In [84]:
```
from fastai.text import *
import html
```

In [106]:
```
data_path1 = Path("/home/paperspace/data/training.1600000.processed.noemoticon.csv")
data_path2 = Path("/home/paperspace/data/testdata.manual.2009.06.14.csv")
lm_path_tt = Path("/home/paperspace/data/twitter/lm")
```

In [67]:
```
twitter_lm_path = Path("/home/paperspace/data/twitter_lm")
twitter_lm_path.mkdir(exist_ok=True)
```

In [12]:
```
data = pd.read_csv(data_path, encoding = "latin1",header=None)
```

In [22]:
```
data2 = pd.read_csv(data_path2, encoding = "latin1",header=None)
```

data.head() len(data)

In [15]:
```
data_txt = data.iloc[:,[5]]
```

In [23]:
```
data2_txt = data2.iloc[:,[5]]
```

In [24]:
```
data2_txt.head(5)
```

Out[24]:

	5
0	@stellargirl I loooooooovvvvvveee my Kindle2. ...
1	Reading my kindle2... Love it... Lee childs i...
2	Ok, first assesment of the #kindle2 ...it fuck...
3	@kenburbary You'll love your Kindle2. I've had...
4	@mikefish Fair enough. But i have the Kindle2...

In [17]:
```
data_txt.head(5)
```

Out[17]:

	5
0	@switchfoot http://twitpic.com/2y1zl - Awww, t...
1	is upset that he can't update his Facebook by ...
2	@Kenichan I dived many times for the ball. Man...
3	my whole body feels itchy and like its on fire
4	@nationwideclass no, it's not behaving at all....

In [51]:

```
data_txt;
```

Out[51]:

	5
0	@switchfoot http://twitpic.com/2y1zl - Awww, t...
1	is upset that he can't update his Facebook by ...
2	@Kenichan I dived many times for the ball. Man...
3	my whole body feels itchy and like its on fire
4	@nationwideclass no, it's not behaving at all....
5	@Kwesidei not the whole crew
6	Need a hug
7	@LOLTrish hey long time no see! Yes.. Rains a...
8	@Tatiana_K nope they didn't have it
9	@twittera que me muera ?
10	spring break in plain city... it's snowing
11	I just re-pierced my ears
12	@caregiving I couldn't bear to watch it. And ...
13	@octolinz16 It it counts, idk why I did either...
14	@smarrison i would've been the first, but i di...
15	@iamjazzyfizzle I wish I got to watch it with ...
16	Hollis' death scene will hurt me severely to w...
17	about to file taxes
18	@LettyA ahh ive always wanted to see rent lov...
19	@FakerPattyPattz Oh dear. Were you drinking ou...
20	@alydesigns i was out most of the day so didn'...
21	one of my friend called me, and asked to meet ...
22	@angry_barista I baked you a cake but I ated it
23	this week is not going as i had hoped
24	blagh class at 8 tomorrow
25	I hate when I have to call and wake people up
26	Just going to cry myself to sleep after watchi...
27	im sad now Miss.Lilly
28	ooooh.... LOL that leslie.... and ok I won't ...
29	Meh... Almost Lover is the exception... this t...
...	...
1599970	Thanks @eastwestchic & @wangyip Thanks! Th...
1599971	@marttn thanks Martin. not the most imaginativ...
1599972	@MikeJonesPhoto Congrats Mike Way to go!
1599973	http://twitpic.com/7jp4n - OMG! Office Space.....

