1 Ingénierie des données textuelles

De nombreuses applications utilisent des données textuelles pour faire de la prédiction : détection d'opinions, classification automatique de documents en fonction du contenu : spam - no spam, article sport vs article économie, etc...

La classification se fait de manière tout à fait classique par contre il est indispensable de traiter les documents pour pouvoir les faire interpréter par un classifieur. Le traitement des données textuelles est particulièrement difficile car il dépend des données disponibles et tout traitement n'est pas forcément justifié. Par exemple le fait de convertir tout le texte en minuscule peut faire perdre de l'information (Mr Play indique une personne et play un verbe), la suppression des ponctuations peut avoir des conséquences (! est très souvent utilisé pour la détection d'opinions), etc. En outre chaque langue possède aussi ses particularités et les librairies disponibles considèrent souvent l'anglais.

Il existe de nombreuses librairies qui aident à effectuer les prétraitements. Nous en présentons ici quelques unes en mettant en avant la librairie NLTK (Natural Language Toolkit) : http://www.nltk.org (http://www.nltk.org)
Il est conseillé de faire un :

import nltk
nltk.download('all')

pour avoir toutes les librairies.

Encodage des données

Les données textuelles sont souvent sujettes à des problèmes d'encodage (“Latin”, “UTF8” etc). Le plus simple est de les convertir dans un format classique (UTF8).

In [ ]:

```
import unicodedata
chaine = u"Klüft skräms inför på fédéral électoral große"
chaine=unicodedata.normalize('NFKD', chaine).encode('ascii','ignore')
print (chaine)

```

Suppression des tags XML/HTML

Les données textuelles peuvent être issues de pages web, contenir des entêtes, etc..
L'une des premières étapes consistent à les nettoyer pour ne retenir que le texte.

La librairie BeautifulSoup permet de récupérer directement le texte en supprimant les tags :
https://www.crummy.com/software/BeautifulSoup/bs4/doc/
(https://www.crummy.com/software/BeautifulSoup/bs4/doc/)

In [ ]:

```
page = """
<!DOCTYPE html>
<html> <head> <title>Data-Driven Science: A New Paradigm?</title> </head>
  <body>
<h1>Michael L. Nelson</h1> (<a href=mln@cs.odu.edu>) is [here an example]
</body> </html>"""
```

In [ ]:

```
print (page)
```

In [ ]:

```
from bs4 import BeautifulSoup

def strip_html(text):
    soup = BeautifulSoup(text, "html.parser")
    return soup.get_text()

page=strip_html (page)
print (page)
```

De nombreuses modifications peuvent être réalisées à ce niveau notamment en utilisant des expressions régulières. Par exemple la fonction suivante permet de supprimer les textes entre crochets [].

In [ ]:

```
import re
def remove_between_square_brackets(text):
    return re.sub('\[[^]]*\]', '', text)

page=remove_between_square_brackets(page)
print (page)
```

Découpage en phrase

La tokenisation consiste à découper un document en phrases ou une phrase en mots (tokens). Dans un premier temps nous découpons notre document en phrase à l'aide de sent_tokenize.

In [ ]:

```
document = """ Data Science is proving to be of paramount importance to the "

This is due to the increased need more than 100000 for understanding the insu

I'm not sure that we couldn't have a great success. """
```

In [ ]:

```
import nltk
from nltk import sent_tokenize

phrases = sent_tokenize(document)
print("Phrase 0 : ",phrases[0])
print("Phrase 1 : ",phrases[1])
print("Phrase 2 : ",phrases[2])
```

Découpage en mot

La phrase est découpée en tokens à l'aide de word_tokenize

In [ ]:

```
from nltk.tokenize import word_tokenize
tokens = word_tokenize(phrases[0])
print(tokens)
```

Nous pouvons constater que les ponctuations sont considérées comme des tokens.

