

---

---

# Outbrain Click Prediction

[Kaggle link](#)

---

---

# Outline

Task definition

Dataset

File description

Evaluation

Rules

Some tips

Pros and Cons

Link

# Task definition

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 context (i.e. a set of recommendations) is given a `display_id`.

Your task is to **rank** the recommendations in each group by decreasing predicted likelihood of being clicked.

As a warning, this is a **very large** relational dataset.

Source [edition.cnn.com/2016/08/23/middleeast/iraq-nineveh-mosul-scene/index.html](http://edition.cnn.com/2016/08/23/middleeast/iraq-nineveh-mosul-scene/index.html) Publisher Document

or quick access, place your bookmarks here on the bookmarks bar. [Import bookmarks now...](#)

**CNN** Regions » Battle looming: Iraqi troops, militia inch towards ISIS-held Mosul


mas.

Promoted Content Set


"I am so happy for them," the man said. "But I am heartbroken myself. My parents were not able to come with me. I don't know how I am going to get them out."




**Paid Content** Recommended by **outbrain**




**Mapping the Startup Nation: The 12 most popular Tech Hubs in...**  
Viola Notes




**First time in Israel: Business degrees in Ramat Gan and New...**  
Israel News




**The most addictive game of the year! Play with 15 million Players...**  
Forge Of Empires



**How to Avoid Everyday Pain Landmines**  
Womens Health



**How One Brand is Disrupting the \$63 Billion Makeup Industry**  
The Huffington Post



**Find out what special ingredient makes this omelette so tasty**  
HomeMadebyYou

Promoted Content Item

# Dataset

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

You can download files from kaggle or download it from Google Drive TA provided. [\(Link\)](#)

# File description (1/5)

## page\_views.csv

the log of users visiting documents.

- uuid (unique user id)
- 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)

DMA = Designated Market Area ([Wiki Link](#))

# File description (2/5)

## clicks\_train.csv

It 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

It 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.

```
display_id, ad_id, clicked
1, 42337, 0
1, 139684, 0
1, 144739, 1
1, 156824, 0
2, 125211, 0
2, 156535, 0
2, 308455, 1
3, 71547, 0
3, 95814, 0
3, 228657, 1
3, 250082, 0
...
```

clicks\_train.csv

## File description (3/5)

### **events.csv**

It provides information on the display\_id context.

It covers both the training and test set.

- display\_id
- uuid
- document\_id
- Timestamp
- Platform
- geo\_location



## File description (4/5)

### **documents\_meta.csv**

It provides details on the documents.

- document\_id
- source\_id
- 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.

# File description (5/5)

## **promoted\_content.csv**

It provides details on the ads.

- ad\_id
- document\_id
- campaign\_id
- advertiser\_id

# Evaluation

Submissions are evaluated according to the [Mean Average Precision @12](#) (MAP@12):

$$MAP@12 = \frac{1}{|U|} \sum_{u=1}^{|U|} \sum_{k=1}^{\min(12,n)} P(k)$$

where  $|U|$  is the number of display\_ids,  $P(k)$  is the precision at cutoff  $k$ ,  $n$  is the number of predicted ad\_ids.

# Rules

- [Follow the rules of this competition](#)
- **2** submissions per day
- Submission file format

```
display_id,ad_id
```

```
16874594,66758 150083 162754 170392 172888 180797
```

```
16874595,8846 30609 143982
```

```
16874596,11430 57197 132820 153260 173005 288385 289122 289915
```

```
etc.
```

# Some Tips

Statistic

How to extract features?

How to evaluate your model before you make submission?

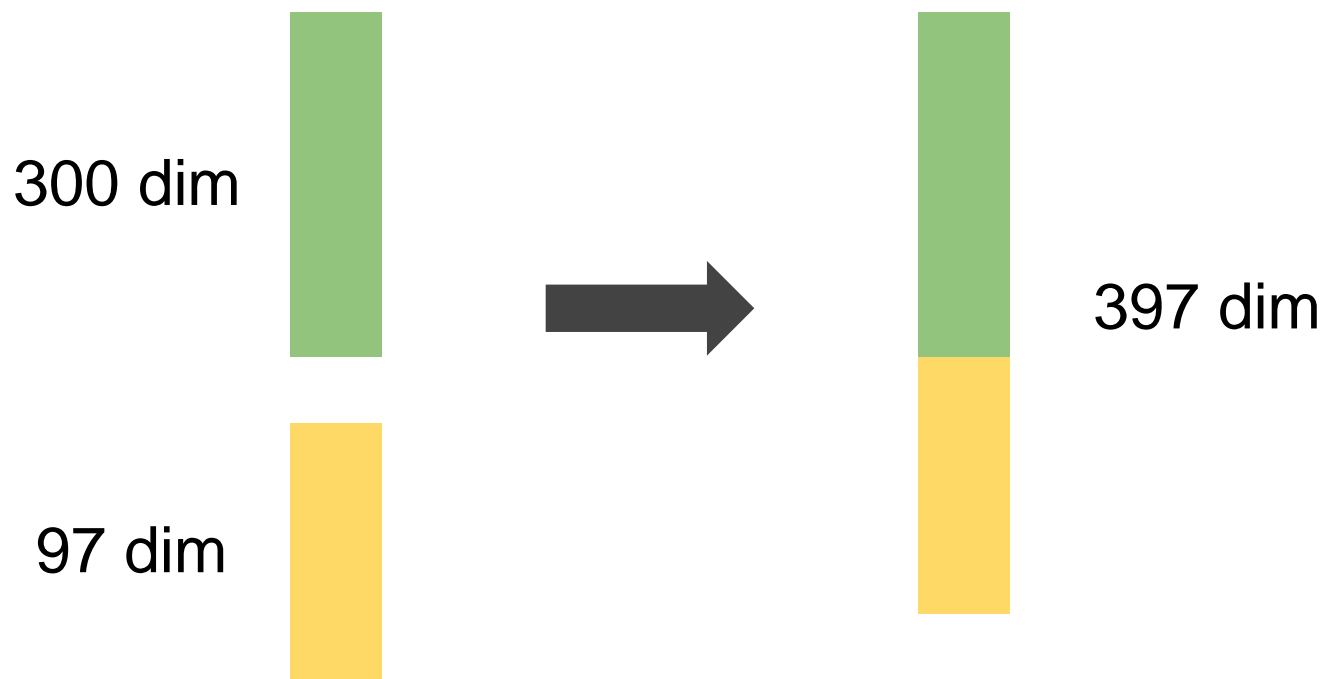
How to consider relation between ads in a display?

## Some Tips - How to extract features?

Each document belongs to a topic and category.

You can use one-hot vector to represent a document.

e.g. 300 topics, 97 category



## Pros & Cons

- 1st Prize: **\$12,000 USD** ( = 383,521 NTD)
- 2nd Prize: \$8,000 USD
- 3rd Prize: \$5,000 USD
  
- The size of this dataset is very large. (about **100GB**)
- Only **2** submission per day

# Link

- [Kaggle link](#)
- [Dataset Link](#) (google drive, TAs provide)