	5
1599974	@yrclndstnlvr ahaha nooo you were just away fr...
1599975	@BizCoachDeb Hey, I'm baack! And, thanks so m...
1599976	@mattycus Yeah, my conscience would be clear i...
1599977	@MayorDorisWolfe Thats my girl - dishing out t...
1599978	@shebbs123 i second that
1599979	In the garden
1599980	@myheartandmind jo jen by nemuselo zrovna tÃ© ...
1599981	Another Commenting Contest! [;: Yay!!! http:/...
1599982	@thrillmesoon i figured out how to see my twee...
1599983	@oxhot theri tomorrow, drinking coffee, talkin...
1599984	You heard it here first -- We're having a girl...
1599985	if ur the lead singer in a band, beware fallin...
1599986	@tarayqueen too much ads on my blog.
1599987	@La_r_a NEVEER I think that you both will get...
1599988	@Roy_Everitt ha- good job. that's right - we g...
1599989	@Ms_Hip_Hop im glad ur doing well
1599990	WOOOOO! Xbox is back
1599991	@rmedina @LaTati Mmmm That sounds absolutely ...
1599992	ReCoVeRiNg FrOm ThE lOnG wEeKeNd
1599993	@SCOOBY_GRITBOYS
1599994	@Cliff_Forster Yeah, that does work better tha...
1599995	Just woke up. Having no school is the best fee...
1599996	TheWDB.com - Very cool to hear old Walt interv...
1599997	Are you ready for your MoJo Makeover? Ask me f...
1599998	Happy 38th Birthday to my boo of alll time!!! ...
1599999	happy #charitytuesday @theNSPCC @SparksCharity...

1600000 rows × 1 columns

In [61]:
```
trn_lm_text, test_lm_text = sklearn.model_selection.train_test_split(np.concat
enate([data_txt.iloc[:,0], data2_txt.iloc[:,0]]), test_size = 0.1)
```

In [62]:
```
print(len(trn_lm_text));
print(len(test_lm_text))
print(trn_lm_text[1:5])
```

```
1440448
160050
['is sad to see Andi go  Come Monday, interns are on our own... check out Gad
get Deals of the Day for a screwup!!'
 "Everything that exists today: despite philosophical rational/personal reduc
tion:  are all scaffold (pro or con) on Womb's Gravitation "
 '@danielooi nice avatar!!! Veli the hensem! ' 'Were going to go get snowcone
s! ']
```

In [1]:
```
#[len(i) for i in trn_lm_text]
```

In [63]:
```
col_names = ["labels","text"]
```

In [68]:
```
lm_trn = pd.DataFrame({"text": trn_lm_text, "labels": [0] * len(trn_lm_text)},
 columns=col_names)
lm_test = pd.DataFrame({"text": test_lm_text, "labels": [0] * len(test_lm_text
)}, columns=col_names)

lm_trn.to_csv(twitter_lm_path/"train.csv", header = False, index = False, enco
ding = "utf-8")
lm_test.to_csv(twitter_lm_path/"test.csv", header = False, index = False, enco
ding = "utf-8")
```

In [78]:
```
## functions pulled from the fast.ai notebook for text tokenization

re1 = re.compile(r'  +')


def fixup(x):
    x = x.replace('#39;', "'").replace('amp;', '&').replace('#146;', "'").repl
ace(
        'nbsp;', ' ').replace('#36;', '$').replace('\\n', "\n").replace('quo
t;', "'").replace(
        '<br />', "\n").replace('\\"', '"').replace('<unk>','u_n').replace('
 @.@ ','.').replace(
        ' @-@ ','-').replace('\\', ' \\ ')
    return re1.sub(' ', html.unescape(x))


def get_texts(df, n_lbls=1):
    labels = df.iloc[:,range(n_lbls)].values.astype(np.int64)
    texts = f'\n{BOS} {FLD} 1 ' + df[n_lbls].astype(str)
    for i in range(n_lbls+1, len(df.columns)): texts += f' {FLD} {i-n_lbls} '
+ df[i].astype(str)
    texts = list(texts.apply(fixup).values)

    tok = Tokenizer().proc_all_mp(partition_by_cores(texts))
    return tok, list(labels)


def get_all(df, n_lbls):
    tok, labels = [], []
    for i, r in enumerate(df):
        print(i)
        tok_, labels_ = get_texts(r, n_lbls)
        tok += tok_;
        labels += labels_
    return tok, labels
```

