

TREC 2022

Crisis Facts and Cross-stream Temporal Summaries (CrisisFACTS)

Guidelines v1.0: 17 June 2022

Coordinators

- Cody Buntain, University of Maryland, College Park
- Richard McCreadie, University of Glasgow

Track Introduction

Tracking new developments in topics and events has been studied at TREC and other venues for several decades (e.g., from DARPA's early Topic-Detection and Tracking initiative to the more recent Temporal Summarization and Real-Time Summarization TREC tracks). These prior initiatives were generally restricted to single streams of information, but the modern information ecosystem is intrinsically multi-stream, with new developments in events surfacing across any number of news sources and social media streams. Complexity and velocity in this multi-stream space leads stakeholders to miss potentially critical information or new developments, as is particularly evident in the crisis-response domain. In that context, stakeholders need insights into new developments without being inundated with stale news stories and redundant information from disparate streams. While modern search engines are adept at providing users with search results relevant to an event, they are ill-suited to these multi-stream fact-finding and summarization information needs.

The CrisisFACTS track aims to foster research that closes these gaps, using the context of disaster response as a model. To this end, the track is making available multi-stream datasets from several disasters, covering Twitter, Reddit, Facebook, and online news sources. We supplement these datasets with a set of queries defining the information needs of disaster-response stakeholders (as extracted from incident-response reporting systems used by FEMA). Participants' systems are asked to integrate these streams into temporally ordered summaries for situational awareness and attention- allocation support for disaster response personnel.

Track organisers will then evaluate participants' systems against several event summaries (e.g., real-world incident reports, manually developed summaries from NIST assessors, and event summaries from Wikipedia). We intend to use this data to answer crucial questions about the utility of social media and news streams during crises, establish the current state of the art in multi-stream processing, and make available test collections for future crisis-summarization efforts.

Crisis Fact-Extraction and Summarization Tasks

Broadly, our users' information need is: "What critical new developments have occurred that I need to know about". Many pieces of information posted during a disaster are not essential for responders or disaster-response managers, however, so we make these needs more explicit by listing out what an responder might consider as important information:

- Damage to key infrastructure, evacuations, or emerging threats
- Changes to affected areas
- Reports and statistics regarding civilians and responders, such as casualties, those in temporary shelters, needing immunizations, or those missing
- Approaching threats to life, property, infrastructure, response operations, etc.
- Critical-resource needs, such as food, water or medicine
- Weather concerns, e.g. high wind, temperatures, humidity, floods, or watches/warnings
- Risks from hazardous materials, such as chemicals, fuels, infectious agents or radiation
- Changes to damage assessments, or restriction to the use or availability of resources
- Progress made and accomplishments by responders
- Incident command transitions (setup or transfer of command and control capabilities)

Emergency response staff typically want to receive a *summary* of this information at particular points during the emergency. For instance, such a summary might be generated at the start of a new staff shift, such that the new team members arriving can be updated on new developments. Or it might be on request from local government or media agencies reporting on the emergency.

Currently, these information needs are fulfilled via manual summarization, e.g. by filling an incident report such as an ICS 209 form.¹

2022 Tasks – Fact Extraction for Downstream Summarization

The 2022 track will have a single, core fact-extraction task, where systems consume our multi-stream datasets for a disaster-day pair and produce a minimally redundant list of atomic facts extracted from these streams, with importance scores denoting how critical the fact is for responders.

System Input

- **Events:** For CrisisFACTs 2022, we provide content for a set of eight Wildfire, Hurricane and/or Flood events. For each event we provide an event (or topic) definition:

¹ [ICS 209 Form](#)

```

{
  "eventID": "CrisisFACTS-001",
  "trecisID": "TRECIS-CTIT-H-092",
  "dataset": "2017_12_07_lilac_wildfire.2017",
  "title": "Lilac Wildfire 2017",
  "type": "Wildfire",
  "url": "https://en.wikipedia.org/wiki/Lilac_Fire",
  "description": "The Lilac Fire was a fire that burned in
  northern San Diego County, California, United States, and the
  second-costliest one of multiple wildfires that erupted in
  Southern California in December 2017."
}, {

