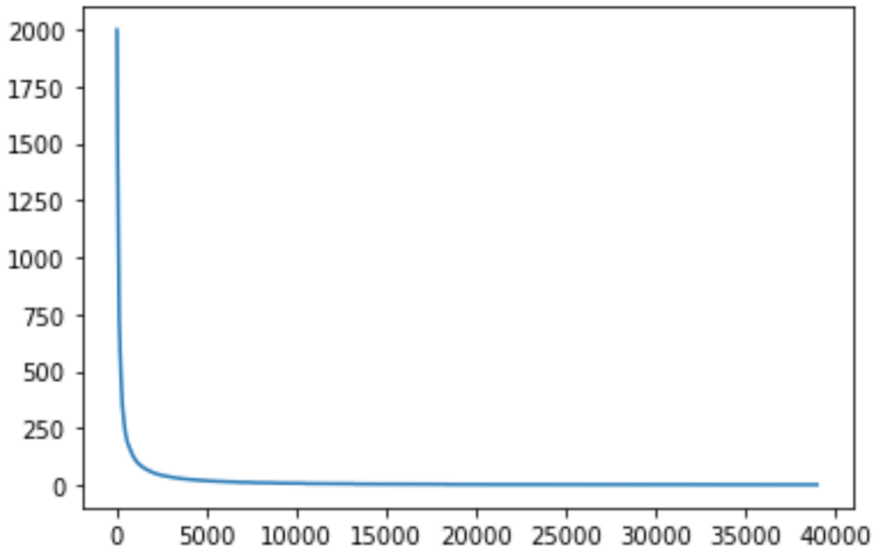
Classification de revues de film

Introduction

Le jeu de données est composé de 20 000 revues, dont 10 000 positives et 10 000 négatives. Les classes sont donc équilibrées.
Le vocabulaire, après suppression de la ponctuation, des accents et majuscules, est composé de **38890 mots** répartis selon leur nombre d'apparitions de la manière suivante :



Fréquence documentaire des mots du vocabulaire
(Triés par fréquence décroissante)

Peu de mots ont une fréquence importante.

Méthode

Nous avons testé trois prétraitements :

1- Prétraitements basiques (ponctuation, accents, majuscules), suppression des stop-words et stemming
2- Prétraitement 1 + uni/bi/trigramme au niveau mot
3- Prétraitement 1 + bi/tri/quadrigramme au niveau caractères

Pour chaque prétraitement, nous avons entrainé et optimisé (selon les paramètres de régularisation) trois classifieurs : un SVM linéaire, une régression logistique et un classifieur bayésien naïf. L'optimisation a été réalisée via une validation croisée à cinq blocs.

Résultats

Prétraitements effectués	SVM	Régression logistique	Bayésien naïf
Prétraitement 1	0.99	0.99	0.96
Prétraitement 2	1	1	1
Prétraitement 3	0.99	0.99	0.92

Taux de bonne classification sur les données d'entrainements (70% des données) avec validation croisée à cinq blocs

On va donc sélectionner le prétraitement 2. Les trois modèles ayant la même performance sur nos données d'entrainement avec les 3 modèles, on va créer un modèle final qui consiste à faire voter les 3 modèles pour prédire la classe.

Ce modèle final donne sur les données test (30% des données) un taux de bonne classification de **84%.**