

DataLab Cup 1: Predicting News Popularity


Outline




- Competition Info
- Feature Engineering




Competition Info


- News Popularity

EDITION: UNITED STATES

 **REUTERS**

 [Business](#) [Markets](#) [World](#) [Politics](#) [Tech](#) [Breakingviews](#) [Wealth](#) [Life](#)  [Pictures](#)  [Video](#)



THE WIRE

2m ago

Wirecard rejects FT report as shares drop

6m ago

HSBC taps Lazard to sell French retail business: source

6m ago


Trump sanctions fail to slow Turkey assault; Syrian troops move on Manbij

9m ago

Two killed at Saudi Aramco's SASREF refinery during maintenance

9m ago

UK's Dover is ready for Brexit



Neighbors turn enemies over Trump in Michigan suburbs

A kind of suburban trench warfare is simmering amid the small detached houses in this pivotal state in the 2020 election where diehard Trump lovers live next to Trump haters.

- Impeachment inquiry turns to career diplomat
- Hunter Biden defends overseas work
- Impeachment, Warren-Biden matchup highlight Democratic debate

How Amazon.com moved into the business of U.S. elections

9:10AM EDT

Exclusive: Trump lawyer Giuliani was paid \$500,000 to consult on indicted associate's firm

5:49AM EDT

U.S. pension funds took positions in blacklisted Chinese surveillance company

6:21AM EDT

Exclusive: Deutsche Bank took years to flag suspect Danske money flows - source

7:34AM EDT

MARKETS

STOCKS

BONDS

CURRENCIES

COMMODITIES

S&P »	2,988.00	+0.74%
Dow »	26,991.51	+0.76%
Nasdaq »	8,111.97	+0.79%
FTSE 100 »	7,189.89	-0.33%
Nikkei 225 »	22,207.21	+1.87%

SPONSORED CONTENT

Competition Info

- Dataset

- Training data(27643)

Id	Popularity	Page content
0	-1	<html><head><div class="article-info"> <span c...
1	1	<html><head><div class="article-info"><span cl...
2	1	<html><head><div class="article-info"><span cl...
3	-1	<html><head><div class="article-info"><span cl...
4	-1	<html><head><div class="article-info"><span cl...

- Testing data(11847)

Id	Page content
27643	<html><head><div class="article-info"><span cl...
27644	<html><head><div class="article-info"><span cl...
27645	<html><head><div class="article-info"><span cl...
27646	<html><head><div class="article-info"><span cl...
27647	<html><head><div class="article-info"><span cl...

Competition Info

- Dataset

 By [Sara Roncero-Menendez](#) 2014-01-03 13:02:48 UTC

Mobile Advertising Projected to Increase 64% in 2014



As our web presence expands, so does the advertising space. Agencies are using [mobile](#) and [native advertising](#) to catch consumers' attention on a variety of online platforms.

Companies nearly tripled the amount of money spent on mobile advertising, from \$1.2 billion in 2012 to \$3 billion in 2013, according to [LinkedIn Marketing Solutions](#). Roughly 65% of both ad agencies and marketers plan to invest in native advertising, for an estimated total of \$4.3 billion, in 2014.

See also: [10 Tips for Improving Your Mobile Advertising Campaign](#)

Social and mobile marketing go hand-in-hand, since at least 17% of the time people spend on their mobile devices is on a social network. It's no wonder then that analysts predict mobile and social advertising will increase 64% and 47%, respectively.

Marketers are expected to spend nearly \$47.6 billion on online ads alone in 2014, with \$13.1 billion of that figure allocated for mobile ads.


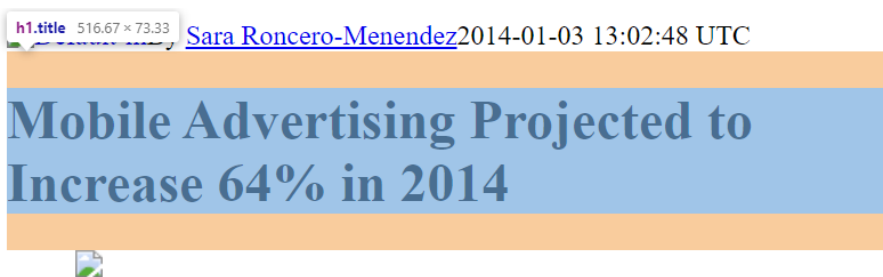
 123013-LinkedIn-Mobile-Ads.nr.KF.jcf

Image: *LLUIS GENE/AFP/Getty Images*

Topics: [Advertising](#), [Business](#), [infographics](#), [Marketing](#), [Mobile](#), [mobile advertising](#)

