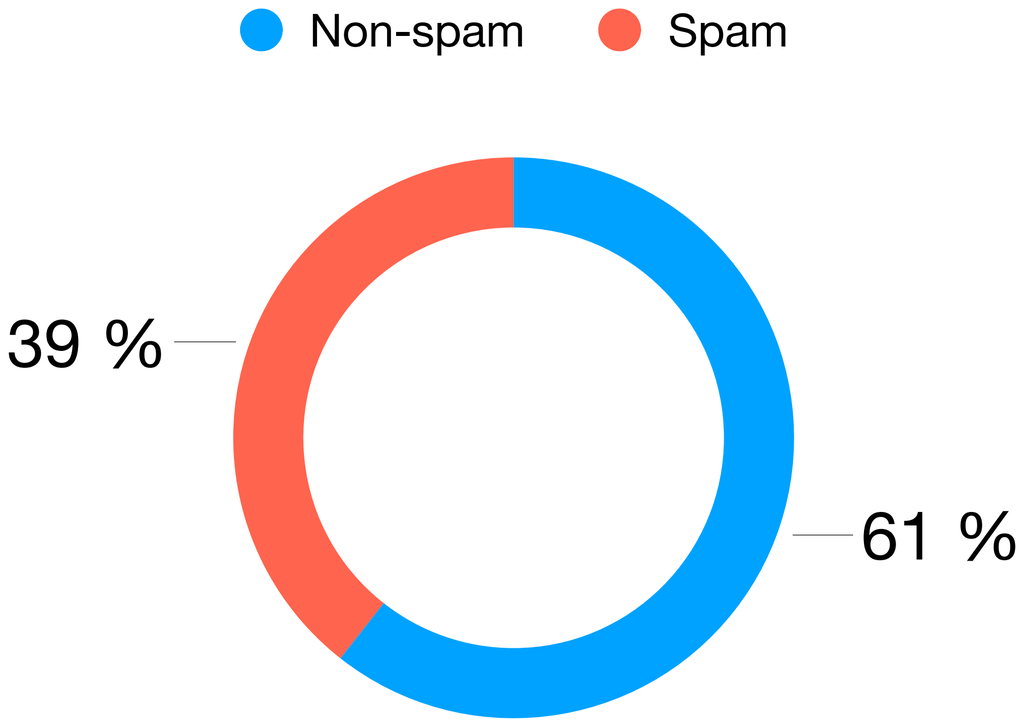
Mail spam detection

Python for Data Analysis

Nicolas Dalbard — DIA 2

Dataset

- This is a classification problem.
- Given a mail, we should predict if it’s a spam or not.
- The dataset is composed of statistics of 4600 mails.



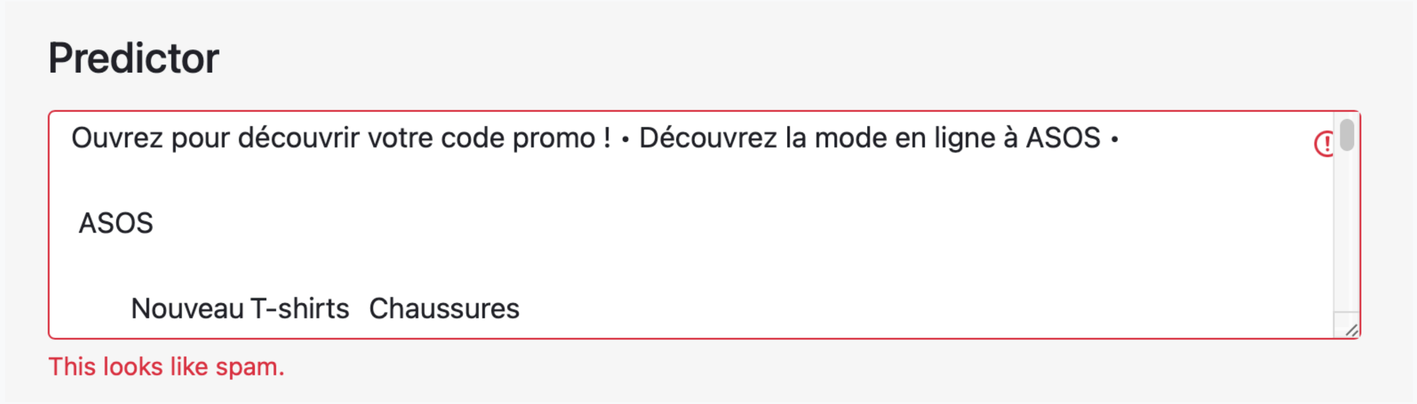
Data analysis

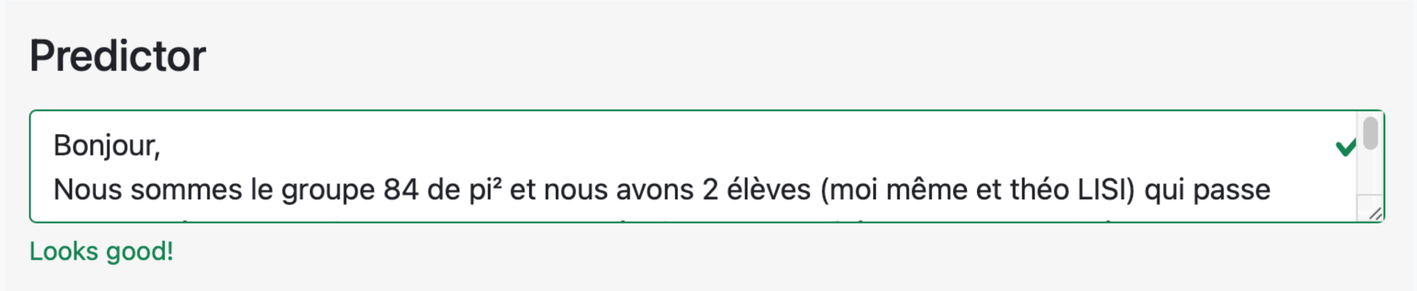
- For 60 words/caracters, it gives their frequency over the total words/ caracters.
- The 3 last variables gives information on capitals letters in the mail.
- In average, a non-spam email has 161 capitals whereas a spam email has 470 capitals.
- Moreover, 80% of spam emails have more than 78 capitals versus 14 for non-spam emails.

Using the trained model

- For the API to give predictions, we need to compute the input excepted by the model by analyzing the content of the mail.
- We must count the required words & get the statistics on capital letters.

Testing


Predictor
Ouvrez pour découvrir votre code promo ! • Découvrez la mode en ligne à ASOS •
ASOS
Nouveau T-shirts Chaussures
This looks like spam.


Predictor
Bonjour,
Nous sommes le groupe 84 de pi² et nous avons 2 élèves (moi même et théo LISI) qui passe
Looks good!

Conclusion

- The model gives satisfying results with English spam emails.
- With French emails, it can only rely on the variables based on the capital letters which lead to many false negatives.