Fo ille de donn es TP 2

M2 Informa iq e, Uni ersi Paris Didero
Anne-Claire Ha r
2016/2017

L objec if de ce second TP es de o s essa er a cl s ering e la classifica ion s per is e. A par ir d n je de donn es q e o s alle ransformer, o s alle po oir es er diff ren s mod les e comprendre commen choisir le meille r.
Vo re alli e dans ce e affaire es la page d aide de l API sklearn. To es les fonc ions ilis es ici son e pliq es. V rifie a pr alable q elle ersion de sklearn es ins all e s r o re ordina e r :

```
          a
         a
```

R c p ra ion e ransforma ion des donn es

R c p a ion de donn e

Po r ce TP no s iliserons n je de donn es compos de messages SMS q e o s po e l charger s r le si e d co rs:
h ps://si es.google.com/si e/da aminingp7/s ppor s-de-co rs/SMSSpamCollec ion. .
Il s agi de cons r ire n an i-spam par ir de ces donn es.
Chaq e ligne d fichier repr sen e n message e son label (spam/ham). No s iliserons ces donn es po r en ra ner n algori hme de pr dic ion q i fo rnira po r n message ne r ponse en re `0` (ham) e `1` (spam).
A pr alable, no s allons iliser ce je de donn es de mani re non s per is e (sans regarder les labels) e l i appliq er de algori hmes de cl s ering.

E e cice 1 - T an fo ma ion de donn e

1. cri e ne fonc ion `read_dataset(filename)` q i prend en arg men le chemin ers le fichier e re o rne ne lis e de paires pe/ e e. Le pe es n en ier `0` po r les hams (non-spams) e `1` po r les spams.

```
    a   a a         a
      R a           a           a     a           a
       a              a a    b  a  ab     a        a
   a

  A
         a    a       a
  R
      a                           Ea
```

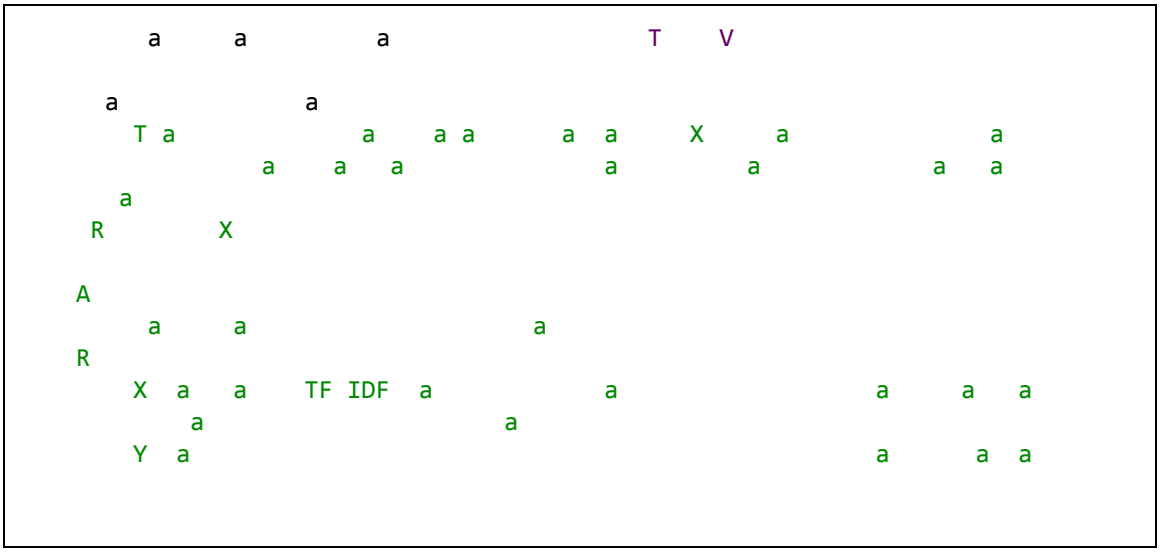
2. cri e ne fonc ion `spams_count(comments)` q i prend en arg men la lis e ren o e par la fonc ion pr c den e e re o rne le nombre de spams dans le je de donn es.

```
    a
      R             b        a        a

  A
           a
  R
      a                           b        a
```

3. Impor e la classe `sklearn.feature_extraction.text.TfidfVectorizer`. Celle-ci o s perme d appliq er l algori hme TF-IDF en co rs. Le cons r c e r accep e n cer ain nombre de param res op ionnels q e o s po e donner po r adap er l algori hme os besoins, e re o rne n obje capable d appliq er ce algori hme. Vo s ro ere ici la doc men a ion de ce e classe. Obser e o s les arg men s e choisisse des ale rs per inen es. Il n a pas ne se le bonne r ponse, mais il fa re capable de j s ifier os choi .

4. cri e ne fonc ion `transform_text(pairs)` q i re o rne ne ma rice `X` (les e es a forma TF-IDF) e n ec e r `y` (0 si le message es n ham, 1 s il s agi d n spam).



Appren issage non s per is (cl s ering)

E e cice 2 - Un imple KMean

Tro e dans la biblio h q e `sklearn` la fonc ion q i o s sera ile po r appliq er n algori hme d KMeans. Lance s r l ensemble `X` n KMeans a ec K = 10. Vo s erre n e emple de scrip s r la page d aide de l algori hme. Inspire - o s en !

al e ce algori hme en regardan la ale r de la fonc ion `sklearn.metrics.silhouette_score`. E pliq e ce q e repr sen e ce e ale r. Bon s si o s es en a ance : cri e ne fonc ion q i ro e les 10 mo s les pl s ilis s dans chaq e cl s er.

E e cice 3 - Choi i le meille K

Ce e fois, o s ne connaisse pas la ale r de K. Commen ro er la meille re ale r de K ? Impl men e ce e sol ion.

E e cice 4 - Compa e de algo i hme de cl e ing

Vo s n a e presq e rien faire po r lancer ce e fois n algori hme de classifica ion hi rarchiq e s r ces donn es. L algori hme s appelle `AgglomerativeClustering` dans la biblio h q e `sklearn`.

Appren issage s per is (classifica ion)

Ce e fois, on cherche pr dire le pe d n message (spam o ham). On a donc iliser `y` q on a ai laiss de c dans la par ie pr c den e.

E e cice 5 - Cla ifica ion pe i e: le pl p oche oi in

1. U ilise la fonc ion `sklearn.model_selection.train_test_split` po r ob enir des ensembles d en ra nemen e de es .
2. Lance n algori hme des pl s proches oisins a ec k = 1 s r l ensemble d en ra nemen . E al e le s r l ensemble de es .
3. Faire de m me a ec k = 3. Leq el donne les meille rs r s l a s?
4. Vo s ene de faire q elq e chose d in erdi . Po rq oi les apes 2 e 3 son -elles in erdi es ?
5. U ilise `sklearn.model_selection.StratifiedKFold` po r choisir le meille r k en re 1 e 100 (niq emen les nombres impairs) s r l ensemble d en ra nemen . Une fois q e o s a e ro le meille r k, relance l algori hme a ec ce e ale r s r la o ali de l ensemble d en ra nemen e es e s r l ensemble de es .

E e cice 6 - Cla ifica ion pe i e: a b e o fo ?

Reprene la derni re q es ion de l e ercice 5 po r choisir les meille rs param res d n arbre de d cision, p is d ne for al a oire. Q el algori hme donne le meille r r s l a ?