Mise en minuscule

In [ ]:

```
tokens = [w.lower() for w in tokens]
print (tokens)
```

Transformation des numériques en mots

In [ ]:

```
#inflect est une librairie qui permet de convertir les nombres en mots
import inflect

tokens = word_tokenize(phrases[1])

print ("Nombre à convertir \n")
words = [word for word in tokens if word.isdigit()]
print(words)
p = inflect.engine()
numbertransf = [p.number_to_words(word) for word in tokens if word.isdigit()]

print ("Nombre après conversion \n")
print(numbertransf)


```

Suppression des ponctuations

In [ ]:

```
# Suppression de tous les termes qui ne sont pas alphanumériques
words = [word for word in tokens if word.isalpha()]
print(words)
```

Traitement des contractions

La dernière phrase contient des contractions. Voici ce que cela donne lorsque l'on obtient des tokens et que l'on supprime les caractères non alphabétiques. Des parties de négation disparaissent.

In [ ]:

```
tokens = word_tokenize(phrases[2])
print(tokens)
words = [word for word in tokens if word.isalpha()]
print(words)
```

In [ ]:

```
import contractions


def replace_contractions(text):
    return contractions.fix(text)

print ("Avant remplacement\n")
print (phrases[2])
print ("\nAprès remplacement\n")
laphrase=replace_contractions(phrases[2])
print (laphrase)
tokens = word_tokenize(laphrase)
print(tokens)

```

Suppression des stop words

Les stopwords sont des mots qui n'ont pas beaucoup d'intérêt dans la classification. Il s'agit des mots comme the, as etc. Attention il faut toujours faire attention à la liste des stopwords. Certains d'entre eux sont peut être utiles pour l'analyse. Le fait de supprimer is peut manquer par la suite. De plus il est parfois utile de mettre dans la liste des stopwords les mots très courant du domaine. Si les données parlent toutes de cinéma il faut mettre cinéma dans la liste des stopwords.

NLTK propose une liste de stopwords pour différentes langues. Ils sont en minuscule et avec des contractions. Pour les utiliser il faut donc avoir fait le même traitement sur nos données.
Il existe de nombreux sites qui proposent des listes de stopwords en différentes langues.

In [ ]:

```
from nltk.corpus import stopwords
stop_words = stopwords.words('english')
print(stop_words)
stop_words = stopwords.words('french')
print(stop_words)
```

In [ ]:

```
from nltk.corpus import stopwords
print ("Exemple d'application des stopwords sur la première phrase\n")
tokens = word_tokenize(phrases[0])
tokens = [w.lower() for w in tokens]
words = [word for word in tokens if word.isalpha()]
print ("Avant transformation \n")
print (words)

stop_words = set(stopwords.words('english'))
words = [w for w in words if not w in stop_words]
print ("\nAprès transformation \n")
print(words)
```

Stemmatisation

La stemmatisation (stemming ou racinisation) est l'opération qui consiste à réduire chaque mot à sa racine. Par exemple malad, maladie, maladive en maladive. Les approches utilisent l'algorithme de Porter qui se base sur le suffixe des mots. NTLK propose la librairie SnowballStemmer pour le français.

In [ ]:

```
from nltk.stem.porter import PorterStemmer

tokens = word_tokenize(phrases[0])
tokens = [w.lower() for w in tokens]

porter = PorterStemmer()
stemmed = [porter.stem(word) for word in tokens]
print(stemmed)
```

In [ ]:

```
# un autre stemmatiser
from nltk.stem.lancaster import LancasterStemmer

lancaster_stemmer = LancasterStemmer()
lstemmed = [lancaster_stemmer.stem(word) for word in tokens]
print(lstemmed)
```

In [ ]:

```
# un autre stemmatiseur qui accepte le français
from nltk.stem.snowball import SnowballStemmer
stemmer = SnowballStemmer("french")
phrase = "malade malades maladie maladies maladive"
tokens = word_tokenize(phrase)
print ("Avant transformation \n")
print (tokens)
stemmed = [stemmer.stem(word) for word in tokens]
print ("\n Après transformation\n")
print (stemmed)

```

La lemmatisation

La lemmatisation consiste aussi à réduire chaque mot à sa racine mais par contre elle ne va pas s'intéresser au suffixe du mot. Elle effectue une première analyse pour mettre les verbes à l'infinitif, supprimer les s pour les pluriels.

Le choix de l'une ou de l'autre méthode dépend bien entendu de la tâche de classification que l'on souhaite faire.