In [79]:
```
BOS = "xboxs"
FLD = "xfld"
```

In [80]:
```
chunksize = 24000
```

In [81]:
```
twitter_lm_train_dl = pd.read_csv(twitter_lm_path/"train.csv", header=None, ch
unksize=chunksize)
twitter_lm_test_dl = pd.read_csv(twitter_lm_path/"test.csv", header=None, chun
ksize=chunksize)
```

In [2]:
```
tok_trn_twitterlm, trn_labels_twittterlm = get_all(twitter_lm_train_dl, 1)
tok_test_twitterlm, test_labels_twittterlm = get_all(twitter_lm_test_dl, 1)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-2-abac1af77d38> in <module>
----> 1 tok_trn_twitterlm, trn_labels_twittterlm = get_all(twitter_lm_train_d
l, 1)
      2 tok_test_twitterlm, test_labels_twittterlm = get_all(twitter_lm_test_
dl, 1)

NameError: name 'get_all' is not defined
```

In [86]:
```
(twitter_lm_path/"tmp").mkdir(exist_ok = True)
np.save(twitter_lm_path/"tmp/tok_trn_twitter.npy", tok_trn_twitterlm)
np.save(twitter_lm_path/"tmp/tok_test_twitter.npy", tok_test_twitterlm)
tok_trn_tt = np.load(twitter_lm_path/"tmp/tok_trn_twitter.npy")
tok_test_tt = np.load(twitter_lm_path/"tmp/tok_test_twitter.npy")
```

In [88]:
```
freq = Counter(p for o in tok_trn_tt for p in o)
freq.most_common(25)
```

Out[88]:
```
[('1', 1403845),
 ('\n', 1392451),
 ('xboxs', 1392449),
 ('xfld', 1392448),
 ('i', 874788),
 ('.', 755229),
 ('!', 698523),
 ('to', 492253),
 ('the', 455330),
 (',', 420699),
 ('a', 343292),
 ('t_up', 295495),
 ('my', 275133),
 ('it', 264137),
 ('and', 263858),
 ('you', 262710),
 ('/', 228098),
 ('is', 214318),
 ('...', 191455),
 ('?', 188774),
 ('in', 188313),
 ('for', 188206),
 ('of', 159853),
 ("'s", 156795),
 ('that', 151947)]
```

In [89]:
```
max_vocab = 60000
min_freq = 2
```

In [91]:
```
#itos = [o for o,c in freq.most_common(max_vocab) if c > min_freq]
itos_tt = [o for o,c in freq.most_common(max_vocab)]

len(itos_tt)
```

Out[91]:
```
60000
```

In [92]:
```
itos_tt.insert(0, "_pad_")
itos_tt.insert(1, "_unk_")
```

In [93]:
```
stoi_tt = collections.defaultdict(lambda:0, {o:c for c,o in enumerate(itos_tt
)})
len(stoi_tt)
```

Out[93]:
```
60002
```

In [96]:
```
trn_lm_tt = np.array([[stoi_tt[o] for o in p] for p in tok_trn_tt])
test_lm_tt = np.array([[stoi_tt[o] for o in p] for p in tok_test_tt])
```

In [97]:
```
np.save(twitter_lm_path/"tmp/trn_ids.npy", trn_lm_tt)
np.save(twitter_lm_path/"tmp/test_ids.npy", test_lm_tt)
```

In [98]:
```
pickle.dump(itos_tt, open(twitter_lm_path/"tmp/itos.pkl", "wb"))
```

In [100]:
```
trn_lm_tt = np.load(twitter_lm_path/"tmp/trn_ids.npy")
test_lm_tt = np.load(twitter_lm_path/"tmp/test_ids.npy")
itos_tt = pickle.load(open(twitter_lm_path/"tmp/itos.pkl", "rb"))
```

