NLP Coding Assignment
V-Labs

Jitendra Chautharia
(jitenchautharia@gmail.com)

Email classification

- [ ] **Dataset: UC Berkeley Enron Email Analysis Project**

[A subset of about 1700 labeled email messages]()

Contains 8 folders and an introductory file, in each folder there are two types of files one is ".txt" file which has a whole email body and another is ".cat" file in which format of ".cat" file is given.

This format is like
Format of each line in .cats file:
n1,n2,n3

n1 = top-level category
n2 = second-level category
n3 = frequency with which this category was assigned to this message

Here are the categories:

1 Coarse genre

1.1 Company Business, Strategy, etc. (elaborate in Section 3 [Topics])
1.2 Purely Personal
1.3 Personal but in professional context (e.g., it was good working with you)
1.4 Logistic Arrangements (meeting scheduling, technical support, etc)
1.5 Employment arrangements (job seeking, hiring, recommendations, etc)
1.6 Document editing/checking (collaboration)
1.7 Empty message (due to missing attachment)
1.8 Empty message

- [ ] **Selected Dataset:** According to the problem statement we have to classify emails in

1.1 Company Business, Strategy
1.2 Purely Personal
1.3 Personal but in professional context (e.g., it was good working with you)

1.4 Logistic Arrangements (meeting scheduling, technical support, etc)
1.5 Employment arrangements (job seeking, hiring, recommendations, etc)
1.6 Document editing/checking (collaboration)

classes. So we have selected the six folders having format [1,1,n3], [1,2,n3], [1,3,n3], [1,4,n3], [1,5,n3], [1,6,n3]. Here n3 frequency with which this category was assigned to this message.

- ☐ **Data Pre-processing:**

1. I have made a ".csv" file in which email body of message with respective class ('Business', 'Personal', 'Personmal_professional', 'logistic_arrangements', 'Employment_arrangement', 'Document-editing')

2. Cleaning data to extract only words from the email body. For this purpose i have performed various operations of email body txt.

   2.1. Making words in lower case
   2.2. Remove_urls
   2.3. Remove_html
   2.4. Removing Numbers
   2.5. Remove_punctuation
   2.6. Tokenization
   2.7. Stopwords removal with ntlk library
   2.8. Applied stemming
   2.9. Lemmatization
   2.10. Spelling correction (not included in the experiment because taking much processing time)

A Sample Before Preprocessing
****
Message-ID: <24956808.1075847598551.JavaMail.evans@thyme>
Date: Sun, 15 Apr 2001 13:02:00 -0700 (PDT)
From: steven.kean@enron.com
To: ray.alvarez@enron.com
Subject: Re: ISO Market Stabilization Plan
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Steven J Kean
X-To: Ray Alvarez
X-cc:
X-bcc:
X-Folder: \Steven_Kean_June2001_1\Notes Folders\All documents
X-Origin: KEAN-S
X-FileName: skean.nsf

Thanks for taking the lead on this. Note Tim's question about
handicapping
the liklihood of approval. Prices will move in the West based on these
odds. We need to have a better view than anyone else.

Ray Alvarez
04/13/2001 01:19 PM
To: James D Steffes/NA/Enron@Enron
cc: Tim Belden/HOU/ECT@ECT, Joe Hartsoe/Corp/Enron@ENRON, Steven J
Kean/NA/Enron@Enron, Alan Comnes/PDX/ECT@ECT, Steve
Walton/HOU/ECT@ECT, Susan
J Mara/NA/Enron@ENRON

Subject: Re: ISO Market Stabilization Plan

Tim, although there's always a "chance" my impression is that the FERC
won't
buy the ban on exports, as this would appear to run afoul of the
Commerce
Clause and certainly goes counter to everything that FERC hopes to
accomplish
with their own Order 888 and 2000 initiatives. I am less certain about
the
direction FERC will go on pricing, since even the staff has recognized
stumbling blocks in their own recommendation and offers possible

variants.

The ISO has not submitted revised tariff sheets for approval yet, so
it is
unlikely they would try to implement their own plan in the near term.
If
they try to do so without FERC approval, possible legal avenues might
include
the filing of a complaint at FERC, asking for fast track processing
(this
"fast" is measured in weeks, not days) and/or seeking injunctive
relief in
court (faster), which can be hard to obtain but not impossible,
depending
entirely on the circumstances.

Will keep you posted if I learn anything new on this. Ray
James D Steffes
04/12/2001 11:21 PM
To: Tim Belden/HOU/ECT@ECT
cc: Joe Hartsoe/Corp/Enron@ENRON, Ray Alvarez/NA/Enron@ENRON, Steven J
Kean/NA/Enron@Enron, Alan Comnes/PDX/ECT@ECT, Steve
Walton/HOU/ECT@ECT, Susan
J Mara/NA/Enron

Subject: Re: ISO Market Stabilization Plan

Ray --

Can you please take the lead in responding to Tim re: FERC v. state
actions?

