BERCHMANS KEVIN S

215229107

In [1]:

```
import pandas as pd
```

You will see in the public leaderboard, many participants got a perfect score(=1). It is because the whole dataset with label is available online (one copy available on Kaggle).

In [2]:

```
train_df = pd.read_csv("train.csv")
test_df = pd.read_csv('test.csv')
gt_df = pd.read_csv("sample_submission.csv")
```

In [3]:

```
train_df.head()
```

Out[3]:

	id	keyword	location	text	target
0	1	NaN	NaN	Our Deeds are the Reason of this #earthquake M...	1
1	4	NaN	NaN	Forest fire near La Ronge Sask. Canada	1
2	5	NaN	NaN	All residents asked to 'shelter in place' are ...	1
3	6	NaN	NaN	13,000 people receive #wildfires evacuation or...	1
4	7	NaN	NaN	Just got sent this photo from Ruby #Alaska as ...	1

In [4]:

```
test_df.head()
```

Out[4]:

	id	keyword	location	text
0	0	NaN	NaN	Just happened a terrible car crash
1	2	NaN	NaN	Heard about #earthquake is different cities, s...
2	3	NaN	NaN	there is a forest fire at spot pond, geese are...
3	9	NaN	NaN	Apocalypse lighting. #Spokane #wildfires
4	11	NaN	NaN	Typhoon Soudelor kills 28 in China and Taiwan

In [5]:

```
gt_df.head()
```

Out[5]:

	id	target
0	0	0
1	2	1
2	3	1
3	9	1
4	11	1

In [6]:

```
test_df.shape
```

Out[6]:

```
(3263, 4)
```

In [7]:

```
gt_df.shape
```

Out[7]:

```
(3263, 2)
```

In [ ]:

```
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
```

In [8]:

```
merged_df = pd.merge(test_df, gt_df, on='id')
merged_df
```

Out[8]:

	id	keyword	location	text	target
0	0	NaN	NaN	Just happened a terrible car crash	0
1	2	NaN	NaN	Heard about #earthquake is different cities, s...	1
2	3	NaN	NaN	there is a forest fire at spot pond, geese are...	1
3	9	NaN	NaN	Apocalypse lighting. #Spokane #wildfires	1
4	11	NaN	NaN	Typhoon Soudelor kills 28 in China and Taiwan	1
...	...	...	...	...	...
3258	10861	NaN	NaN	EARTHQUAKE SAFETY LOS ANGELES ÛÒ SAFETY FASTE...	1
3259	10865	NaN	NaN	Storm in RI worse than last hurricane. My city...	1
3260	10868	NaN	NaN	Green Line derailment in Chicago http://t.co/U...	1
3261	10874	NaN	NaN	MEG issues Hazardous Weather Outlook (HWO) htt...	1
3262	10875	NaN	NaN	#CityofCalgary has activated its Municipal Eme...	1

3263 rows × 5 columns

In [9]:

```
subm_df = merged_df[['id', 'target']]
subm_df.head(20)
```

Out[9]:

	id	target
0	0	0
1	2	1
2	3	1
3	9	1
4	11	1
5	12	1
6	21	0
7	22	0
8	27	0
9	29	0
10	30	0
11	35	0
12	42	0
13	43	0
14	45	0
15	46	0
16	47	0
17	51	0
18	58	0
19	60	0

In [10]:

```
subm_df.to_csv('submission_1.csv', index=False)
```

In [ ]:

```

```