Competition Info

- Dataset

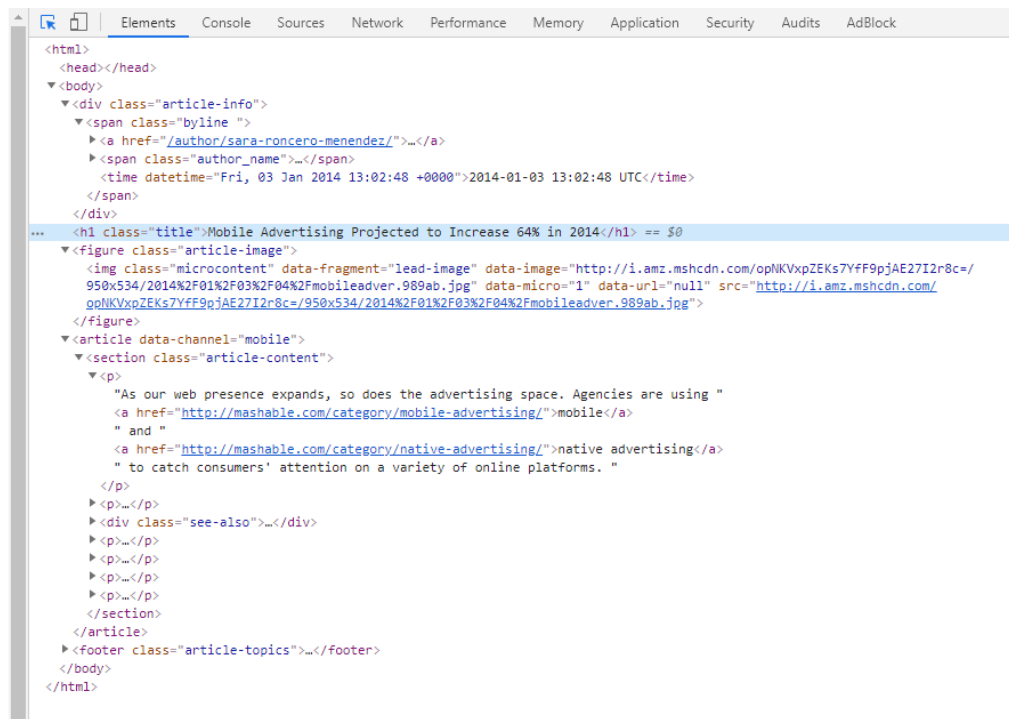


As our web presence expands, so does the advertising space. Agencies are using [mobile](#) and [native advertising](#) to catch consumers' attention on a variety of online platforms.

Companies nearly tripled the amount of money spent on mobile advertising, from \$1.2 billion in 2012 to \$3 billion in 2013, according to [LinkedIn Marketing Solutions](#). Roughly 65% of both ad agencies and marketers plan to invest in native advertising, for an estimated total of \$4.3 billion, in 2014.

See also: [10 Tips for Improving Your Mobile Advertising Campaign](#)

Social and mobile marketing go hand-in-hand, since at least 17% of the time people spend on their mobile devices is on a social network. It's no wonder then that analysts predict mobile and social advertising will increase 64% and 47%,



Competition Info

- Evaluation metric
 - AUC

The screenshot shows the header of a competition page. At the top left is a logo with a graduation cap and the text "InClass Prediction Competition". Below this is the main title "DataLab Cup1: Predicting News Popularity" and the subtitle "Competition for CS565600 Deep Learning". To the left of the title, it says "23 days to go". A navigation bar at the bottom contains links: Overview, Data, Notebooks, Discussion, Leaderboard, Rules, Team, Host, My Submissions, and Submit Predictions. The "Leaderboard" link is underlined. Three red arrows point from text annotations to specific elements: "Download data" points to the "Data" link, "View results and rankings" points to the "Leaderboard" link, and "Submit the results" points to the "Submit Predictions" button.

