text_class_scimed.py (Python file)

```
#!/usr/bin/env python3

# In iPython, use 'run' to run a Python script:
#    In [1]: run text_class_scimed.py
#
# Or use this plain Python command:
#    exec(open('text_class_scimed.py').read())
#
# Text classification using Transformers and DistilBERT
# Adapted from https://towardsdatascience.com/
# text-classification-with-hugging-face-transformers-in-tensorflow-2-
# without-tears-ee50e4f3e7ed
#
import pdb                    # pdb.set_trace() as breakpoint for
debugging
from datetime import datetime
startTime = datetime.now()

# Use text data from Python scikit-learn, easier.  Extract 4 known
categories.
# These are newsgroup postings. The goal is to train BERT to classify
# verbatim free text in the postings to these 4 categories.
categories = ['sci.med', 'rec.sport.baseball', 'comp.sys.mac.hardware',
             'comp.graphics']

from sklearn.datasets import fetch_20newsgroups
train_b = fetch_20newsgroups(subset='train',
   categories=categories, shuffle=True, random_state=42)
test_b = fetch_20newsgroups(subset='test',
   categories=categories, shuffle=True, random_state=42)
print('size of training set: %s' % (len(train_b['data'])))
print('size of validation set: %s' % (len(test_b['data'])))
print('classes: %s' % (train_b.target_names))
x_train = train_b.data
y_train = train_b.target
x_test = test_b.data
y_test = test_b.target

import numpy as np
np.random.seed(23)      # set a random seed so that you get the
                        # identical random sequence every time
idx = np.random.randint(low = 0, high = len(x_train)-1, size = 3)
for i in idx:
    print("*\n*\n*\n********** Here is the verbatim text **********")
    print(x_train[i])
    print("********* Text above is from this newsgroup:")
    print(train_b.target_names[y_train[i]], "\n\n")

pdb.set_trace()         # pause program here

import ktrain           # ktrain helps to train a neural network
from ktrain import text
MODEL_NAME = 'distilbert-base-uncased'
```

```
t = text.Transformer(MODEL_NAME, maxlen=500,
class_names=train_b.target_names)
trn = t.preprocess_train(texts=x_train, y=y_train)
val = t.preprocess_test(texts=x_test, y=y_test)
model = t.get_classifier()  # lots of warnings, ignore for now.
print("----------------------------------------")
print("Starting fitting the training model. This takes a while")
print("----------------------------------------")
learner = ktrain.get_learner(model, train_data=trn, val_data=val,
batch_size=6)
###
learner.fit_onecycle(lr = 5e-5, epochs = 4)
###
# The line above fits a distilbert model with a specific learning rate
# and number of epochs (number of learning cycles).
# lr (learning rate) is a 'hyper parameter'. It affects how a
# neural network adjusts network weights as it learns from data.
# The idea is to tune the learning rate to minimize a loss
# function. Training will be slow or can stall if lr is too low.
# Loss can be high if the learning rate is too high. To find an
# optimal learning rate for your model, one can simulate the training
# by starting with a low learning rate and gradually increasing it.
# This is called 'tuning the learning rate'.  See the paper by
# Leslie Smith (https://arxiv.org/abs/1506.01186)
#
# Ktrain has a learning rate finder to estimate the model's optimal
# learning rate.
#   learner.lr_find()
#   learner.lr_plot()
#
# There are various learning rate schedules such as triangular policy
# and SGDR. Details are beyond the scope here.
# See, e.g., https://towardsdatascience.com/ktrain-a-lightweight-
# wrapper-for-keras-to-help-train-neural-networks-82851ba889c

# examine the newsgroup posting that our model is getting the most wrong
print("----------------------------------------")
print("Print an example of wrong classification")
print("----------------------------------------")
learner.view_top_losses(n = 1, preproc = t)  # takes a while
# newgroup posting associated with ID 1355 in the validation set
# is in the comp.graphics newsgroup (i.e., computer graphics),
# but it is really about color blindness, which is a medical topic.
# Thus, our model's prediction of sci.med for this post is
# not surprising.
print(x_test[0])

import re
pattern = 'interested if anyone knows of any current research that is
going on into the subject'
for idx, val in enumerate(x_train):
     if re.search(pattern, val):
          print('found a match of {} at index {}\n with this text
{}'.format(pattern, idx, val))
```

```
print("----------------------------------------")
print("Print an example of regex match")
print("----------------------------------------")
print(x_train[0])

# Instantiate a Predictor object to make predictions on new examples
predictor = ktrain.get_predictor(learner.model, preproc=t)
pred_text = 'She is hospitalized after complications of a cystectomy
surgery for bladder cancer.'
predictor.predict(pred_text)
# You might run into this warning:
# UserWarning: ktrain requires a forked version of eli5 to support
# tf.keras. Install with
#    pip install git+https://github.com/amaiya/eli5@tfkeras_0_10_1
predictor.explain(pred_text)
# Predictor object can be saved to disk and reloaded later in
# real-world deployment scenarios.
predictor.save('/tmp/my_20newsgroup_predictor')
reloaded_predictor =
ktrain.load_predictor('/tmp/my_20newsgroup_predictor')
reloaded_predictor.predict(pred_text)
reloaded_predictor.predict_proba(pred_text)
reloaded_predictor.get_classes()
# ktrain uses eli5 to render the prediction in HTML.  It won't show up
# on a text terminal. Instead, you see on the text terminal
# <IPython.core.display.HTML object>. Googling it shows that you
# need to save it to an HTML file.  From there you can use a web
# browser to display it, or to print it as PDF to be included in
# a Word file.
print("----------------------------------------")
print("Load \'text_explain.htm\' with a browser to view the prediction.")
print("----------------------------------------")
html_obj = reloaded_predictor.explain(pred_text)
with open('text_explain.htm','wb') as f:
    f.write(html_obj.data.encode("UTF-8"))

# Python 3 timing
print("text classification took")
print(datetime.now() - startTime)
```