In [102]:
```
vs=len(itos_tt)
vs,len(trn_lm_tt)
```

Out[102]: (60002, 1392448)

Language model

In [104]:
```
wd=1e-7
bptt=70
bs=52
opt_fn = partial(optim.Adam, betas=(0.8, 0.99))
```

In [111]:
```
em_sz, nh, nl = 400, 1150, 3
```

In [107]:
```
trn_dl_tt = LanguageModelLoader(np.concatenate(trn_lm_tt), bs, bptt)
test_dl_tt = LanguageModelLoader(np.concatenate(test_lm_tt), bs, bptt)
md_tt = LanguageModelData(lm_path_tt, 1, vs, trn_dl_tt, test_dl_tt, bs=bs, bpt
t=bptt)
```

In [108]:
```
drops = np.array([0.25, 0.1, 0.2, 0.02, 0.15])*0.7
```

In [112]:
```
learner= md_tt.get_model(opt_fn, em_sz, nh, nl,
    dropouti=drops[0], dropout=drops[1], wdrop=drops[2], dropoute=drops[3], dr
opouth=drops[4])

learner.metrics = [accuracy]
learner.freeze_to(-1)
```

In [113]:
```
lr=1e-3
lrs = lr
```

In [114]:
```
learner.fit(lrs/2, 1, wds=wd, use_clr=(32,2), cycle_len=1)
```

```
epoch      trn_loss   val_loss   accuracy
    0      5.726489   5.68029    0.18256
```

Out[114]: [array([5.68029]), 0.18256042742239648]

In [122]:
```
learner.fit(lrs, 1, wds=wd, use_clr=(32,2), cycle_len=1)
```

```
epoch      trn_loss   val_loss   accuracy
    0      5.261247   5.170833   0.238439
```

Out[122]: [array([5.17083]), 0.2384387330461157]

In [123]:
```
learner.save('lm_last_tt_ft')
learner.load('lm_last_tt_ft')
learner.unfreeze()
```

In [124]:

```
learner.fit(lrs, 1, wds=wd, use_clr_beta=(20,20, 0.95,0.85), cycle_len=10)
```

```
epoch      trn_loss   val_loss   accuracy
    0      4.151435   4.045434   0.347076
  5%|▌          | 428/7853 [01:19<23:05,  5.36it/s, loss=4.14]
```

```
---------------------------------------------------------------------------
KeyboardInterrupt                         Traceback (most recent call last)
<ipython-input-124-b9b39cee8b0c> in <module>
----> 1 learner.fit(lrs, 1, wds=wd, use_clr_beta=(20,20, 0.95,0.85), cycle_le
n=10)

~/fastai/courses/dl2/fastai/text.py in fit(self, *args, **kwargs)
    209
    210     def _get_crit(self, data): return F.cross_entropy
--> 211     def fit(self, *args, **kwargs): return super().fit(*args, **kwarg
s, seq_first=True)
    212
    213     def save_encoder(self, name): save_model(self.model[0], self.get_
model_path(name))

~/fastai/courses/dl2/fastai/learner.py in fit(self, lrs, n_cycle, wds, **kwar
gs)
    300         self.sched = None
    301         layer_opt = self.get_layer_opt(lrs, wds)
--> 302         return self.fit_gen(self.model, self.data, layer_opt, n_cycle
, **kwargs)
    303
    304     def warm_up(self, lr, wds=None):

~/fastai/courses/dl2/fastai/learner.py in fit_gen(self, model, data, layer_op
t, n_cycle, cycle_len, cycle_mult, cycle_save_name, best_save_name, use_clr,
 use_clr_beta, metrics, callbacks, use_wd_sched, norm_wds, wds_sched_mult, us
e_swa, swa_start, swa_eval_freq, **kwargs)
    247             metrics=metrics, callbacks=callbacks, reg_fn=self.reg_fn,
 clip=self.clip, fp16=self.fp16,
    248             swa_model=self.swa_model if use_swa else None, swa_start=
swa_start,
--> 249             swa_eval_freq=swa_eval_freq, **kwargs)
    250
    251     def get_layer_groups(self): return self.models.get_layer_groups()

~/fastai/courses/dl2/fastai/model.py in fit(model, data, n_epochs, opt, crit,
 metrics, callbacks, stepper, swa_model, swa_start, swa_eval_freq, visualize,
 **kwargs)
    139             batch_num += 1
    140             for cb in callbacks: cb.on_batch_begin()
--> 141             loss = model_stepper.step(V(x),V(y), epoch)
    142             avg_loss = avg_loss * avg_mom + loss * (1-avg_mom)
    143             debias_loss = avg_loss / (1 - avg_mom**batch_num)

~/fastai/courses/dl2/fastai/model.py in step(self, xs, y, epoch)
     55         if self.loss_scale != 1: assert(self.fp16); loss = loss*self.
loss_scale
     56         if self.reg_fn: loss = self.reg_fn(output, xtra, raw_loss)
---> 57         loss.backward()
     58         if self.fp16: update_fp32_grads(self.fp32_params, self.m)
     59         if self.loss_scale != 1:

~/anaconda3/envs/fastai/lib/python3.6/site-packages/torch/autograd/variable.p
y in backward(self, gradient, retain_graph, create_graph, retain_variables)
    165                 Variable.
    166         """
--> 167         torch.autograd.backward(self, gradient, retain_graph, create_
graph, retain_variables)
    168
    169     def register_hook(self, hook):

~/anaconda3/envs/fastai/lib/python3.6/site-packages/torch/autograd/__init__.p
```