In [ ]:

```
from nltk.stem import WordNetLemmatizer

tokens = word_tokenize(phrases[0])
tokens = [w.lower() for w in tokens]

print ("Lemmatisation \n")
wordnet_lemmatizer = WordNetLemmatizer()
lstemmed = [wordnet_lemmatizer.lemmatize(word,pos='v') for word in tokens]
print("Lemmatisation : \n",lstemmed)

print ("\n A comparer avec la Stemmatisation\n")
lancaster_stemmer = LancasterStemmer()
lstemmed = [lancaster_stemmer.stem(word) for word in tokens]
print(lstemmed)


print("Lemmatisation : \n",lstemmed)
print ("\n autre exemple avec la phrase\n")
sentence = "have had having dogs crying"
print (sentence)
tokens = word_tokenize(sentence)


porter = PorterStemmer()
stemmed = [porter.stem(word) for word in tokens]
print("Stemmatisation : \n",stemmed)


wordnet_lemmatizer = WordNetLemmatizer()
lstemmed = [wordnet_lemmatizer.lemmatize(word,pos='v') for word in tokens]
print("Lemmatisation : \n",lstemmed)


```

Part of speech tagging

Cette étape consiste à appliquer un analyseur morpho-syntaxique pour déterminer le genre du mot dans la phrase. Elle peut être très utile pour, par exemple, ne considérer que les adjectifs, les verbes ou les adverbes dans le cas de la détection de l'opinion.

In [ ]:

```
tokens = word_tokenize(phrases[0])
tokens = [w.lower() for w in tokens]

nltk.pos_tag(tokens)
```

Il est possible de connaître l'intitulé de chaque Tag par la fonction nltk.help.upenn_tagset('XX') où XX représente le Tag recherché.

In [ ]:

```
print (nltk.help.upenn_tagset('RB'))
print (nltk.help.upenn_tagset('VB'))

```

Il est possible d'intégrer le genre directement avec le mot à l'aide de la fonction pos_tag

In [ ]:

```
from nltk import pos_tag

tokens = word_tokenize(phrases[0])

tokens=pos_tag(tokens)
print (tokens)
```

Données de type tweet

Les tweets ont une syntaxe très particulière et généralement les traitements se font à l'aide d'expressions régulières.

In [ ]:

```
tweet = '#NLP is thus an example :D ;) RT @theUser: see http://example.com'
```

In [ ]:

```
import re
#traitement des émoticones
emoticons_str = r"""
    (?:
        [:=;] # Eyes
        [oO\-]? # Nose (optional)
        [D\)\]\(\]/\\OpP] # Mouth
    )"""

#Prise en compte des éléments qui doivent être regroupés
regex_str = [
    emoticons_str,
    r'<[^>]+>', # HTML tags
    r'(?:@[\w_]+)', # @-mentions
    r"(?:\#+[\w_]+[\w\'_\-]*[\w_]+)", # hash-tags
    r'http[s]?://(?:[a-z]|[0-9]|[$-_@.&amp;+]|[!*\(\),]|(?:%[0-9a-f][0-9a-f])

    r'(?:(?:\d+,?)+(?:\.?\d+)?)', # nombres
    r"(?:[a-z][a-z'\-_]+[a-z])", # mots avec - et '
    r'(?:[\w_]+)', # autres mots
    r'(?:\S)' # le reste
]

tokens_re = re.compile(r'('+'|'.join(regex_str)+')', re.VERBOSE | re.IGNORECA
emoticon_re = re.compile(r'^'+emoticons_str+'$', re.VERBOSE | re.IGNORECASE)

def tokenize(s):
    return tokens_re.findall(s)

def preprocess(s, lowercase=False):
    tokens = tokenize(s)
    if lowercase:
        tokens = [token if emoticon_re.search(token) else token.lower() for t
    return tokens

# un example de tweet
tweet = '#NLP is thus an example :D RT @theUser: see http://example.com'
print ("Un exemple de tweet : \n",tweet)

print ("\nLe tweet avec un processus normal de transformation\n")
print (word_tokenize(tweet))
print ("\nLe tweet avec des expressions régulières\n")
words=preprocess(tweet)
print(words)
```