text_class_scimed.py (Output file)

```
[Lz try.py]% source transformers/bin/activate
(transformers) [Lz try.py]% python3
Python 3.8.5 (default, May 27 2021, 13:30:53)
[GCC 9.3.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
entering ~/.pystartup
>>> exec(open('text_class_scimed.py').read())
size of training set: 2353
size of validation set: 1567
classes: ['comp.graphics', 'comp.sys.mac.hardware', 'rec.sport.baseball',
'sci.med']
preprocessing train...
language: en
train sequence lengths:
        mean : 233
        95percentile : 583
        99percentile : 1243
Is Multi-Label? False
preprocessing test...
language: en
test sequence lengths:
        mean : 261
        95percentile : 652
        99percentile : 1630
----------------------------------------
Starting fitting the training model. This takes a while
----------------------------------------


begin training using onecycle policy with max lr of 5e-05...
Epoch 1/4
393/393 [==============================] - 1986s 5s/step - loss: 0.5183 -
accuracy: 0.8411 - val_loss: 0.1915 - val_accuracy: 0.9483
Epoch 2/4
393/393 [==============================] - 1981s 5s/step - loss: 0.1304 -
accuracy: 0.9601 - val_loss: 0.1699 - val_accuracy: 0.9432
Epoch 3/4
393/393 [==============================] - 1978s 5s/step - loss: 0.0664 -
accuracy: 0.9775 - val_loss: 0.2269 - val_accuracy: 0.9368
Epoch 4/4
393/393 [==============================] - 1981s 5s/step - loss: 0.0227 -
accuracy: 0.9945 - val_loss: 0.1562 - val_accuracy: 0.9572
----------------------------------------
Print an example of wrong classification
----------------------------------------
----------
id:172 | loss:7.09 | true:sci.med | pred:rec.sport.baseball)

Distribution: world
From: Thomas_n.a._Krebs@mcontent.apana.org.au
Organization: MacContent BBS, Doncaster, Victoria, Australia
Return-Receipt-To: Thomas_n.a._Krebs@mcontent.apana.org.au
Subject: Re: Why the drive speeds differ??
```