```

- A **user profile**, containing an itemised list of queries representing the user's information needs. The union of these queries comprises an explicit representation of user interests and should be used to create a directed summary containing only matching information. For each event, we provide a general set of queries/information needs that are common across emergencies (General) and a smaller set that are specific to the event type (Wildfire, Hurricane or Flood). An example of such queries is as follows:

```

[
  {
    "queryID": "CrisisFACTS-General-q001",
    "indicativeTerms": "airport closed",
    "query": "Have airports closed",
    "trecisCategoryMapping": "Report-Factoid"
  },
  {
    "queryID": "CrisisFACTS-General-q002",
    "indicativeTerms": "rail closed",
    "query": "Have railways closed",
    "trecisCategoryMapping": "Report-Factoid"
  },
  {
    "queryID": "CrisisFACTS-General-q003",
    "indicativeTerms": "water supply",
    "query": "Have water supplies been contaminated",
    "trecisCategoryMapping": "Report-EmergingThreats"
  },
  {
    "queryID": "CrisisFACTS-Wildfire-q001",
    "indicativeTerms": "acres size",
    "query": "What area has the wildfire burned",
    "trecisCategoryMapping": "Report-Factoid"
  },
  {
    "queryID": "CrisisFACTS-Wildfire-q002",
    "indicativeTerms": "wind speed",
    "query": "Where are wind speeds expected to be high",
    "trecisCategoryMapping": "Report-Weather"
  },
  {
    "queryID": "CrisisFACTS-Wildfire-q003",
    "indicativeTerms": "helicopters",
    "query": "Are helicopters available",
    "trecisCategoryMapping": "Report-ServiceAvailable"
  }
]

```

Figure 1. Example Queries. We show three general and three type-specific queries.

- A list of **summary requests**, where each request lists the event ID, date to summarise, and the start and end timestamps bounding the requested summary. A multi-day disaster will have multiple such summary requests:

```
{
  "eventID": "CrisisFACTS-001",
  "requestID": "CrisisFACTS-001-r3",
  "dateString": "2017-12-07",
  "startUnixTimestamp": 1512604800,
  "endUnixTimestamp": 1512691199
}
{
  "eventID": "CrisisFACTS-001",
  "requestID": "CrisisFACTS-001-r4",
  "dateString": "2017-12-08",
  "startUnixTimestamp": 1512691200,
  "endUnixTimestamp": 1512777599
}
```

Figure 2. Two Summary Requests. Each request corresponds to a day in the disaster timeline

- A set of **content streams from online sources**, containing itemised text snippets (e.g., sentences) that contain content related to the event (though a specific snippet may not be a fact relevant to the disaster's user profile). Each item has an *event*, *item identifier*, *source*, *timestamp* and a *piece of text*. These items contain the content that your system should use to produce your summary, and the union of these snippets should reproduce the original content stream from Twitter, Reddit, etc.

```
{
  "event": "CrisisFACTS-001",
  "streamID": "CrisisFACTS-001-Twitter-14023-0",
  "unixTimestamp": 1512604876,
  "text": "Big increase in the wind plus drop in humidity tonight into Thursday for San Diego County #SanDiegoWX https://t.co/1pV0ZAhSJH",
  "sourceType": "Twitter"
}
{
  "event": "CrisisFACTS-001",
  "streamID": "CrisisFACTS-001-Twitter-27052-0",
  "unixTimestamp": 1512604977,
  "text": "Prayers go out to you all! From surviving 2 massive wild fires in San Diego and California in general we have all c... https://t.co/B5Y7KLY0uS",
  "sourceType": "Twitter"
}
{
  "event": "CrisisFACTS-001",
  "streamID": "CrisisFACTS-001-Twitter-43328-0",
  "unixTimestamp": 1512691164,
  "text": "If you're in the San Diego area (or north of it), you should probably turn on tweet notifs from @CALFIRESANDIEGO fo... https://t.co/hNjEuEfKaB",
  "sourceType": "Twitter"
}
```

Figure 3. Three Event Snippets for Event CrisisFACTS-001. Each snippet is extracted from the content stream associated with the listed event.

System Output

Your system should produce one summary for each request using only the content provided for that event and only between the starting and ending timestamps. This task differs from traditional summarization in that you should not simply produce a block of text of a set length. Instead, this track's "summaries" are composed of a collection of facts describing the target disaster's evolution. As such, your summaries consist of an itemised list of 'facts' that match one or more of the user information needs.

Each fact should contain the following:

Required:

- **RequestID:** The identifier of the summary request.
- **Fact Text:** A short single-sentence statement that conveys an important, atomic and self-contained piece of information about the event.
- **Timestamp:** A UNIX epoch timestamp indicating the earliest point in time that your system detected the fact.
- **Importance:** A numerical score between 0 and 1 that indicates how important your system thinks it is for this fact to be included in the final summary.
- **Sources:** A list of StreamID identifiers that your system believes support the fact (what you used as input to produce the fact text).

