

2019-10-25 Lab meeting 2

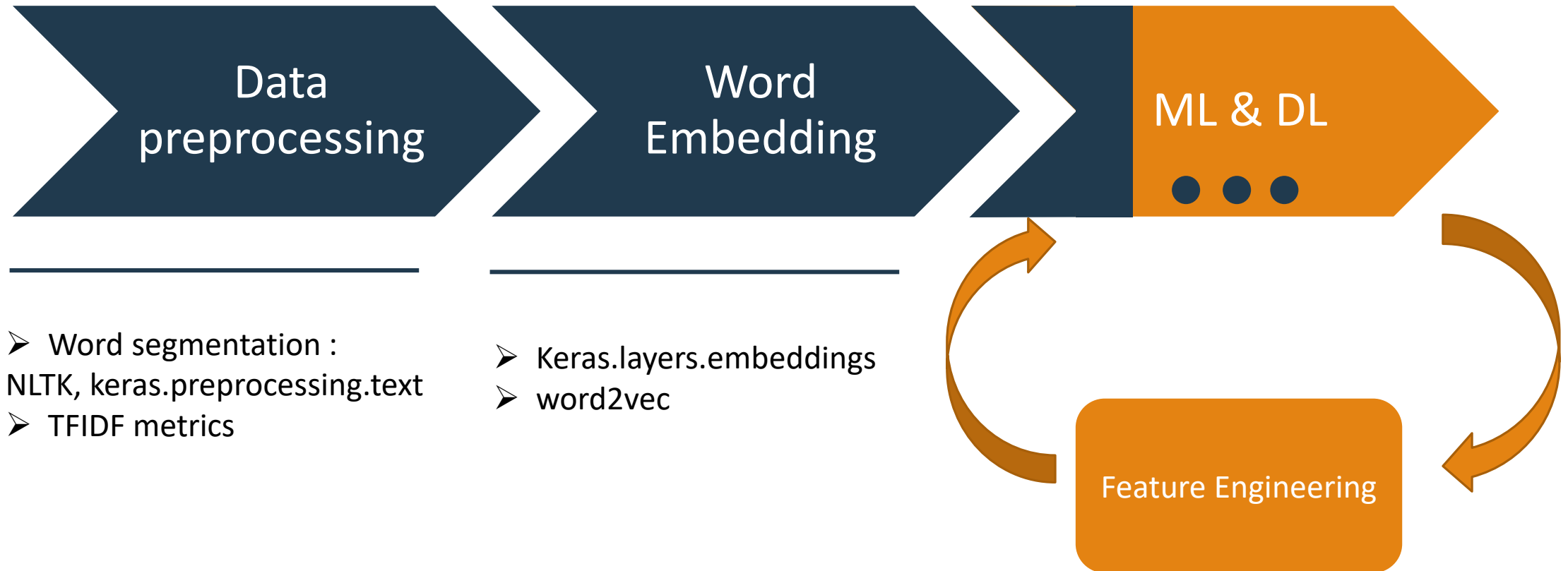
黃彥鈞 (WEBER, HUANG)

Outline

1. AI cup
2. Data release survey : Harvard Data Verse

2019 AI cup

AI cup progress (1/2)



AI cup progress (2/2)

- ML model works
- Classification Method :
 - SVM
 - Naïve Bayes
 - Decision Tree
 - Logistic Regression
 - Random Forest
 - XG Boost
- Deep Neuron Network :
 - MLP
 - RNN
 - CNN
 - CRNN

AI cup Model : MLP example (1/4)

```
from sklearn.model_selection import train_test_split

X_train,X_test,y_train,y_test = train_test_split(k_x_train,K_y_train,test_size=0.4, random_state=
0)

# Try about keras preprocessing
from keras.preprocessing import sequence
from keras.preprocessing.text import Tokenizer

token = Tokenizer(num_words=500)
token.fit_on_texts(X_train)
print(token.document_count)

4200

x_train_seq = token.texts_to_sequences(X_train)
x_test_seq = token.texts_to_sequences(X_test)

xtrain = sequence.pad_sequences(x_train_seq,maxlen=100)
xtest = sequence.pad_sequences(x_test_seq,maxlen=100)

print('Before pad_sequences length=', len(x_train_seq[0]))
print(x_train_seq[0])
print('After pad_sequences length=', len(xtrain[0]))
print(xtrain[0])
```