Lines: 11

The most likely explanation may have something to do with the fact that a greater density of information exists on the larger capacity disk drive than the smaller one. If your running the drive on a Mac I would recommend a shareware utility called Time drive which tests seek, SCSI throughput and rotational speed. This utility should let you know what the differences are between the drives.
****
**
 The views expressed in this posting those of the individual author only.
[BBS Number:(613) 848-1346   MacContent is VictoriaÕs first Iconic BBS!]
****
**

found a match of interested if anyone knows of any current research that is going on into the subject at index 724
 with this text From: uabdpo.dpo.uab.edu!gila005 (Steve Holland)
Subject: Re: Crohn's Disease
Organization: UAB - Gastroenterology
Lines: 32

In article <1993Apr14.174824.12295@westminster.ac.uk>, kxaec@sun.pcl.ac.uk
(David Watters) wrote:
>
> Dear all,
>
> I am a Crohn's Disease sufferer and I'm interested if anyone knows of any current research that is going on into the subject. I've done some investigation myself so you don't need to spare me any details. I've had the fistulas, the ileostomy, etc..
>
> Is a "cure" on the horizon ?
>
> I am not in the medical profession so if you do reply I would appreciate plain speak.
>
> I'd prefer to be mailed direct as I don't always get a chance to read the news.
>
> Thank you in advance.
>
> Dave.
The best group to keep you informed is the Crohn's and Colitis Foundation of America. I do not know if the UK has a similar organization. The address of
the CCFA is

CCFA
444 Park Avenue South
11th Floor
New York, NY 10016-7374

USA

They have a lot of information available and have a number of newsletters.

Good Luck.

Steve

Print an example of regex match

From: robin@ntmtv.com (Robin Coutellier)
Subject: Critique of Pressure Point Massager
Originator: robin@volans
Nntp-Posting-Host: volans
Reply-To: robin@ntmtv.com (Robin Coutellier)
Organization: Northern Telecom Inc, Mountain View, CA
Distribution: na
Lines: 141

As promised, below is a personal critique of a Pressure Point Massager I recently bought from the Self Care Catalog. I am very pleased with the results. The catalog description is as follows:

> The Pressure Point Massager is an aggressive physical massager that actually kneads the tension out of muscles ... much like a professional shiatsu masseur. The powerful motor drives two counter-rotating "thumbs" that move in one-inch orbits -- releasing tension in the neck, back, legs and arms.
>
> Pressure Point Massager A2623 $109

To order or receive a catalog, call (24 hours, 7 days) 1-800-345-3371 or fax at 1-800-345-4021.

****
NOTE:
When I ordered the massager, the item number was different, and the price was $179, not $109. When I received it, I glanced thru the newer catalog enclosed with it to see anything was different from the first one. I was QUITE annoyed to see a $70 difference in price. I called them about it, and the cust rep said that they had switched manufacturers, although it looks and works exactly the same. He told me to go ahead and return the first one and order the cheaper one, using the price difference as a reason for return. In fact, since the newer ones might take a while to ship from the factory (I received this one in 3 days), he told me I could use the one I already have until the new one arrives, then return the old one. VERY reasonable people.
****

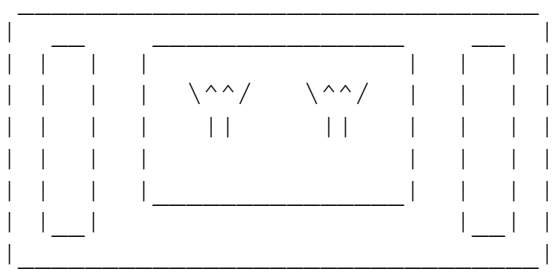
I have long-term neck, shoulder and back pain (if I were a building, I would be described as "structurally unsound :-)). I have stretches and exercises to do that help, but the problem never really goes away.

If, for whatever reason, I do not exercise for a while (illness, not enough
time, lazy, etc.), the muscles become quite stiff and painful and, thus, more prone to further strain. Even with exercise, I sometimes require physical therapy to get back on track, which 1st requires a doctor visit to get the prescription for p.t.

The tension in my neck, if not released, eventually causes a headache (sometimes confused with a sinus headache) over my left eye. When my physical therapist has massaged my neck, and the sub-occipital muscles in particular (the 2 knobby areas near the base of the skull), the headache usually eased within a day, although it hurts like hell to while it is being massaged.

