kaggle Q Search

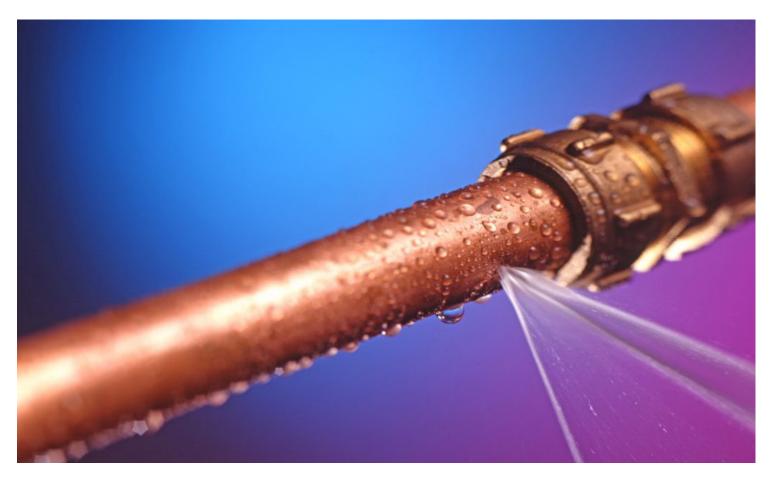
Competitions Datasets Notebooks Discussion Courses

Submission

✓ Ran successfully

Submitted by szelee 3 hours ago

Public Score 1.00000



Hate to spoil the fun, but I've checked the origin (https://www.figure-eight.com/data-for-everyone/) of the dataset (look for 'Disasters on social media') provided in the overview page (https://www.kaggle.com/c/nlp-getting-started/overview) and found the dataset that holds ground truth for the test set.

If I can discover this so easily, I am sure it's just a matter of time before someone else does the same. Or most likely I am not the first, just sayin'.

```
In [1]:
    import pandas as pd

In [2]:
    test_df = pd.read_csv('/kaggle/input/nlp-getting-started/test.csv')
    gt_df = pd.read_csv("../input/disasters-on-social-media/socialmedia-disaster-tweets-DFE.cs
    v")
```

```
In [3]:
    gt_df = gt_df[['choose_one', 'text']]
    gt_df['target'] = (gt_df['choose_one']== 'Relevant').astype(int)
    gt_df['id']= gt_df.index
    gt_df
```

Out[3]:

	choose_one	text	target	id
0	Relevant	Just happened a terrible car crash	1	0
1	Relevant	Our Deeds are the Reason of this #earthquake M	1	1
2	Relevant	Heard about #earthquake is different cities, s	1	2
3	Relevant	there is a forest fire at snot nond, deese are	1	3

9	Relevant	there is a forest file at spot polita, geese are	'	3	Г
4	Relevant	Forest fire near La Ronge Sask. Canada	1	4	
	•••				
10871	Relevant	M1.94 [01:04 UTC]?5km S of Volcano Hawaii. htt	1	10871	
10872	Relevant	Police investigating after an e-bike collided	1	10872	
10873	Relevant	The Latest: More Homes Razed by Northern Calif	1	10873	
10874	Relevant	MEG issues Hazardous Weather Outlook (HWO) htt	1	10874	
10875	Relevant	#CityofCalgary has activated its Municipal Eme	1	10875	

10876 rows × 4 columns

```
In [4]:
    merged_df = pd.merge(test_df, gt_df, on='id')
    merged_df
```

Out[4]:

	id	keyword	location	text_x	choose_one	text_y	target
0	0	NaN	NaN	Just happened a terrible car crash	Relevant	Just happened a terrible car crash	1
1	2	NaN	NaN	Heard about #earthquake is different cities, s	Relevant	Heard about #earthquake is different cities, s	1
2	3	NaN	NaN	there is a forest fire at spot pond, geese are	Relevant	there is a forest fire at spot pond, geese are	1
3	9	NaN	NaN	Apocalypse lighting. #Spokane #wildfires	Relevant	Apocalypse lighting. #Spokane #wildfires	1
4	11	NaN	NaN	Typhoon Soudelor kills 28 in China and Taiwan	Relevant	Typhoon Soudelor kills 28 in China and Taiwan	1
3258	10861	NaN	NaN	EARTHQUAKE SAFETY LOS ANGELES □ÛÒ SAFETY FASTE	Not Relevant	EARTHQUAKE SAFETY LOS ANGELES ‰ÛÒ SAFETY FASTE	0
3259	10865	NaN	NaN	Storm in RI worse than last hurricane. My city	Relevant	Storm in RI worse than last hurricane. My city	1
3260	10868	NaN	NaN	Green Line derailment in Chicago http://t.co/U	Relevant	Green Line derailment in Chicago http://t.co/U	1
3261	10874	NaN	NaN	MEG issues Hazardous Weather Outlook (HWO) htt	Relevant	MEG issues Hazardous Weather Outlook (HWO) htt	1
3262	10875	NaN	NaN	#CityofCalgary has activated its Municipal Eme	Relevant	#CityofCalgary has activated its Municipal Eme	1

