New entrants are currently prohibited. You must wait until after the deadline (6.7 days) to download the data or submit.

$25,000 • 959 teams

Outbrain Click Prediction

Merger and Entry Deadline

Wed 5 Oct 2016

Wed 18 Jan 2017 (6.7 days to go)

Public Leaderboard

1. Three Data Points
2. code monkey
3. brain-afk
4. Neuron
5. FG Knight
6. Andrii Cherednychenko
7. CV
8. Sangxia
9. rokh
10. Igor Pasechnik

1,451 Kernels

Evaluate map score without groupby in python
14 Votes / 16 days ago / Python

Outbrain EDA
191 Votes / 3 months ago / Python

Data Files

File Name	Available Formats
documents_categories.csv	.zip (32.34 mb)
clicks_test.csv	.zip (135.43 mb)
documents_meta.csv	.zip (15.51 mb)
documents_entities.csv	.zip (125.67 mb)
promoted_content.csv	.zip (2.52 mb)
sample_submission.csv	.zip (99.57 mb)
documents_topics.csv	.zip (120.91 mb)
clicks_train.csv	.zip (389.75 mb)
events.csv	.zip (477.74 mb)
page_views.csv	.zip (29.71 gb)
page_views_sample.csv	.zip (148.51 mb)

The dataset for this challenge contains a sample of users' page views and clicks, as observed on multiple publisher sites in the United States between 14-June-2016 and 28-June-2016. Each viewed page or clicked recommendation is further accompanied by some semantic attributes of those documents. For full details, see data specifications below.

The dataset contains numerous sets of content recommendations served to a specific user in a specific context. Each context (i.e. a set of recommendations) is given a display_id. In each such set, the user has clicked on at least one recommendation. The identities of the clicked recommendations in the test set are not revealed. Your task is to rank the recommendations in each group by decreasing predicted likelihood of being clicked.

Unveiling page_views.csv with PySpark
87 Votes / 2 months ago / Python

pypy implementation of fm(with adam)
6 Votes / 5 days ago / Python

FTRL Starter (with leakage vars)
33 Votes / 2 months ago / Python

pandas is cool! LB: 0.63714
25 Votes / 2 months ago / Python

Forum (116 topics)

FFM Input
4 minutes ago

Best single model performance
14 minutes ago

let's break 0.7
3 hours ago

Why pandas read both strings and integers value for platforms in events.csv
5 hours ago

Future publish_time in documents_meta.csv file
8 hours ago

Benchmark 0.65251 using BTB with 95GB page_views.csv
12 hours ago

teams

players

entries

As a warning, this is a very large relational dataset. While most of the tables are small enough to fit in memory, the page views log (page_views.csv) is over 2 billion rows and 100GB uncompressed. We have also uploaded a sample version of this file with the first 10,000,000 rows. The MD5 checksum of page_views.csv.zip is 3742c116bab4030e0a7ea1c0be623bd9.

Data Fields

Each user in the dataset is represented by a unique id (uuid). A person can view a document (document_id), which is simply a web page with content (e.g. a news article). On each document, a set of ads (ad_id) are displayed. Each ad belongs to a campaign (campaign_id) run by an advertiser (advertiser_id). You are also provided metadata about the document, such as which entities are mentioned, a taxonomy of categories, the topics mentioned, and the publisher.

File Descriptions

page_views.csv is a the log of users visiting documents. To save disk space, the timestamps in the entire dataset are relative to the first time in the dataset. If you wish to recover the actual epoch time of the visit, add 1465876799998 to the timestamp.

- uuid
- document_id
- timestamp (ms since 1970-01-01 - 1465876799998)
- platform (desktop = 1, mobile = 2, tablet =3)
- geo_location (country>state>DMA)
- traffic_source (internal = 1, search = 2, social = 3)

clicks_train.csv is the training set, showing which of a set of ads was clicked.

- display_id
- ad_id
- clicked (1 if clicked, 0 otherwise)

clicks_test.csv is the same as clicks_train.csv, except it does not have the clicked ad. This is the file you should use to predict. Each display_id has only one clicked ad. Note that test set contains display_ids from the entire dataset timeframe. Additionally, the public/private sampling for the competition is uniformly random, not based on time. These sampling choices were intentional, in spite of the possibility that participants can look ahead in time.

sample_submission.csv shows the correct submission format.

events.csv provides information on the display_id context. It covers both the train and test set.