InClass Prediction Competition

DataLab Cup1: Predicting News Popularity

Competition for CS565600 Deep Learning

23 days to go

Download data

View results and rankings

Submit the results

Overview Data Notebooks Discussion Leaderboard Rules Team Host My Submissions **Submit Predictions**

Competition Info

- Evaluation metric
 - AUC

[Public Leaderboard](#)

[Private Leaderboard](#)

This leaderboard is calculated with approximately 50% of the test data.

The final results will be based on the other 50%, so the final standings may be different.

[Raw Data](#) [Refresh](#)

#	Team Name	Notebook	Team Members	Score ?	Entries	Last
📍	BenchMark-80.csv			0.57298		
📍	BenchMark-60.csv			0.54396		


Competition Info

- Kaggle

How to Submit Results?

You have to predict the correct labels of data points in `test.csv` and submit your predictions to the [Kaggle-In-class](#) online judge system to get scores. Following are some example actions:

Action	Description
Data	Get the dataset.
Make a Submission	Your testing performance will be evaluated immediately and shown on the leaderboard.
Leaderboard	The current ranking of participants. Note that this ranking only reflects the performance on part of the testset and may not equal to the final ranking (see below).
Forum	You can ask questions or share findings here.
Kernels	You can create your jupyter notebook, run it, and keep it as private or public here.



Competition Info

- Rules
 - What you can do
 - Use untaught APIs: you can use any machine learning tools you like as well as models/techniques that are not taught in the class.

Competition Info

- Rules
 - What you **can't** do
 - Attempt to make predictions by means other than "learning" from the given dataset X or related sources.
 - Train models using representation learning based on neural networks.

Competition Info

- Honor code
 - Cheating is forbidden
 - Attempting to use datasets and references beyond those made available by the competition
 - Attempting to abuse the competition infrastructure to gain an edge
 - Attempting to copy code from other teams

Competition Info

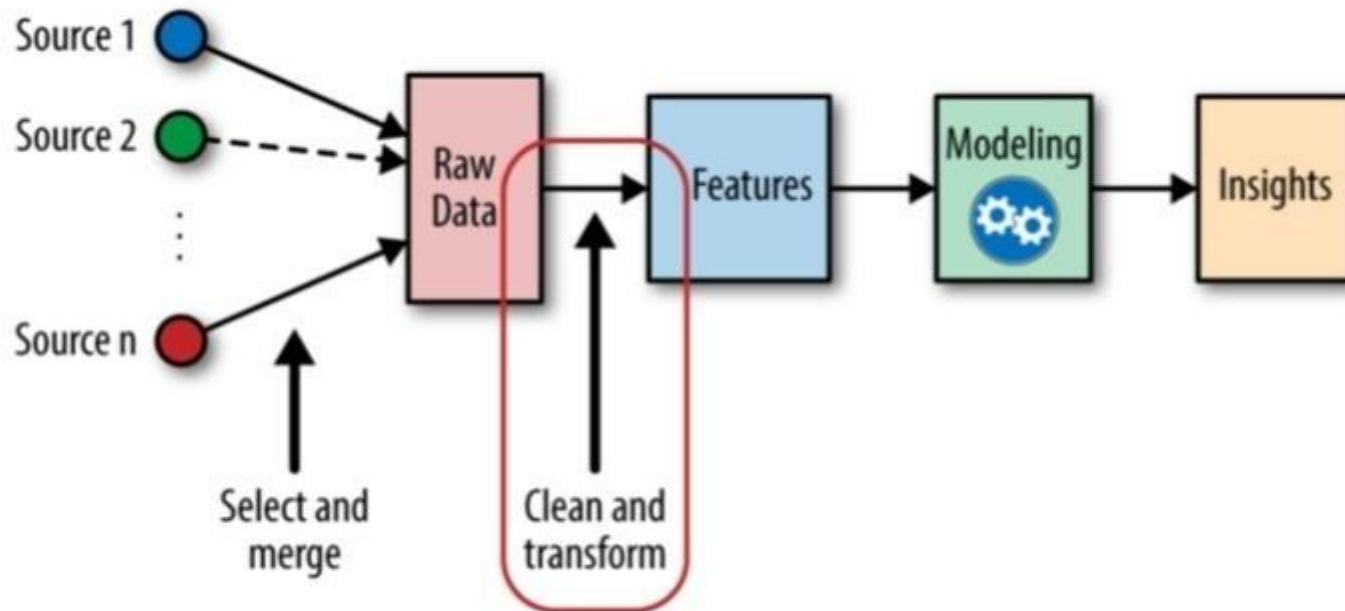
- Important Dates
 - 2019/10/17 (Thur) - competition starts
 - 2019/11/07 (Thur) 23:59pm - competition ends, final score announcement
 - 2019/11/10 (Sun) 23:59pm - report submission (to iLMS)
 - 2019/11/12 (Tue) - competition 1 show off

