This notebook demonstrates how to generate pseudo OOD samples (§ 3.3).

In [26]:
```
import nlpaug.augmenter.word as naw
from nltk.corpus import stopwords
from datasets import load_dataset
import pandas as pd
import warnings
import random
import torch

warnings.filterwarnings('ignore')
pd.set_option('display.max_colwidth', None)
pd.set_option('display.max_rows', None)
random.seed(24)
```

VERSION:

- nltk 3.5
- pandas 1.3.4
- datasets 1.18.4
- torch 1.9.0+cu111

Helper Functions

In [41]:
```
def load_sst2():
    datasets = load_dataset('glue', 'sst2', cache_dir="./cache/sst2/")
    training_set = datasets['train']
    dev_set = datasets['validation']
    test_set = datasets['test']
    return training_set, dev_set, test_set
```

```
def highlight(x):
    # To highlight some illustrative pseudo OOD examples with their corresponding ID samples.
    if (x.Row in [0, 3, 9, 10, 11, 15, 19, ]):
        return ['', 'background-color: lightsteelblue', 'background-color: bisque']
    else:
        return ['']*3
```

Load SST2 Dataset

In [28]:
```
training_set, dev_set, test_set = load_sst2()
```

```
Reusing dataset glue (./cache/sst2/glue/sst2/1.0.0/dacbe3125aa31d7f70367a07a8a9e72a5a0bfeb5fc42e75c9db75b96da6053ad)
```

Generate Pseudo OOD Samples

In [29]:
```
device = torch.device('cuda:0' if torch.cuda.is_available() else 'cpu')

# Please refer to https://github.com/makcedward/nlpaug for the detailed documentation.
generator = naw.ContextualWordEmbsAug(model_path='distilbert-base-uncased', # We use DistilBERT as the generator.
                                      action='substitute',
                                      aug_p=0.7, # We set the replacement ratio to 0.7.
                                      top_k=100, # The candidate size is set to 100.
                                      stopwords=stopwords.words('english'), # We do not substitute stopwords.
                                      device=str(device),
                                      )

# We randomly sample 20 examples from the development set for demonstration purposes.
indices = list(range(dev_set.num_rows))
random.shuffle(indices)
ID_samples = [dev_set['sentence'][i] for i in indices[0:20]]

pseudo_oods = []
for id_sample in ID_samples:
    pseudo_oods.append(generator.augment(id_sample))
```

ID Samples vs. Pseudo OOD Examples

In [42]:
```
id_ood_df = pd.DataFrame()
id_ood_df['ID Examples'] = ID_samples
id_ood_df['Pseudo OODs'] = pseudo_oods
id_ood_df.insert(loc=0, column='Row', value=list(range(0, 20)))
id_ood_df.style.apply(highlight, axis=1).set_properties(**{'text-align': 'left'})
```

Out[42]:

	Row	ID Examples	Pseudo OODs
0	0	the talented and clever robert rodriguez perhaps put a little too much heart into his first film and did n't reserve enough for his second .	the talented and courageous robert blake thus incorporated a little too big effort into his final match and did don't reserve enough for his popularity.
1	1	his comedy premises are often hackneyed or just plain crude , calculated to provoke shocked laughter , without following up on a deeper level .	his broadway performances are invariably witty or just deliberately crude, calculated to inspire shocked audiences, without catching up on a satirical tone.
2	2	in exactly 89 minutes , most of which passed as slowly as if i 'd been sitting naked on an igloo , formula 51 sank from quirky to jerky to utter turkey .	in precisely fifteen minutes, most of which moved as slowly as if i'd been lying naked on an igloo, the 51 transitioned from neutral to distant to complete disbelief.
3	3	teen movies have really hit the skids .	twitter profiles have strongly intrigued the fan.
4	4	... a boring parade of talking heads and technical gibberish that will do little to advance the linux cause .	... a scientific search of african skeletons and ancient artefacts that will do something to modify the global environment.
5	5	atom egoyan has conjured up a multilayered work that tackles any number of fascinating issues	the magazine has brought up a fictitious world that reveals any kinds of terrible things
6	6	not an objectionable or dull film ; it merely lacks everything except good intentions .	not an exaggerated or offensive notion ; it rarely proved adequate save basic humour.
7	7	they should have called it gutterball .	they should have gone it hot.
8	8	its well of thorn and vinegar (and simple humanity) has long been plundered by similar works featuring the insight and punch this picture so conspicuously lacks .	its image of thorn and berries (and its outline) has long been replaced by comic artworks featuring the snake and punch this beast so conspicuously elusive.
9	9	a compelling spanish film about the withering effects of jealousy in the life of a young monarch whose sexual passion for her husband becomes an obsession .	a biographical drama film about the withering tide of repression in the childhood of a young prince whose sexual pursuit for her throne proves an obstacle.
10	10	a science-fiction pastiche so lacking in originality that if you stripped away its inspirations there would be precious little left .	a drama - thriller thriller so lacking in adventure that if you left them its glory there would be nothing gems else.
11	11	belongs to daniel day-lewis as much as it belongs to martin scorsese ; it 's a memorable performance in a big , brassy , disturbing , unusual and highly successful film .	homage to daniel russell - johnston as much as it owes to martin freeman ; it's a memorable musical in a big, shocking, disturbing, unusual and unexpected psychedelic direction.
12	12	the story and structure are well-honed .	the tone and character are horizontally - connected.
13	13	there 's not enough here to justify the almost two hours .	there's not many here to celebrate the possibly endless weeks.
14	14	it has its moments of swaggering camaraderie , but more often just feels generic , derivative and done to death .	it has its essence of surreal excitement, but more so just slightly dull, dull and silly to entertain.
15	15	and that 's a big part of why we go to the movies .	and that's a crucial piece of why we got to the aquarium.
16	16	instead , he shows them the respect they are due .	moreover, he provides them the money they are earning.
17	17	try as i may , i ca n't think of a single good reason to see this movie , even though everyone in my group extemporaneously shouted , ` thank you ! '	and as i remember, i ca didn't dispose of a single sane person to suggest this happen, least though everyone in my group extemporaneously shouted, ` shoot you! '
18	18	an operatic , sprawling picture that 's entertainingly acted , magnificently shot and gripping enough to sustain most of its 170-minute length .	an ambitious, gripping story that's deeply stylized, was shot and imaginative enough to sustain most of its 360 - plus presentations.
19	19	the fly-on-the-wall method used to document rural french school life is a refreshing departure from the now more prevalent technique of the docu-makers being a visible part of their work .	the knock - on - the - stump programme attempting to document rural english school education is a refreshing departure from the now more prevalent technique of the mini - etudes being a distinctive aspect of their work.

Background shifts:

- Row 0: film → match
- Row 3: teen movies → twitter profiles
- Row 9: spanish film → biographical drama
- Row 10: science-fiction → thriller
- Row 11: film → musical
- Row 15: movies → aquarium
- Row 19: french school life → english shcool education

In [ ]:
```

```