PROJECT 5

ANALYSIS

* Should I filter english questions ? Lot of false positives.
* Should I remove the body with too many characters ? Check boxplot.
* How should I use Bigram and Trigram ? Ca reduirait la taille de mon corpus. Should I lemmatize the words there ?
* Should I convert the numbers to their letters ? For now I remove them.s
* Bag of words with dimensionality reduction ? TSNE following tf-idf ?
* Should I remove the href links from my text ? Or http links ? I could use the filter about this earlier in the cleaning method.
* What else should I show in my analysis ?
* How many questions should I use overall, now it’s 2500.

UNSUPERVISED

* C\_V good method for evaluation ? Ligne 112.
* Should I also compare here the 3 approches de Word/Sentence Embedding ?

SUPERVISED

* I’ll use MultiOutputClassifier Binary Relevance scheme and the Classifier Chain Scheme, should I use another one as well ?
* Is Hamming Loss a good way to evaluate a model ? Subset accuracy seems to extreme as I don’t want the exact orders and all tags necessarily.
* What scoring algo should I use in my GridsearchCV like roc\_auc, accuracy ratio, F1,, … ?
* Word2Vec(Continuous Bag of Words or The skip gram model) or Sentence2Vec or Doc2Vec or Glove ? Doc2Vec seems to also consider the context of the words.
* Should my word embedding methods be in the pipeline or can I do it at the beginning like now ?
* How to compare unsupervised vs supervised ? A validation set of 20 %?

MLOPS

* What should be unit tested here ? The words embedding functions ?

training set size:800, test set size:200

Using MultiOutputClassifier now:

Fitting 2 folds for each of 48 candidates, totalling 96 fits

Best mean squared score:-0.003448087431693989 with params:{'estimator\_\_max\_depth': 7, 'estimator\_\_max\_features': 8}

Hamming loss:0.0034153005464480873

Using BinaryRelevance now:

Fitting 2 folds for each of 48 candidates, totalling 96 fits

Best mean squared score:0.0025 with params:{'classifier\_\_max\_depth': 6, 'classifier\_\_max\_features': 8}

Hamming loss:0.0034207650273224044

Using ClassifierChain now:

Fitting 2 folds for each of 48 candidates, totalling 96 fits

Best mean squared score:0.00125 with params:{'classifier\_\_max\_depth': 5, 'classifier\_\_max\_features': 9}

Hamming loss:0.003469945355191257

100 rows dataset

Using MultiOutputClassifier now:

Fitting 2 folds for each of 1 candidates, totalling 2 fits

[CV 1/2] END estimator\_\_max\_depth=5, estimator\_\_max\_features=6;, score=(train=1.000, test=0.000) total time= 36.2s

[CV 2/2] END estimator\_\_max\_depth=5, estimator\_\_max\_features=6;, score=(train=1.000, test=0.000) total time= 36.6s

Best mean squared score:0.0 with params:{'estimator\_\_max\_depth': 5, 'estimator\_\_max\_features': 6}

Hamming loss:0.021739130434782608

Using BinaryRelevance now:

Fitting 2 folds for each of 1 candidates, totalling 2 fits

[CV 1/2] END classifier\_\_max\_depth=5, classifier\_\_max\_features=6;, score=(train=1.000, test=0.000) total time= 35.4s

[CV 2/2] END classifier\_\_max\_depth=5, classifier\_\_max\_features=6;, score=(train=1.000, test=0.000) total time= 35.6s

Best mean squared score:0.0 with params:{'classifier\_\_max\_depth': 5, 'classifier\_\_max\_features': 6}

Hamming loss:0.021739130434782608

Using ClassifierChain now:

Fitting 2 folds for each of 1 candidates, totalling 2 fits

[CV 2/2] END classifier\_\_max\_depth=5, classifier\_\_max\_features=6;, score=(train=1.000, test=0.000) total time= 39.4s

[CV 1/2] END classifier\_\_max\_depth=5, classifier\_\_max\_features=6;, score=(train=1.000, test=0.000) total time= 39.7s

Best mean squared score:0.0 with params:{'classifier\_\_max\_depth': 5, 'classifier\_\_max\_features': 6}

Hamming loss:0.021739130434782608