I ordered this device because it seemed to be exactly what I was wishing someone would invent --a machine that would massage, NOT VIBRATE, my neck and sub-occipital muscles like my physical therapist has done in the past, that I could use by myself. No doctor visit or inconvenient p.t. appts for a week later would be needed to use it. I could get up in the middle of the night and use it, if necessary.

I have been using it for about a week or so now, and LOVE it. The base unit is about a 14" x 9" rectangle, about 3-3/4" high, with handles on each side, and it plugs into an average outlet. The two metal "thumbs" are about 1-1/2" in diameter and protrude about 2-1/2" above the base. The thumbs are covered with a gray cloth that is non-removable. They are located more toward one end, rather than centered (see figure below). They move in either clockwise or counter-clockwise directions, depending on which side of the switch is pushed, and are very quiet. It can be used from either side. For instance, the thumbs can be positioned at the base of the neck or the top of the neck, depending on which direction you approach it.



For the neck/head, the user varies the amount of pressure used by (if laying down) allowing all or part of the full weight of the head and/or neck to rest on the thumbs. The handles can also be used if sitting or standing, applying pressure with the arms/wrists. Since my wrists are also impaired (I'm typing this over an extended period of time), and I don't have someone living with me who can apply it, laying down works well for me.

For my back, I sit in a high-backed kitchen chair, position the massager behind me at whatever point I want massaged, and lean back lightly (or not so lightly) against it. The pressure of leaning back holds it in place.
If I want to massage the entire spine, I simply move it down a few inches whenever I feel like it. For my back, this machine is far superior to use than the commonly used "home-made" massager of 2 tennis balls taped together (with the balls, position (against a wall or door) them over the spine and move the body up and down against them). The tennis balls are better than nothing, but difficult to use for very long, especially if your quads are not in good shape, and my long hair gets (painfully) in the way if I don't pin it up first. As far as I'm concerned, the easier something like this is to use, the more likely I'll use/do it. If there are multiple considerations/hassles, I'm more likely to not bother with it.

Not only has this machine helped with my headaches, but my range of motion for my neck and back are greatly increased. The first time I used it on my neck/sub-occipital muscles, however, I overdid it and pressed too hard against it, which resulted in a very tender, almost bruised area for a few days. I laid off it for about 3 days and applied ice, which helped.
After that, I was more gradual about applying pressure. At this point, the pain in the sub-occipital area is now minimal while being massaged. I also learned to use VERY LIGHT pressure on my lower back, which is the most vulnerable point for me.

It also eased some painful knots of tension between my shoulder blades, although, again, it took a few days of massaging (just a few minutes at a time) to really work it out.

I highly recommend this product if you have similar problems, although I cannot vouch for its durability (it seems pretty sturdy), since I've had it such a short time. I plan to use it not only to ease tension, but also to loosen the muscls BEFORE exercising (and maybe after, too). I have been ill recently and not able to exercise much for a few weeks, so this was very timely for me.

This is the 1st product I've ordered from this company and only recently became aware of it thru a co-worker. The catalog states they have been in business since 1976. It contains quite a few health care products and, while they appear to be more expensive than the average health care catalog products, they also appear to be of much higher quality with more thought put into what they actually do. Definitely a step above some other ones
I've seen such as "Dr. Leonards Health Care Catalog" or "Mature Wisdom". I'm only 37, but have ended up on some geriatric-type mailing lists (no big surprise here :-)). I consider many of those products to be rip-offs,
particularly targeted toward the elderly, with dubious health benefits.

I apologize for the length of this, but it's the kind of info _I_ would like to know before ordering something thru the mail.

Robin Coutellier
Northern Telecom, Mountain View, CA
INTERNET: robin@ntmtv.com
UUCP:portal!ntmtv!robin

```
----------------------------------------
Load 'text_explain.htm' with a browser to view the prediction.
----------------------------------------
text classification took
2:36:52.577039
>>>
>>>
>>>
```

y=sci.med (probability **0.998**, score **5.956**) top features

Contribution[?]	Feature
+5.789	Highlighted in text (sum)
+0.167	<BIAS>

she is hospitalized after complications of a cystectomy surgery for bladder cancer.