The pseudo dataset containing the cosine similarity difference and the tweet sentiment predictions are in the form of a list.

Data format : [index,|cosine similarity difference|, predicted label, {neg_coefficient,pos_coefficient,true_label,'cos_sim_bad','cos_sim_good'}]

In order to easily work with the data and derive insights on confidence, we export it as a csv file.

This notebook explains the data manipulation done on this list to get a final csv file.

```
print(len(pseudo_dataset))
```

```
386000
```

```
type(pseudo_dataset)
```

```
list
```

```
for data in pseudo_dataset[:10]:
  print(data)
```

```
[0, 0.027175323124258965, 1, "Unfortunately, the frustration of being Dr. Goldberg's pat
[4, 0.02363800632857238, 1, 'All the food is great here. But the best thing they have is
[8, 0.0231286657595563, 1, "Before I finally made it over to this range I heard the same
[12, 0.02436210483020207, 1, "After a morning of Thrift Store hunting, a friend and I we
[16, 0.021526030194333523, 1, 'Used to go there for tires, brakes, etc.  Their prices ha
[20, 0.02402625126055724, 1, "Don't waste your time.  We had two different people come t
[24, 0.025843451733099987, 1, "Two meals, on the recommendation of a friend who lives ne
[28, 0.023052479524365288, 1, "The biggest breakfast in Pittsburgh, as far as I can tell
[32, 0.024813901753635625, 1, "Classic breakfast joint.  Grimy looking hole in the wall
[36, 0.024260021998294024, 1, 'Great little place. Treats you like a local.Eaten here 3
```

```
data
```

```
[36,
 0.024260021998294024,
 1,
 'Great little place. Treats you like a local.Eaten here 3 times a week for a month.
Same overtime. Barb is always here.',
 {'neg_coefficient': 0.5,
  'pos_coefficient': 0.5,
  'cos_sim_bad': 0.9449997755480609,
  'cos_sim_good': 0.9692597975463549,
  'true_label': 1}]
```

```
def predicted(a):
  return a[2]
```

```
pred_list = map(predicted,pseudo_dataset)
```

```
pred_list = list(pred_list)
type(pred_list)
```

```
list
```

```
print(pred_list)
```

```
[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
```

```
def csd(a):
  return a[1]

csd_list = map(csd,pseudo_dataset)
```

```
csd_list = list(csd_list)
```

```
def true_label(a):
  pseudo_dict = a[4]
  return pseudo_dict['true_label']

truth_list = map(true_label,pseudo_dataset)
```

```
truth_list = list(truth_list)
print(truth_list)
```

```
[0, 1, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0,
```

```
len(pred_list)
len(truth_list)
```

```
386000
```

Turning these lists into data frame columns

```
#calling DataFrame constructor after zipping both lists, with columns specified

plot_data = pd.DataFrame(list(zip(pred_list, truth_list, csd_list)),
               columns=['predictions', 'truth', 'cosine_similarity_difference'])
plot_data.head()
```

```
plot_data.head()
```

	predictions	truth	csd_list
0	1	0	0.027175
1	1	1	0.023638
2	1	1	0.023129
3	1	1	0.024362
4	1	0	0.021526

```
plot_data['correct_guess'] = np.where(plot_data['predictions'] == plot_data['truth'], 1, 0)
plot_data.head()
```

	predictions	truth	csd_list	correct_guess
0	1	0	0.027175	0
1	1	1	0.023638	1
2	1	1	0.023129	1
3	1	1	0.024362	1
4	1	0	0.021526	0

```
plot_data.shape
```

```
(386000, 4)
```

Some data manipulation to get a desired dataframe

```
# add an empty columns
plot_data['null1'] = ' '
plot_data['null2'] = ' '
plot_data['null3'] = ' '
```

```
plot_data.head()
```

	predictions	truth	csd_list	correct_guess	null1	null2	null3

```
plot_data = plot_data[['predictions', 'null1', 'truth', 'null2', 'csd_list', 'null3', 'correc
```

1	1		1	0.023638		1	

```
plot_data.head()
```

	predictions	null1	truth	null2	csd_list	null3	correct_guess
0	1		0		0.027175		0
1	1		1		0.023638		1
2	1		1		0.023129		1
3	1		1		0.024362		1
4	1		0		0.021526		0

```
plot_data.to_csv('csd_cg_dataset_normalised.csv', sep='\t', encoding='utf-8', index=False)
```

Further data cleaning of newly-created csv file in Microsoft Excel by following these steps:

https://www.extendoffice.com/documents/excel/1786-excel-split-text-by-space.html