- display_id
- uuid
- document_id
- timestamp
- platform
- geo_location

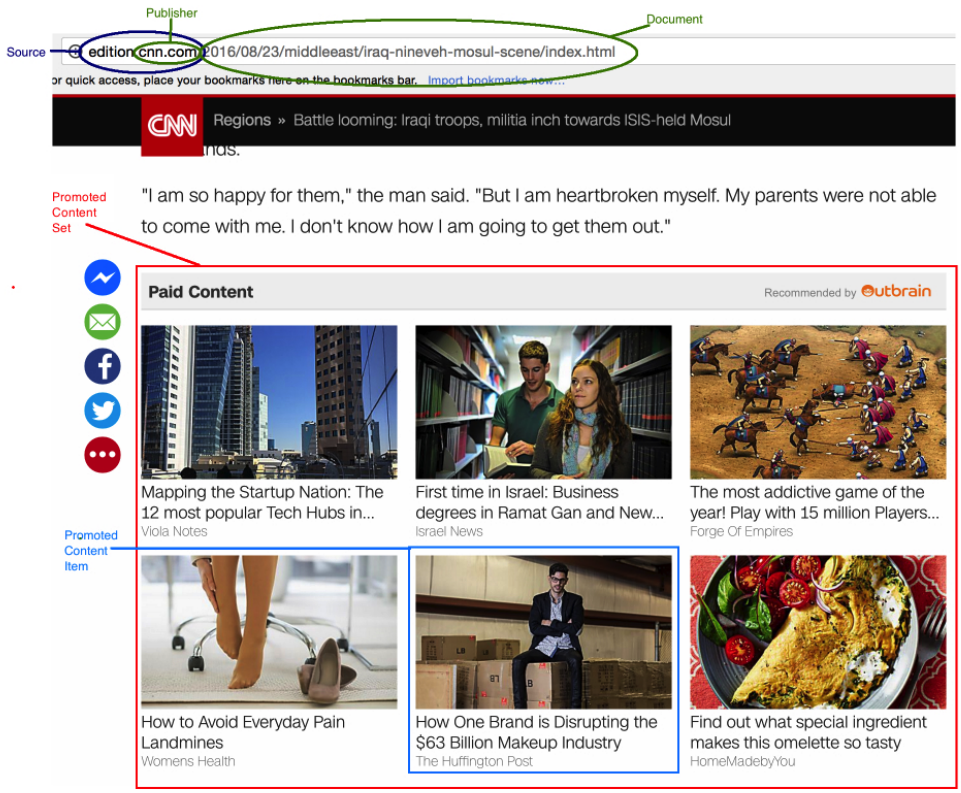
promoted_content.csv provides details on the ads.

- ad_id
- document_id
- campaign_id
- advertiser_id

documents_meta.csv provides details on the documents.

- document_id
- source_id (the part of the site on which the document is displayed, e.g. edition.cnn.com)
- publisher_id
- publish_time

documents_topics.csv, **documents_entities.csv**, and **documents_categories.csv** all provide information about the content in a document, as well as Outbrain's confidence in each respective relationship. For example, an entity_id can represent a person, organization, or location. The rows in documents_entities.csv give the confidence that the given entity was referred to in the document.



Privacy Reminder

Outbrain is releasing 2 Billion page views and 16,900,000 clicks of 700 Million unique users, across 560 sites. The data is anonymized. Please remember that participants are prohibited from de-anonymizing or reverse engineering data or combining the data with other publicly available information. Outbrain does not collect or hold PII (personally identifiable information), and the user identifiers we are releasing here are obscured. To protect its publisher partners, Outbrain is not releasing URLs of viewed or

clicked stories, but rather anonymized document and site identifiers. The task at hand is click prediction, and by downloading the dataset, participants agree to use the data for that task alone, and will not attempt to reverse engineer the mapping from document, site, and user identifiers to URLs, site names or actual users.