Sue --

Any info on whether the ISO would do this unilaterally?

Jim

To: Joe Hartsoe/Corp/Enron@ENRON, James D Steffes/NA/Enron@Enron, Ray
Alvarez/NA/Enron@ENRON, Steven J Kean/NA/Enron@Enron, Alan
Comnes/PDX/ECT@ECT, Steve Walton/HOU/ECT@ECT, Susan J
Mara/NA/Enron@ENRON
cc:

Subject: ISO Market Stabilization Plan

The recent plan filed at FERC is horrible. The two most aggregious
parts are
the cost based standing bids and the ban on exports. I know that we
are
commenting on this proposal. I am also looking for intellegence on
whether
the ISO proposal has any chance of getting approved by FERC. If it is
not
approved by FERC, what can the Californians do? California has ignored
FERC
before. If they attempt to unilaterally implement changes what is the
likelihood that the Feds step in to intervene? If you hear anything on
this
matter please keep me posted. The proposed plan will have a huge
impact on
the California market and we need as much advance notice as possible.

****

Same Sample after Preprocessing

****

stevenkeanenroncom rayalvarezenroncom subject iso ket stabil plan mimevers contenttyp textplain charsetusascii contenttransferencod bit xfrom steven j kean xto ray alvarez xcc xbcc xfolder stevenkeanenot foldersal document xorigin kean xfilenam skeannsf thank take lead note tim question handicap liklihood approv price move west base odd need better view anyon el ray alvarez pm jame steffesnaenronenron cc tim beldenhouectect joe hartsoecorpenronenron steven j keannaenronenron alan comnespdxectect steve waltonhouectect susan j anaenronenron subject iso ket stabil plan tim although there alway chanc impress ferc wont buy ban export would appear run afoul commerc claus certainli goe counter everyth ferc hope accomplish order initi le certain direct ferc go price sinc even staff recogn stumbl block recommend offer possibl variant iso submit revis tariff sheet approv yet unlik would tri implement plan near term tri without ferc approv possibl legal avenu might includ file complaint ferc ask fast track process fast measur week day andor seek inct relief court faster hard obtain imposs depend entir circumst keep post learn anyth new ray jame steff pm tim beldenhouectect cc joe hartsoecorpenronenron ray alvareznaenronenron steven j keannaenronenron alan comnespdxectect steve waltonhouectect susan j anaenron subject iso ket stabil plan ray plea take lead respond tim ferc v state action sue info whether iso would unilater jim joe hartsoecorpenronenron jame steffesnaenronenron ray alvareznaenronenron steven j keannaenronenron alan comnespdxectect steve waltonhouectect susan j anaenronenron cc subject iso ket stabil plan recent plan file ferc horribl two aggregi part cost base stand bid ban export know comment propos also look intelleg whether iso propos chanc get approv ferc approv ferc californian california ignor ferc attempt unilater implement chang likelihood fed step interven hear anyth matter plea keep post propos plan huge impact california ket need much advanc notic possible

****

- ☐ **Modeling Method:**

Splitted the data between 80-20 among train and test. Then we checks the value count of class.
As we can see from the Fig:1 that the value count of the classes are not balanced so i assigned the weight to the classes.

```
Business                    834
logistic_arrangements       476
Document_editing            143
Personmal_professional      100
Employment_arrangement      74
Personal                    36
Name: CATEGORY, dtype: int64
```

Fig:1 Value count of classes.

We have assign the lower value to the class having high value count and higher value to the class having lower value count as you can see in the below fig:2 .

```
{0: 0.33233413269384493,
 1: 0.5822829131652661,
 2: 1.9382284382284383,
 3: 2.7716666666666665,
 4: 3.7454954954954953,
 5: 7.699074074074074}
```

Fig:2 Weights for class balancing.

Now we have assign the keys('Business', 'Personal', 'Personmal_professional','logistic_arrangements', 'Employment_arrangement', 'Document-editing')
And values(0, 1, 2, 3, 4, 5) as representation of the classes.

Now we use Text embedding based on feed-forward Neural-Net Language Models[1]. It Maps from text to 128-dimensional embedding vectors. Model summary is given below in fig:3. We use 6 neuron dense layer with "softmax" activation function for our model. The module takes a batch of sentences in a 1-D tensor of strings as input.

