Datamining : TP6

M2 Informatique, Universite Paris 7
Anne-Claire Haury

Exercice 1

1. Reprendre les donnees de spams texto. Appliquer leur la transformation TF-IDF.
2. Ecrire une fonction qui prend en entree les donnees TF-IDF, les labels et un entier k et renvoie les k mots avec les valeurs les plus signi cativement di erentes entre les spams et les non-spams.
3. Reduire la matrice de donnees a cet ensemble pour $k = 100$.
4. En utilisant la classe pca de la library sklearn.decomposition.PCA, appliquer a ces donnees une transformation en composantes principales.
5. Representer graphiquement les donnees sur les deux premieres composantes principales.

Exercice 2 : travail en groupe sur vos projets