AI cup Model : MLP example (2/4)

```
# labels' one_hot() encoding
from keras.utils import np_utils
y_train_OneHot = np_utils.to_categorical(y_train)
y_test_OneHot = np_utils.to_categorical(y_test)

from keras.models import Sequential
from keras.layers.core import Dense, Dropout, Activation, Flatten
from keras.layers.embeddings import Embedding

model = Sequential()
model.add(Embedding(output_dim=32,input_dim=500,input_length=100))
model.add(Dropout(0.15))
model.add(Flatten())
model.add(Dense(units=256,activation='relu'))
model.add(Dropout(0.1))
model.add(Dense(units=128))
model.add(Dense(units=8,activation='sigmoid'))
model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])
train_history = model.fit(xtrain, y_train_OneHot, nb_epoch=100, batch_size=100,
                          verbose=2, validation_split=0.2)
```

AI cup Model : MLP example (3/4)

```
# prediction
prediction = model.predict_classes(xtest)
scores = model.evaluate(xtest,y_test_OneHot,verbose=0)

from sklearn import metrics
from sklearn.metrics import accuracy_score

print("Classification report for classifier:\n%s" % ( metrics.classification_report(y_test, prediction.tolist())))
accuracy = accuracy_score(y_test, prediction.tolist())
print("Accuracy: %.2f%%" % (accuracy * 100.0))
print("=====\n")
# output confusion_matrix
import pandas_ml
from pandas_ml import ConfusionMatrix
confusion_matrix = ConfusionMatrix(y_test, prediction.tolist())
print("Confusion matrix:\n%s" % confusion_matrix)
```


AI cup Model : MLP example (4/4)

Classification report for classifier:

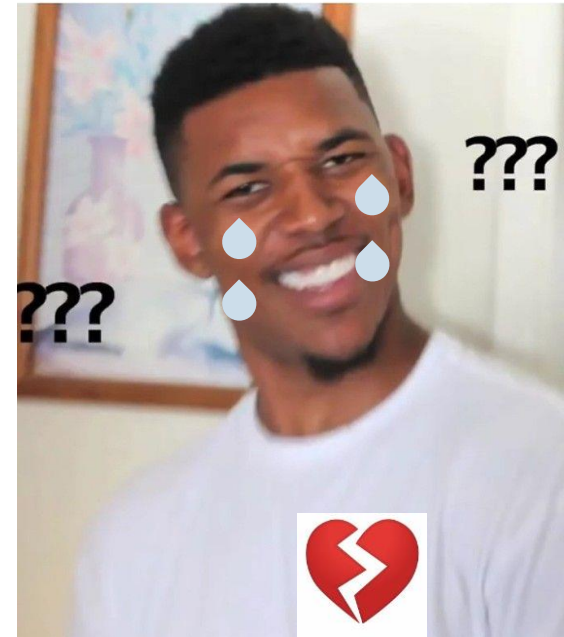
	precision	recall	f1-score	support
0	0.45	0.48	0.47	796
1	0.38	0.40	0.39	783
2	0.22	0.21	0.22	387
3	0.13	0.09	0.10	148
4	0.15	0.15	0.15	243
5	0.12	0.13	0.12	293
6	0.08	0.07	0.08	97
7	0.00	0.00	0.00	53
accuracy			0.31	2800
macro avg	0.19	0.19	0.19	2800
weighted avg	0.30	0.31	0.31	2800

Accuracy: 31.18%

=====

Confusion matrix:

Predicted	0	1	2	3	4	5	6	7	__all__
Actual									
0	386	152	71	29	34	104	17	3	796
1	157	311	97	20	79	85	28	6	783
2	65	121	82	18	50	32	14	5	387
3	49	30	22	13	12	15	7	0	148
4	39	80	53	9	37	16	8	1	243
5	101	89	25	8	22	37	6	5	293
6	32	20	18	3	8	9	7	0	97
7	23	15	3	3	5	3	1	0	53
__all__	852	818	371	103	247	301	88	20	2800