Optional:

- **StreamID:** If your system is extractive (i.e., your facts are direct copies of text found in the input streams), include the StreamID of your fact's original item.
- **InformationNeeds:** If your system directly uses the queries/information-needs to decide what to include in your summary, and these can act as an explanation for what your system included an item, then list the queries that were core to this item's selection.

Examples of system output are as follows:

```
[
  {
    "requestID": "CrisisFACTS-001-r3",
    "factText": "Increased threat of wind damage in the San Diego area.",
    "unixTimestamp": 1512604876,
    "importance": 0.71,
    "sources": [
      "CrisisFACTS-001-Twitter-14023-0"
    ],
    "streamID": null,
    "informationNeeds": [
      "CrisisFACTS-General-q015"
    ]
  },
  {...}
]
```

Figure 4. Example System Output with Abstractive Facts. The streamID field is empty as this fact may not appear in the dataset verbatim. It is, however, supported by one Twitter message.

```
[
  {
    "requestID": "CrisisFACTS-001-r3",
    "factText": "Big increase in the wind plus drop in humidity
tonight into Thursday for San Diego County #SanDiegoWX
https://t.co/1pV0ZAhsJH",
    "unixTimestamp": 1512604876,
    "importance": 0.71,
    "sources": [
      "CrisisFACTS-001-Twitter-14023-0"
    ],
    "streamID": "CrisisFACTS-001-Twitter-14023-0",
    "informationNeeds": [
      "CrisisFACTS-General-q015"
    ]
  },
  { ... }
]
```

Figure 5. Example System Output with Extractive Facts. The streamID field is populated with the Twitter document from which this text was taken.

Participant systems may produce as many facts as they wish for a specific summary request. However, to handle variable summary length, each fact may not contain more than 200 characters. For days after the first, your system should avoid returning information that has been reported in previous summaries for the same event. Furthermore, evaluation will be performed at a predetermined number of facts (not revealed in advance). To truncate your list of facts, we will rank them by importance score and cut at a specific rank k – which will vary across event-day pairs.

We recommend that you return at least 100 facts per summary request.

Track Data

Datasets and Sources

For each day during an event, we are making available content from the following sources:

- **Twitter:** We are re-using tweets collected as part of the TREC Incident Streams track (<http://trecis.org>). These tweets were crawled by keyword and should be relevant to the

event (though some noise likely exists) but are not necessarily good candidates for inclusion into a summary of what is happening.

- **Reddit:** Discussions regarding what happens during events also occurs on the forum platform Reddit. We collected relevant Reddit threads to each event, where we include both the original submission and subsequent comments within those threads. We searched for relevant top-level posts and collected all comments associated with the post, so some comments may not be informative.
- **News:** Traditional news agencies are often a good source of information during an emergency, and we have included a small number of news articles collected during each event as well.
- **Facebook:** We collected Facebook/Meta posts from public pages that are relevant to each event using CrowdTangle. We cannot share the content of these posts, however, we have included the post and page ids of this content within the stream for those who have access to the CrowdTangle API and can retrieve this data separately.

We provide these data streams to participants to use as the source of content for inclusion into their summaries. The statistics for each of the 8 2022 events are listed below:

eventID	Title	Type	Description	Tweets	Reddit	News	Facebook
CrisisFACTS-001	Lilac Wildfire 2017	Wildfire	The Lilac Fire was a fire that burned in northern San Diego County, California, United States, and the second-costliest one of multiple wildfires that erupted in Southern California in December 2017.	41346	1738	2494	5437
CrisisFACTS-002	Cranston Wildfire 2018	Wildfire	The Cranston Fire was a wildfire that burned in southwest Riverside County, California, in the United States. The fire was started on July 25, 2018 by Brandon M. McGlover, in an act of arson. The fire burned a total of 13,139 acres (53 km ²), before it was fully contained on August 10.	22974	231	1967	5386
CrisisFACTS-003	Holy Wildfire 2018	Wildfire	The Holy Fire was a wildfire that burned in the Cleveland National Forest in Orange and Riverside Counties, California.	23528	459	1495	7016
CrisisFACTS-004	Hurricane Florence 2018	Hurricane	Hurricane Florence was a powerful and long-lived Cape Verde hurricane that caused catastrophic damage in the Carolinas in September 2018, primarily as a result of freshwater flooding due to torrential rain.	41187	120776	18323	196281
CrisisFACTS-005	Maryland Flood 2018	Flood	In the afternoon of May 27, 2018, after over 8 inches (20 cm) of rain in a span of two hours, the historic Main Street in Ellicott City, Maryland was flooded, just before the new flood emergency alert	33584	2006	2008	4148