Competition Info

- Report
 - Student ID, name of each team member
 - How did you preprocess data
 - How did you build the classifier
 - Conclusions

Feature Engineering

- Feature Engineering is More Important Than You Expected



Feature Engineering

- Preprocessing
 - Data Cleaning
 - Extended abbreviation
 - Word Stemming
 - Stop-Word Removal
- Word2Vec
 - BoW
 - TF-IDF
 - Feature Hashing
- Out-of-Core Learning

Feature Engineering

- Preprocessing
 - Data Cleaning

I know that Chill Wills usually played lovable old sorts in Westerns. But his role in this segment is something I've remembered for a long time. Wills could be a first rate villain. Yes, Burgess Meredith's Fall was correct! That look in Hepplewhite's eye! It expressed porcine greed, ignorance, and the threat of violence all at once. Quite a performance, I think.

The segment itself was a good one, too. Question: couldn't the little black bag cure alcoholism? I guess it did, sort of, with Fall. But the doctor would have been wise to apply the cure, if he had it, as quickly as possible to Hepplewhite.

There is one moment that was annoying but also necessary. And it is something that appears to recur in these Night Gallery segments. It's Serling's constant need to sermonize. For that's what we got, one more time, with Dr. Fall. I don't know what was more frustrating, losing the black bag and all its miracles or not being to stop Fall from preaching about the bag's benefit for humanity, all while rubbing Hepplewhite's greedy face in the mud, and, therefore, all but begging for Hepplewhite to strike out at him. But as I say, it was necessary. At least it was for me. Otherwise, we wouldn't have been able to see Wills' performance discussed above. All done without moving a muscle or speaking a word.

Feature Engineering

- Preprocessing
 - Data Cleaning
 - removing all HTML tags
 - removing punctuation marks but emoticons
 - converting all characters to lowercase

```
<a href="example.com">Hello, This :-( is a sanity check ;P!</a>
```

```
hello this is a sanity check :( ;P
```

Feature Engineering

- Preprocessing
 - Extended abbreviation
 - don't -> do not
 - I'd -> i would

Feature Engineering

- Preprocessing
 - Word Stemming
 - watches/watching/watched -> watch

runners like running and thus they run

['runner', 'like', 'run', 'and', 'thu', 'they', 'run']

Feature Engineering

- Preprocessing
 - Stop-Word Removal
 - a/an/the
 - am/is/are

Feature Engineering

- Word2Vec
 - BoW

John likes to watch movies, Mary likes movies too

John also likes to watch football games

(also,0) (football,1)
(games,2) (john,3) (likes,4)
(mary,5) (movies,6) (to,7)
(too,8) (watch,9)

	also	football	games	john	likes	mary	movies	to	too	watch
s1 =	0	0	0	1	2	1	2	1	1	1
s2 =	1	1	1	1	1	0	0	1	0	1

Feature Engineering

- Word2Vec
 - TF-IDF
 - TF: Term frequency
 - IDF: Inverse document frequency

$$TF-IDF = TF \cdot \left(\log \left(\frac{1 + N_{\text{doc}}}{1 + DF} \right) + 1 \right)$$

Feature Engineering

- Word2Vec
 - Feature Hashing
 - (+) no need to store vocabulary dictionary in memory anymore
 - (-) no way to map token index back to token
 - (-) no IDF weighting

Feature Engineering

- Out-of-Core Learning
 - data streaming
 - partial update