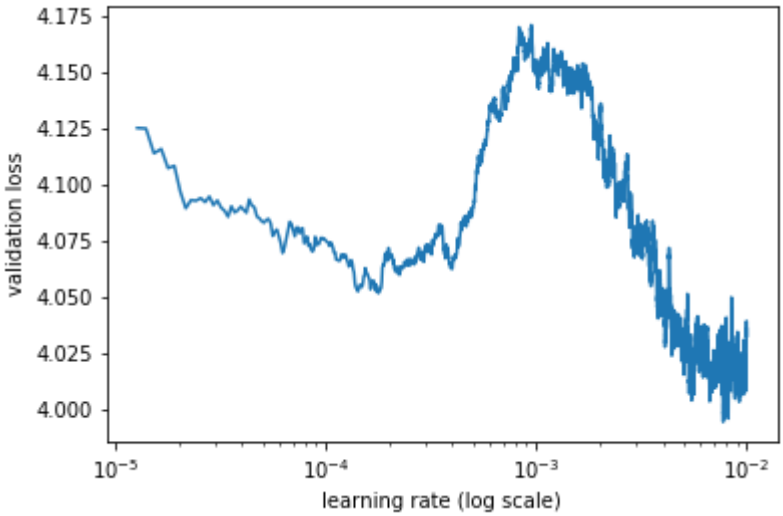
```
y in backward(variables, grad_variables, retain_graph, create_graph, retain_v
ariables)
     97
     98     Variable._execution_engine.run_backward(
---> 99         variables, grad_variables, retain_graph)
    100
    101

KeyboardInterrupt:
```

In [125]:
```
learner.lr_find(start_lr=lrs/10, end_lr=lrs*10, linear=True)
```

```
epoch      trn_loss   val_loss   accuracy
    0      4.034484   3.907411   0.353202
```