Model: "sequential_6"

Layer (type)	Output Shape	Param #
keras_layer_4 (KerasLayer)	(None, 128)	124642688
dense_30 (Dense)	(None, 128)	16512
dropout_24 (Dropout)	(None, 128)	0
dense_31 (Dense)	(None, 128)	16512
dropout_25 (Dropout)	(None, 128)	0
dense_32 (Dense)	(None, 64)	8256
dropout_26 (Dropout)	(None, 64)	0
dense_33 (Dense)	(None, 32)	2080
dropout_27 (Dropout)	(None, 32)	0
dense_34 (Dense)	(None, 6)	198

Total params: 124,686,246
Trainable params: 124,686,246
Non-trainable params: 0

Fig:3 Model summary

For training we use 'adam' optimizer, 'CategoricalCrossentropy' loss function and for evaluation purpose precision, recall, accuracy and confusion matrix.

Default learning rate

batch size = 64

epochs=12

333 is the count of validation data while 1330 is the count for training data. Total data count is 1663

Training stats are given in fig:4 below.

```
Epoch 1/12
/usr/local/lib/python3.7/dist-packages/tensorflow/python/util/dispatch.py:1082: UserWarning: "`categorical_crossentropy` received `from_logits=True`,
  return dispatch_target(*args, **kwargs)
21/21 [==============================] - 19s 828ms/step - loss: 1.8056 - accuracy: 0.2820 - val_loss: 1.2556 - val_accuracy: 0.3393
Epoch 2/12
21/21 [==============================] - 16s 770ms/step - loss: 1.4386 - accuracy: 0.3782 - val_loss: 1.2654 - val_accuracy: 0.4084
Epoch 3/12
21/21 [==============================] - 16s 781ms/step - loss: 1.2258 - accuracy: 0.3940 - val_loss: 1.1598 - val_accuracy: 0.5015
Epoch 4/12
21/21 [==============================] - 16s 740ms/step - loss: 1.0358 - accuracy: 0.5173 - val_loss: 1.0699 - val_accuracy: 0.6547
Epoch 5/12
21/21 [==============================] - 15s 736ms/step - loss: 0.9080 - accuracy: 0.6308 - val_loss: 1.1545 - val_accuracy: 0.6757
Epoch 6/12
21/21 [==============================] - 16s 750ms/step - loss: 0.7951 - accuracy: 0.7015 - val_loss: 0.9714 - val_accuracy: 0.7027
Epoch 7/12
21/21 [==============================] - 16s 747ms/step - loss: 0.7034 - accuracy: 0.7368 - val_loss: 1.1247 - val_accuracy: 0.7177
Epoch 8/12
21/21 [==============================] - 16s 741ms/step - loss: 0.5845 - accuracy: 0.7932 - val_loss: 1.1095 - val_accuracy: 0.7087
Epoch 9/12
21/21 [==============================] - 16s 764ms/step - loss: 0.4947 - accuracy: 0.8226 - val_loss: 1.1100 - val_accuracy: 0.7207
Epoch 10/12
21/21 [==============================] - 17s 783ms/step - loss: 0.3939 - accuracy: 0.8519 - val_loss: 1.2809 - val_accuracy: 0.7177
Epoch 11/12
21/21 [==============================] - 16s 782ms/step - loss: 0.3251 - accuracy: 0.8504 - val_loss: 1.4897 - val_accuracy: 0.7177
Epoch 12/12
21/21 [==============================] - 16s 770ms/step - loss: 0.2799 - accuracy: 0.8737 - val_loss: 1.5488 - val_accuracy: 0.6937
```

Fig:4 Training stats.

Maximum validation accuracy in 12 epochs achieved at epoch 9 i.e. 0.7207 .

- ☐ **Result and analysis of result:**

Results are evaluated at
 Epoch 12/12
21/21 [==============================] - 16s 770ms/step - loss: 0.2799 - accuracy: 0.8737 - val_loss: 1.5488 - val_accuracy: 0.6937

We evaluated the model by its Precision,recall, f1-score, accuracy. These values are given below fig:5 and fig:6.

	precision	recall	f1-score	support
0	0.83	0.76	0.79	195
1	0.00	0.00	0.00	4
2	0.14	0.20	0.17	15
3	0.60	0.77	0.68	100
4	0.50	0.16	0.24	19
accuracy			0.69	333
macro avg	0.42	0.38	0.38	333
weighted avg	0.70	0.69	0.69	333

Fig:5 Precision,recall, f1-score, accuracy.

```
array([[148,   0,  12,  35,   0],
       [  0,   0,   0,   4,   0],
       [  7,   0,   3,   4,   1],
       [ 18,   0,   3,  77,   2],
       [  5,   0,   3,   8,   3]])
```

Fig:6 Confusion Matrix.

The validation accuracy of the model can be improved by including spelling correction in preprocessing steps and we can experiment with orders of the pre-processing operation. Training with more data with fine-tuning can also improve the validation accuracy.

References

[1] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, Christian Jauvin. A Neural Probabilistic Language Model. Journal of Machine Learning Research, 3:1137-1155, 2003.