			system was supposed to become operational.				
CrisisFACTS -006	Saddleridge Wildfire 2019	Wildfire	The Saddleridge Fire was a wildfire burning near the San Fernando Valley of Los Angeles County, California. It broke out roughly around 9:02 pm on Thursday October 10, 2019.	31969	244	2267	3869
CrisisFACTS -007	Hurricane Laura 2020	Hurricane	Hurricane Laura was a deadly and destructive Category 4 hurricane that is tied with the 1856 Last Island hurricane as the strongest hurricane on record to make landfall in the U.S. state of Louisiana, as measured by maximum sustained winds.	36120	10035	6406	9048
CrisisFACTS -008	Hurricane Sally 2020	Hurricane	Hurricane Sally was a destructive Atlantic hurricane which became the first hurricane to make landfall in the U.S. state of Alabama since Ivan in 2004, coincidentally on the same date in the same place. The eighteenth named storm, and seventh hurricane of the extremely active 2020 Atlantic hurricane season.	40695	11825	15112	48492

For each of these events, participants are to produce a ranking of facts per day, as defined in the summary requests. For reference, a breakdown of the event-day pairs is included in Appendix A.

Accessing the Data

CrisisFACTS has transitioned to the [ir_datasets](#) infrastructure for making data available to the community. We provide a GitHub repository with Jupyter notebooks and a Collab notebook to accelerate participants' access to this data:

- GitHub Repository: <https://github.com/crisisfacts/utilities>
 - Jupyter Notebook: <https://github.com/crisisfacts/utilities/blob/main/00-Data/00-CrisisFACTS.Downloader.ipynb>
- Google Collab Notebook: <https://colab.research.google.com/github/crisisfacts/utilities/blob/main/00-Data/00-CrisisFACTS.Downloader.ipynb>

Submission

Runs will be submitted through the NIST submission system at trec.nist.gov. Runs that do not pass validation will be rejected outright. Submitted runs will be asked to specify whether they

are (a) manual or automatic, whether they use (b) extant TREC-IS labels for Twitter data, and whether the summaries built are (c) abstractive or extractive.

Automatic Runs

Each run submission must indicate whether the run is manual or automatic. An automatic run is any run that receives no human intervention once the system is started and provided with the task inputs. We expect most CrisisFACTS runs to be automatic.

Manual Runs

Results on manual runs will be specifically identified when results are reported. A manual run is any run in which a person manually changes, summarises, or re-ranks queries, the system, or the system's lists of facts. Simple bug fixes that address only format handling do not result in manual runs, but the changes should be described.

Submission Format

The submission format for CrisisFACTS is a newline-delimited JSON file, where each entry in the submitted file contains the fields outlined in the “System Output” section above. Each submission file corresponds to a single submitted run (i.e., all event-day pairs for all events), with the submission's runtag included in the filename.

Example submission (of an extractive summary):

```
[
  {
    "requestID": "CrisisFACTS-001-r3",
    "factText": "Big increase in the wind plus drop in humidity
tonight into Thursday for San Diego County #SanDiegoWX
https://t.co/1pV0ZAhsJH",
    "unixTimestamp": 1512604876,
    "importance": 0.71,
    "sources": [
      "CrisisFACTS-001-Twitter-14023-0"
    ],
    "streamID": "CrisisFACTS-001-Twitter-14023-0",
    "informationNeeds": [
      "CrisisFACTS-General-q015"
    ]
  },
  {...}
]
```

Evaluation

Submitted systems will be evaluated using two sets of approaches for this first year. In both approaches, participant systems' lists of facts will be truncated to a private k value based on NIST assessors' assessments.