- Adjust the parameter
- Try out another DNN model

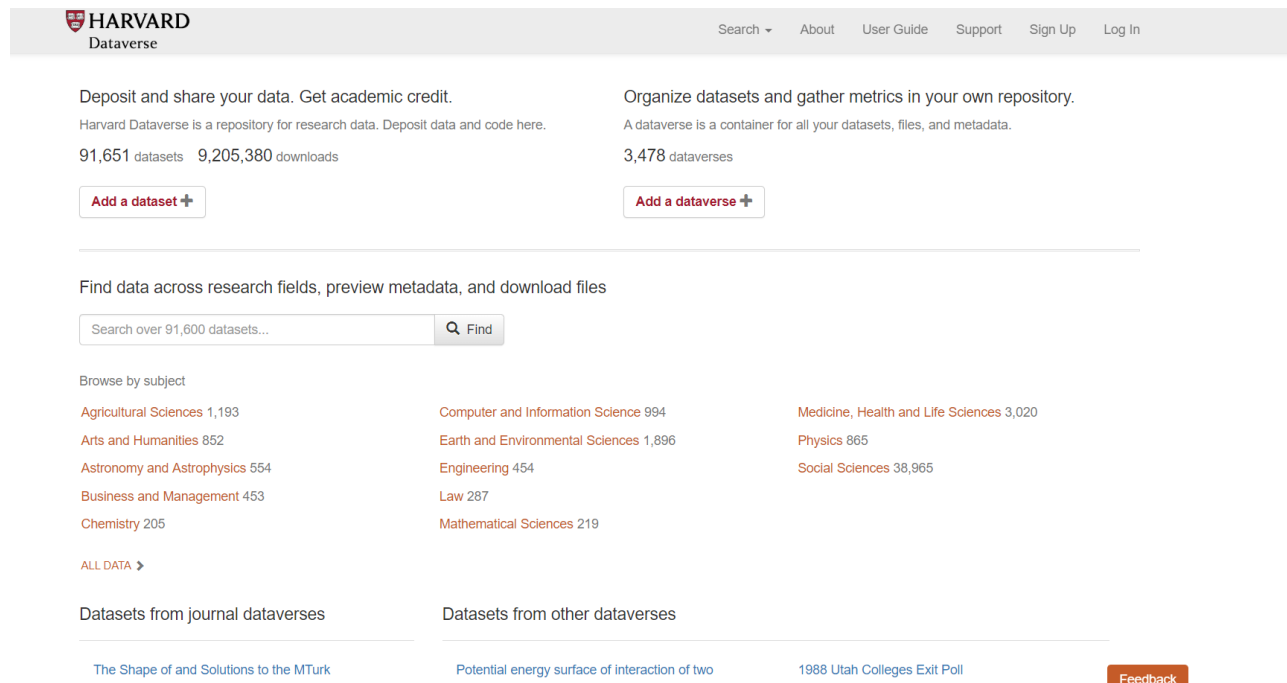
Next week's goals

- Upload submission to leaderboard
- Finish trying every model
- Additional, Go on to feature engineering

Data release survey

Harvard Dataverse

Harvard Dataverse



The screenshot shows the Harvard Dataverse homepage. At the top is a navigation bar with the Harvard Dataverse logo, a search bar, and links for About, User Guide, Support, Sign Up, and Log In. Below the navigation bar, there are two main sections: 'Deposit and share your data. Get academic credit.' and 'Organize datasets and gather metrics in your own repository.' Each section includes a description, statistics (91,651 datasets, 9,205,380 downloads for the first; 3,478 dataverses for the second), and an 'Add a dataset/dataverse' button. A search bar is located below these sections, with the text 'Find data across research fields, preview metadata, and download files'. Below the search bar, there is a 'Browse by subject' section with a grid of subject categories and their respective dataset counts. At the bottom, there are two sections: 'Datasets from journal dataverses' and 'Datasets from other dataverses', each displaying a list of dataset titles. A 'Feedback' button is located in the bottom right corner.

HARVARD
Dataverse

Search ▾ About User Guide Support Sign Up Log In

Deposit and share your data. Get academic credit.
Harvard Dataverse is a repository for research data. Deposit data and code here.
91,651 datasets 9,205,380 downloads
[Add a dataset +](#)

Organize datasets and gather metrics in your own repository.
A dataverse is a container for all your datasets, files, and metadata.
3,478 dataverses
[Add a dataverse +](#)

Find data across research fields, preview metadata, and download files

Search over 91,600 datasets... [Find](#)

Browse by subject

Agricultural Sciences 1,193	Computer and Information Science 994	Medicine, Health and Life Sciences 3,020
Arts and Humanities 852	Earth and Environmental Sciences 1,896	Physics 865
Astronomy and Astrophysics 554	Engineering 454	Social Sciences 38,965
Business and Management 453	Law 287	
Chemistry 205	Mathematical Sciences 219	

[ALL DATA >](#)

Datasets from journal dataverses

Datasets from other dataverses

[The Shape of and Solutions to the MTurk](#)

[Potential energy surface of interaction of two](#)

[1988 Utah Colleges Exit Poll](#)

