**Steps to take beforehand**[**¶**](#gjdgxs)

1. Unzip reddit2012\_2019.zip into directly
2. Install gpt-2-simple for finetuning gpt. pip3 install gpt-2-simple
3. Make sure to install Tensorflow 1. It doesn't work on tensorflow 2.

**concat all dataframes**[**¶**](#30j0zll)

In [1]:

**import** **pandas** **as** **pd**

In [20]:

!ls

10 run3\_results run\_no\_clean\_ccb  
2012\_2016.csv run3\_walao run\_no\_clean\_china  
2017.csv run\_long\_aiya run\_no\_clean\_gay  
2018.csv run\_long\_bt4k run\_no\_clean\_knn  
2018\_clean.csv run\_long\_ccb run\_no\_clean\_lah  
2019.csv run\_long\_china run\_no\_clean\_pap  
GPT - 2 model training.ipynb run\_long\_copypaste run\_no\_clean\_pmd  
checkpoint run\_long\_gay run\_no\_clean\_singlish  
csv\_encoded.txt run\_long\_james run\_no\_clean\_smlj  
gen run\_long\_knn run\_no\_clean\_walao  
gpt2\_gentext\_20191113\_133928.txt run\_long\_lah run\_no\_clean\_yishun  
main.py run\_long\_pap samples  
models run\_long\_pmd schedule.sh  
reddit2012\_2019.zip run\_long\_singapore schedule2.sh  
run3\_ccb run\_long\_singlish schedule\_long.sh  
run3\_knn run\_long\_smlj text\_encoded.npz  
run3\_ns run\_long\_walao train\_df.csv  
run3\_nus run\_long\_yishun try2.csv  
run3\_pap run\_no\_clean try3.csv  
run3\_r run\_no\_clean\_aiya

In [3]:

df = pd.read\_csv("2012\_2016.csv")

/ldap\_home/james.chua/miniconda3/envs/tf2/lib/python3.7/site-packages/IPython/core/interactiveshell.py:3049: DtypeWarning: Columns (0,1,2,3) have mixed types. Specify dtype option on import or set low\_memory=False.  
 interactivity=interactivity, compiler=compiler, result=result)

In [4]:

df2=pd.read\_csv("2017.csv")  
df3=pd.read\_csv("2018.csv")  
df4=pd.read\_csv("2019.csv")

/ldap\_home/james.chua/miniconda3/envs/tf2/lib/python3.7/site-packages/IPython/core/interactiveshell.py:3049: DtypeWarning: Columns (2) have mixed types. Specify dtype option on import or set low\_memory=False.  
 interactivity=interactivity, compiler=compiler, result=result)

In [5]:

all\_df = pd.concat([df,df2,df3,df4]).reset\_index()

In [7]:

all\_df.body[1]

Out[7]:

"Don't pretend that's not a thing you do in Singapore... I've seen 2 photographic examples already."

In [8]:

len(all\_df)

Out[8]:

3515857

In [9]:

*#remove [removed]*  
all\_df = all\_df[all\_df.body != "[removed]"].reset\_index()

In [10]:

all\_df = all\_df.dropna()

In [11]:

len(all\_df)

Out[11]:

2992653

**Shuffle and replace some encoding errors**[**¶**](#1fob9te)

In [12]:

all\_df = all\_df.sample(frac=1).reset\_index(drop=**True**)

In [13]:

all\_df

Out[13]:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **level\_0** | **index** | **author** | **body** | **created\_utc** | **date** |
| 0 | 1426256 | 174318 | reddumpling | Use English to converse all the time (except e... | 1493474831 | 2017-04-29 19:00:00 |
| 1 | 2002658 | 119191 | Neralo | > These can only be done on headphones, not... | 1519722453 | 2018-02-27 17:00:00 |
| 2 | 1268789 | 16851 | asilentfilm9 | Wait one more day. | 1484299223 | 2017-01-13 06:00:00 |
| 3 | 2130345 | 246878 | sraelgaiznaer | What’s with the 3D version? Isn’t it good enou... | 1523704828 | 2018-04-14 19:00:00 |
| 4 | 1234346 | 1234346 | shqippotato | dont u know losing internet connectivity can l... | 1480854504 | 2016-12-04 09:00:00 |
| ... | ... | ... | ... | ... | ... | ... |
| 2992648 | 3136284 | 393654 | WalfAkaiTsuki | Yesterday my wallet dropped out of my pocket a... | 1560216032 | 2019-06-11 06:00:00 |
| 2992649 | 2835463 | 92833 | dontknowla | Audio mixing you need to be a chio bu then con... | 1549463366 | 2019-02-06 22:00:00 |
| 2992650 | 3237747 | 495117 | cmonplsnotnow | I’m just wondering how people love that stuff ... | 1563604484 | 2019-07-20 14:00:00 |
| 2992651 | 1506995 | 255057 | Hurt\_cow | thanks | 1498098367 | 2017-06-22 06:00:00 |
| 2992652 | 2339403 | 455936 | pjepej\_ | Depends on context and what kind of message? | 1531293363 | 2018-07-11 15:00:00 |

2992653 rows × 6 columns

In [18]:

*# Save to csv called train\_df*

In [19]:

all\_df[(all\_df.body.str.len() > 100)].to\_csv("train\_df.csv",index=**False**)

**Run the finetuning**[**¶**](#3znysh7)

You can change the model\_name to "124M" for faster training and less memory requirements but lower quality results. We required 50gb of RAM for the 355M model

In [24]:

**import** **gpt\_2\_simple** **as** **gpt2**  
**import** **os**  
**import** **requests**  
  
model\_name = "355M"  
**if** **not** os.path.isdir(os.path.join("models", model\_name)):  
 print(f"Downloading **{model\_name}** model...")  
 gpt2.download\_gpt2(model\_name=model\_name) *# model is saved into current directory under /models/355M/*  
  
sess = gpt2.start\_tf\_sess()  
csv\_path = "train\_df.csv"  
gpt2.encode\_csv(csv\_path=csv\_path, out\_path="csv\_encoded.txt")  
gpt2.encode\_dataset(file\_path="csv\_encoded.txt", out\_path="text\_encoded.npz")  
gpt2.finetune(sess,  
 dataset="text\_encoded.npz",  
 model\_name=model\_name,  
 steps=8000,  
 run\_name='run\_long',  
 print\_every=10,  
 sample\_every=200,  
 save\_every=500,  
 overwrite=**True**  
 )

**Run inference**[**¶**](#2et92p0)

gpt\_2\_simple provides a command line interface Some examples:

!gpt\_2\_simple generate --temperature 0.8 --nsamples 80 --batch\_size 8 --prefix "In NUS, a group of students took the module BT4222, Mining Web Data for Business Insights, under Assistant Professor Qiao Dandan. The students had to create a project using Natural Language Processing tools. One of her students, James, did not know what to do for his project and so" --truncate "<|endoftext|>" --nfiles 5 --folder run\_long\_james --run\_name run\_long

!gpt\_2\_simple generate --temperature 0.8 --nsamples 80 --batch\_size 8 --prefix "SINGAPORE" --truncate "<|endoftext|>" --nfiles 5 --folder run\_long\_singapore --run\_name run\_long

!gpt\_2\_simple generate --temperature 0.8 --nsamples 80 --batch\_size 8 --prefix "PAP" --truncate "<|endoftext|>" --nfiles 5 --folder run\_long\_pap --run\_name run\_long