3263 rows × 7 columns

```
In [5]:
    subm_df = merged_df[['id', 'target']]
    subm_df
```

Out[5]:

	id	target
0	0	1
1	2	1
2	3	1
3	9	1
4	4.4	4

4	11	1
•••		
3258	10861	0
3259	10865	1
3260	10868	1
3261	10874	1
3262	10875	1

3263 rows × 2 columns

In [6]:
 subm_df.to_csv('submission.csv', index=False) # The holy grail of perfect prediction?

Speaking of real disasters eh?

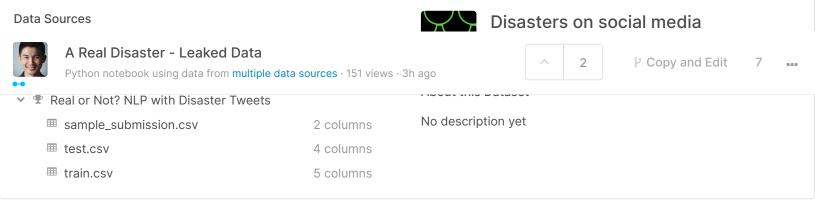
This Notebook has been released under the Apache 2.0 open source license.

Did you find this Notebook useful? Show your appreciation with an upvote





Data



Output Files

Output Files

■ submission.csv

About this file

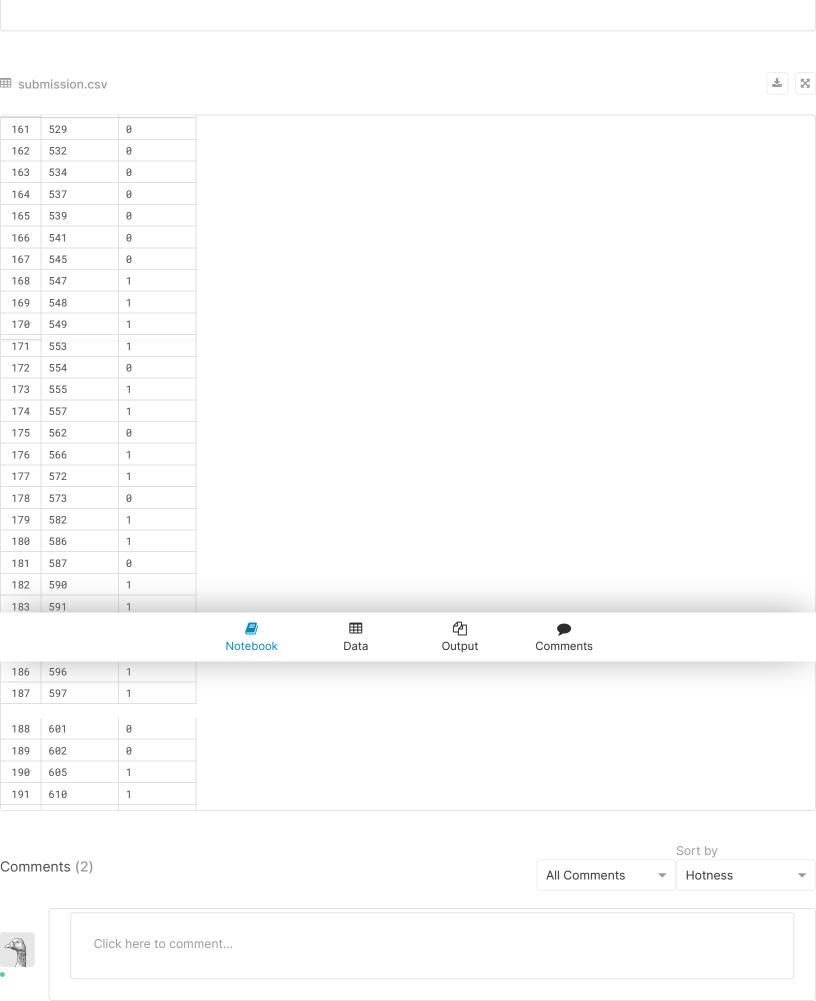
New Dataset

This file was created from a Kernel, it does not have a description.

New Notebook

 Σ

Download All



Uh oh... maybe they will have a new set of data that is much harder to find for the private leaderboard? Great job finding this out @szelee

2 hours ago

This Comment was deleted.

© 2019 Kaggle Inc

Our Team Terms Privacy Contact/Support