Metric Set 1 – Standard Summarization Metrics Against Extant Summaries

- First, the top- k facts for each system's event-day pair will be aggregated into a "summary" of that day's new developments. Then, this summary will be compared against extant summaries using standard summarization metrics (e.g., Rouge-1, Rouge-2, and/or Rouge-L).
- These extant summaries may include:
 - NIST-based assessments of these events and related summaries
 - Summaries extracted from the ICS209 reports from FEMA, which are available for each disaster event
 - Wikipedia articles summarising the events

Metric Set 2 – Matching Facts Between Participant System and Manual Assessments

- For this metric set, we first manually construct a map for matching a system's submitted facts for a given event-day pair to the NIST-assessor-developed set of facts for that event-day.
- For a given event-day pair, we therefore have:
 - A set of "gold standard" facts S produced by NIST assessors
 - A set of submitted facts $n \in N$ from a system
 - A fact-matching function $M(N, S)$ that measures the overlap between submitted facts and the gold-standard fact set. This fact-matching function comes in two forms:
 - Graded matching: $\sum_{s \in S} \text{MaxSim}(s, N)$
 - $\text{MaxSim}(x, Y)$ measures the maximum similarity in the range $[0, 1]$ between fact x and each fact in $y \in Y$.
 - Binary matching: $\sum_{s \in S} \text{BinaryMatch}(s, N)$
 - $\text{BinaryMatch}(s, N)$ returns 1 iff fact s is in set of facts N
 - Using this fact-matching function, we define **comprehensiveness** (essentially a recall-oriented metric):
 - $C(N, S) = M(N, S) / |S|$
 - We likewise define a **verbosity** metric as the number of submitted facts that match a gold-standard fact (i.e., precision):
 - $V(N, S) = M(S, N) / |N|$

- For each disaster event, we calculate the average comprehensiveness and verbosity scores over all days in the event where S contains facts.

Timeline

Guidelines released	17 June 2022
Submissions Due	26 September 2022
NIST-Assessor Evaluation	29 Sept - 27 Oct 2022
Scores returned to participants	28 October 2022
TREC Notebook Drafts Due	7 November 2022 (Tentative)
TREC Conference	14 November 2022

Appendix A. Daily Breakdown By Event

CrisisFACTS-001 - Lilac Wildfire 2017

crisisfacts/001/2017-12-07
crisisfacts/001/2017-12-08
crisisfacts/001/2017-12-09
crisisfacts/001/2017-12-10
crisisfacts/001/2017-12-11
crisisfacts/001/2017-12-12
crisisfacts/001/2017-12-13
crisisfacts/001/2017-12-14
crisisfacts/001/2017-12-15

CrisisFACTS-002 - Cranston Wildfire 2018

crisisfacts/002/2018-07-25
crisisfacts/002/2018-07-26
crisisfacts/002/2018-07-27
crisisfacts/002/2018-07-28
crisisfacts/002/2018-07-29
crisisfacts/002/2018-07-30

CrisisFACTS-003 - Holy Wildfire 2018

crisisfacts/003/2018-08-06
crisisfacts/003/2018-08-07
crisisfacts/003/2018-08-08
crisisfacts/003/2018-08-09
crisisfacts/003/2018-08-10
crisisfacts/003/2018-08-12
crisisfacts/003/2018-08-13

CrisisFACTS-004 - Hurricane Florence 2018

crisisfacts/004/2018-09-01
crisisfacts/004/2018-09-04
crisisfacts/004/2018-09-05
crisisfacts/004/2018-09-07
crisisfacts/004/2018-09-08
crisisfacts/004/2018-09-09
crisisfacts/004/2018-09-10
crisisfacts/004/2018-09-11
crisisfacts/004/2018-09-12
crisisfacts/004/2018-09-13
crisisfacts/004/2018-09-14
crisisfacts/004/2018-09-15
crisisfacts/004/2018-09-16
crisisfacts/004/2018-09-17
crisisfacts/004/2018-09-18

CrisisFACTS-005 - Maryland Flood 2018

crisisfacts/005/2018-05-27
crisisfacts/005/2018-05-28
crisisfacts/005/2018-05-29
crisisfacts/005/2018-05-30

CrisisFACTS-006 - Saddleridge Wildfire 2019

crisisfacts/006/2019-10-10
crisisfacts/006/2019-10-11
crisisfacts/006/2019-10-12
crisisfacts/006/2019-10-13

CrisisFACTS-007 - Hurricane Laura 2020
crisisfacts/007/2020-08-27
crisisfacts/007/2020-08-28

CrisisFACTS-008 - Hurricane Sally 2020
crisisfacts/008/2020-09-11
crisisfacts/008/2020-09-12
crisisfacts/008/2020-09-13
crisisfacts/008/2020-09-14
crisisfacts/008/2020-09-15
crisisfacts/008/2020-09-16
crisisfacts/008/2020-09-17
crisisfacts/008/2020-09-18