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 ?

UNSUPERVISED

* C\_V good method for evaluation ? Ligne 112.

SUPERVISED

* I’ll use Binary Relevance scheme and the Classifyier Chain Scheme, should I use another one as well ?
* Is Hamming Loss a good way to evaluate a model ? Should I use more like accuracy ratio, F1, … ? Subset accuracy seems to extreme as I don’t want the exact orders and all tags necessarily.
* Word2Vec(Continuous Bag of Words or The skip gram model) or Sentence2Vec or Doc2Vec ? Doc2Vec seems to also consider the context of the words.