时间为12m13s，其算法效率较批处理算法提升明显，符合算法预期。

项目排期及参考文献

项目排期

时间节点	完成工作
5.23～6.6	学习原论文工作，学习流处理框架
6.6～6.20	技术选型，讨论复现细节，寻找数据集
6.20～7.4	复现论文，接入数据集
7.4～7.18	优化实现逻辑
7.18～8.1	集群部署，优化实现逻辑
8.1～8.15	可视化结果，整理报告

参考文献

Fung, Gabriel Pui Cheong, et al. "Parameter free bursty events detection in text streams." *Proceedings of the 31st international conference on Very large data base*. 2005.