# Echobox Data Science Challenge

Breaking News Detection

**IMPORTANT: The information and data contained in this challenge remains the property of Echobox and may not be distributed or communicated in any form beyond the intended candidate.**

To better understand whether you’d be a fit for Echobox, we kindly ask you to complete the following data science challenge. We use your submission to evaluate how you would approach/attack any problem during your work at Echobox. We evaluate many factors, such as analysis approach, potential code structure, accuracy, documentation, etc.

Should you run out of time, we are happy for you to submit something ‘incomplete’ providing you describe any incomplete steps, failing test cases and/or additional sophistications you would have investigated or added given further time.

If you have any questions regarding this challenge please feel free to email [simon@echobox.com](mailto:simon@echobox.com) for clarification (Please see below for the correct completed submission email address).

## Background:

The main goal of Echobox is to intelligently automate sharing on Social. When a news article is published in an [RSS](https://en.wikipedia.org/wiki/RSS) feed, Echobox is able to determine how best to share this article on social media by optimising several factors: the time at which it gets shared is one of the most important factors, and determining this optimally is key to maximising the amount of engagement/clicks/traffic this article will get once shared.

This optimal sharing time depends heavily on the content of the article itself: some articles are reporting on events that happened recently and thus these articles are only relevant for a short period of time after being published. These articles are often referred as [breaking news](https://en.wikipedia.org/wiki/Breaking_news) articles. It is particularly important for Echobox to detect such articles and identify to which event they relate, as they need to be prioritized over other articles that might be shared later during the day.

We have provided a sample dataset of english articles published from different news websites over a period of 2 days, which can be downloaded [here](https://docs.google.com/spreadsheets/d/14TsdYfeHU-XRAXRRk8JrxIt1KAB0T8ZO2FLsvqUJLgg/edit#gid=724625444). Each entry in this dataset corresponds to an article. For each of these articles, we have provided publicly available information:

* **ArticleId:** A unique string identifying the article.
* **ArticleURL**: URL of the article.
* **ArticleTitle**: Title of the article.
* **ArticleDescription:** A short description of the article. Note that sometimes we don’t get this information from the RSS feed so it can be empty.
* **ArticlePublishedTime**: Published time of the article (These times are in [Unix time format](https://en.wikipedia.org/wiki/Unix_time))

## The challenge:

Develop a solution, using the tools of your choice, that given this list of articles attempts to identify the breaking news events that these articles relate to, and extract relevant information for these events.

This challenge is broken down into three main parts:

### 1. Identify major breaking news events

Major breaking news events can be identified by noticing that many articles reporting on the same news event have similar content, and are published relatively quickly after the event has happened.

For this first part, we ask you to group articles related to the same breaking news event by assigning an eventId to each article. Articles that don’t relate to a major breaking news event should be labelled as “-1”.

Example outcome:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ArticleURL | ArticleTitle | ArticleDescription | ArticlePublishedTime | EventId |
| [URL1](https://www.smh.com.au/sport/tennis/djokovic-prevails-over-tsitsipas-to-claim-third-madrid-title-20190513-p51mn8.html) | Djokovic prevails over Tsitsipas to claim third Madrid title | [...] | 1557697145 | **1** |
| [URL2](https://www.completesports.com/int/djokovic-claims-madrid-masters-title/) | Djokovic Claims Madrid Masters Title | [...] | 1557697256 | **1** |
| [URL3](https://www.tnp.sg/news/world/myanmar-plane-lands-safely-no-front-wheels) | Myanmar plane lands safely with no front wheels | [...] | 1557698400 | **2** |
| [URL4](https://www.theguardian.com/world/video/2019/may/13/smoke-fills-myanmar-plane-after-emergency-landing-without-front-wheel-video) | Smoke fills Myanmar plane after emergency landing without front wheel – video | [...] | 1557709532 | **2** |
| [URL5](https://triblive.com/opinion/gisele-fetterman-american-dream-should-be-for-everyone/) | Gisele Fetterman: American dream should be for everyone | [...] | 1557664199 | **-1** |

In this example, articles with [URL1](https://www.smh.com.au/sport/tennis/djokovic-prevails-over-tsitsipas-to-claim-third-madrid-title-20190513-p51mn8.html) and [URL2](https://www.completesports.com/int/djokovic-claims-madrid-masters-title/) are related to the same breaking news event: the victory of Novak Dojokovic over Stefanos Tsitsipas at the [2019 Madrid Open](https://en.wikipedia.org/wiki/2019_Mutua_Madrid_Open). They are hence labelled with the same EventId (EventId **1**).

On the contrary, the article with [URL5](https://triblive.com/opinion/gisele-fetterman-american-dream-should-be-for-everyone/) is an opinion piece: it can’t be linked to any other article of the dataset and don’t relate to a breaking news event. The EventId is thus **-1**.

Please use this format as it makes it quicker for us to evaluate your submission.

### 2. Extract information about breaking news events

Once you have identified which articles are linked to their respective events, please attempt to derive meaningful information about these events.

We are open about the outcome here: you could consider identifying the time, or how long ago relatively speaking, at which the breaking news event happened, or where these events occurred.

Example output:

|  |  |  |  |
| --- | --- | --- | --- |
| EventId | EventLocation | EventTime | EventSummary |
| **1** | Madrid, Spain | 1557687600 | Djokovic beats Tsitsipas to win third Madrid Open title |
| **2** | Mandalay, Myanmar | 1557691200 | Pilot lands plane safely without front wheels |

In this example, we have provided some information about the Event 1, which corresponds to the Men's Singles Final of the [2019 Madrid Open](https://en.wikipedia.org/wiki/2019_Mutua_Madrid_Open). This event occurred in Madrid, on the 12th of May and saw Novak Djokovic beat Stephanos Tsitsipas.

### 3. Further research ideas

1. What curated labels could be included in this data set to allow for alternative approaches to the one you used? How would you use these labels?
2. Please elaborate on any other ideas you would have enjoyed exploring within this problem space/data set. If we haven’t already explored your suggestion it may form part of your internship.

## Libraries:

You are free to use whatever libraries you wish.

## Submission:

Submit your completed solution (including all code written) as a single package/file (e.g. ZIP) to [operations@echobox.com](mailto:operations@echobox.com).