In [126]:
```
learner.sched.plot()
```



In [127]:

```
learner.fit(lrs, 1, wds=wd, use_clr_beta=(20,10,0.95,0.85), cycle_len=2)
```

```
epoch      trn_loss   val_loss   accuracy
    0      3.922033   3.822018   0.364298
  1%|          | 89/7853 [00:16<24:18,  5.32it/s, loss=3.91]
```

```
---------------------------------------------------------------------------
KeyboardInterrupt                         Traceback (most recent call last)
<ipython-input-127-a5f6df78919a> in <module>
----> 1 learner.fit(lrs, 1, wds=wd, use_clr_beta=(20,10,0.95,0.85), cycle_len
=2)

~/fastai/courses/dl2/fastai/text.py in fit(self, *args, **kwargs)
    209
    210     def _get_crit(self, data): return F.cross_entropy
--> 211     def fit(self, *args, **kwargs): return super().fit(*args, **kwarg
s, seq_first=True)
    212
    213     def save_encoder(self, name): save_model(self.model[0], self.get_
model_path(name))

~/fastai/courses/dl2/fastai/learner.py in fit(self, lrs, n_cycle, wds, **kwar
gs)
    300         self.sched = None
    301         layer_opt = self.get_layer_opt(lrs, wds)
--> 302         return self.fit_gen(self.model, self.data, layer_opt, n_cycle
, **kwargs)
    303
    304     def warm_up(self, lr, wds=None):

~/fastai/courses/dl2/fastai/learner.py in fit_gen(self, model, data, layer_op
t, n_cycle, cycle_len, cycle_mult, cycle_save_name, best_save_name, use_clr,
 use_clr_beta, metrics, callbacks, use_wd_sched, norm_wds, wds_sched_mult, us
e_swa, swa_start, swa_eval_freq, **kwargs)
    247             metrics=metrics, callbacks=callbacks, reg_fn=self.reg_fn,
 clip=self.clip, fp16=self.fp16,
    248             swa_model=self.swa_model if use_swa else None, swa_start=
swa_start,
--> 249             swa_eval_freq=swa_eval_freq, **kwargs)
    250
    251     def get_layer_groups(self): return self.models.get_layer_groups()

~/fastai/courses/dl2/fastai/model.py in fit(model, data, n_epochs, opt, crit,
 metrics, callbacks, stepper, swa_model, swa_start, swa_eval_freq, visualize,
 **kwargs)
    139             batch_num += 1
    140             for cb in callbacks: cb.on_batch_begin()
--> 141             loss = model_stepper.step(V(x),V(y), epoch)
    142             avg_loss = avg_loss * avg_mom + loss * (1-avg_mom)
    143             debias_loss = avg_loss / (1 - avg_mom**batch_num)

~/fastai/courses/dl2/fastai/model.py in step(self, xs, y, epoch)
     55         if self.loss_scale != 1: assert(self.fp16); loss = loss*self.
loss_scale
     56         if self.reg_fn: loss = self.reg_fn(output, xtra, raw_loss)
---> 57         loss.backward()
     58         if self.fp16: update_fp32_grads(self.fp32_params, self.m)
     59         if self.loss_scale != 1:

~/anaconda3/envs/fastai/lib/python3.6/site-packages/torch/autograd/variable.p
y in backward(self, gradient, retain_graph, create_graph, retain_variables)
    165                 Variable.
    166         """
--> 167         torch.autograd.backward(self, gradient, retain_graph, create_
graph, retain_variables)
    168
    169     def register_hook(self, hook):

~/anaconda3/envs/fastai/lib/python3.6/site-packages/torch/autograd/__init__.p
```

```
y in backward(variables, grad_variables, retain_graph, create_graph, retain_v
ariables)
     97
     98     Variable._execution_engine.run_backward(
---> 99         variables, grad_variables, retain_graph)
    100
    101

KeyboardInterrupt:
```

In [128]:
```
learner.fit(lrs/10, 1, wds=wd, use_clr_beta=(20,20,0.95,0.85), cycle_len=1)
```

```
epoch      trn_loss   val_loss   accuracy
    0      3.762373   3.752207   0.372365
```

Out[128]:
```
[array([3.75221]), 0.3723646476451353]
```

In [129]:
```
learner.save("lm_tt")
```

In [130]:
```
learner.save_encoder("lm_tt_enc")
```

In [131]:
```
learner.sched.plot_loss()
```