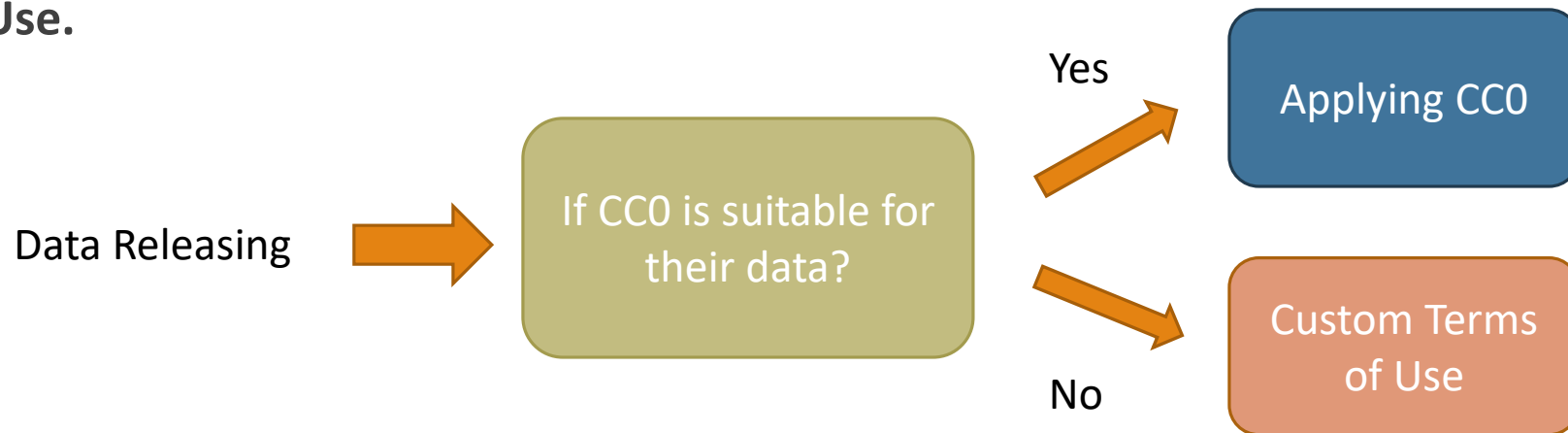
[Feedback](#)

A collaboration with the Institute for Quantitative Social Science (IQSS), the Harvard Library, and Harvard University Information Technology (HUIT):

the Harvard Dataverse is a repository for sharing, citing, analyzing, and preserving research data. It is open to all scientific data from all disciplines worldwide.

The licensing of dataset release (1/3)

- **By default**, all new datasets created through Dataverse's web UI are given a [Creative Commons CC0 Public Domain Dedication](#).
- Additional restrictions, conditions, and terms can still be compatible with CC0, as CC0 only operates in the realm of copyright, which is rather limited when it comes to data.
- If a data owner feels that CC0 is not suitable for their data, they are able to enter **custom Terms of Use**.



The licensing of dataset release (2/3)

Choose a license

This chooser helps you determine which Creative Commons License is right for you in a few easy steps. If you are new to Creative Commons, you may also want to read [Licensing Considerations](#) before you [get started](#).




Allow adaptations of your work to be shared?
Allow commercial uses of your work?

Every license helps creators — we call them licensors if they use our tools — retain copyright while allowing others to copy, distribute, and make some uses of their work — at least non-commercially.



The licensing of dataset release (3/3)

- To add more information about the Terms of Use, we have provided fields like Special Permissions, Restrictions, Citation Requirements, etc.
- Example of data usage agreement 
- Restrict files in dataset

Sample Data Usage Agreement

***Note: This is a sample DUA for datasets that have de-identified human subject data*

This is an agreement ("Agreement") between you the downloader ("Downloader") and the owner of the materials ("User") governing the use of the materials ("Materials") to be downloaded.

I. Acceptance of this Agreement

By downloading or otherwise accessing the Materials, Downloader represents his/her acceptance of the terms of this Agreement.

II. Modification of this Agreement

Users may modify the terms of this Agreement at any time. However, any modifications to this Agreement will only be effective for downloads subsequent to such modification. No modifications will supersede any previous terms that were in effect at the time of the Downloader's download.

III. Use of the Materials

Use of the Materials include but are not limited to viewing parts or the whole of the content included in the Materials; comparing data or content from the Materials with data or content in other Materials; verifying research results with the content included in the Materials; and extracting and/or appropriating any part of the content included in the Materials for use in other projects, publications, research, or other related work products.

1. Representations

In Use of the Materials. Downloader represents that:

Thanks for listening !!!