PROJECT 5

* Should I filter english questions ?
* Should I remove the body with too many characters ? Check boxplot.
* Should I remove the small tokens < 3 characters ?
* 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.
* What else should I show in my analysis ?