data-mining-lab-text-mining

December 2, 2023

1 Data Mining LAB: Text Mining

- réalisé par : chaimaabouabd
- encadré par : Mme Khadija Bouzaachane

1.1 1. Acquisition des données et préparation des données

1.1.1 Chargement du fichier

[37]:
```
import pandas as pd

messagesTwitter = pd.read_csv('./global_warming_tweets.csv', delimiter=',',␣
↪encoding='unicode_escape')
```

[38]:
```
print(messagesTwitter.shape)
```

```
(6090, 3)
```

[39]:
```
messagesTwitter.head(2)
```

[39]:

	tweet	existence	\
0	Global warming report urges governments to act…	Yes	
1	Fighting poverty and global warming in Africa …	Yes	

	existence.confidence
0	1.0
1	1.0

[40]:
```
messagesTwitter['existence'] = (messagesTwitter['existence'] == 'Yes').
↪astype(int)
messagesTwitter.head(100)
```

[40]:

	tweet	existence	\
0	Global warming report urges governments to act…	1	
1	Fighting poverty and global warming in Africa …	1	
2	Carbon offsets: How a Vatican forest failed to…	1	
3	Carbon offsets: How a Vatican forest failed to…	1	
4	URUGUAY: Tools Needed for Those Most Vulnerabl…	1	

```
..                                                  …         …
95  Plants effective way of tackling global warmin…          1
96  Climate change & sustainability will be a key …          1
97  Frederic Hague at #PEN: climate change isn't j…          1
98  US Generals say: Climate Change Threatens Amer…          1
99  Even the generals know climate change is going…          1

    existence.confidence
0                 1.0000
1                 1.0000
2                 0.8786
3                 1.0000
4                 0.8087
..                   …
95                0.7925
96                0.7874
97                0.5778
98                1.0000
99                1.0000

[100 rows x 3 columns]
```

1.1.2 Normalisation

[41]:
```
import re
def normalisation (message):
    message = re.sub('((www\.[^\s]+) | (https?://[^\s]+))', 'URL', message)
    message = re.sub('@[^\s]+', 'USER', message)
    message = message.lower().replace("e", "e")
    message = re.sub('[^a-zA-Za- - 1-9]+', ' ', message)
    message = re.sub(' +',' ', message)
    return message.strip()
```

[42]:
```
messagesTwitter["tweet"] = messagesTwitter["tweet"].apply(normalisation)
print(messagesTwitter.head(10))
```

```
                                               tweet  existence  \
0  global warming report urges governments to act…          1
1  fighting poverty and global warming in africa …          1
2  carbon offsets how a vatican forest failed to …          1
3  carbon offsets how a vatican forest failed to …          1
4  uruguay tools needed for those most vulnerable…          1
5  rt user rt user ocean saltiness shows global w…          1
6  global warming evidence all around us|a messag…          1
7  migratory birds new climate change strategy st…          1
8  southern africa competing for limpopo water cl…          1
9  global warming to impact wheat rice production…          1
```

```
   existence.confidence
0                1.0000
1                1.0000
2                0.8786
3                1.0000
4                0.8087
5                1.0000
6                1.0000
7                1.0000
8                1.0000
9                1.0000
```

[43]:
```
messagesTwitter.head()
```

[43]:
```
                                               tweet  existence  \
0  global warming report urges governments to act…          1
1  fighting poverty and global warming in africa …          1
2  carbon offsets how a vatican forest failed to …          1
3  carbon offsets how a vatican forest failed to …          1
4  uruguay tools needed for those most vulnerable…          1

   existence.confidence
0                1.0000
1                1.0000
2                0.8786
3                1.0000
4                0.8087
```

1.1.3 Suppression des stop words

[44]:
```
import nltk
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data…
[nltk_data]   Package stopwords is already up-to-date!
```

[44]:
```
True
```

[45]:
```
stopWords = stopwords.words('english')
```

[46]:
```
messagesTwitter["tweet"] = messagesTwitter["tweet"].apply(
    lambda message: ' '.join([word for word in message.split() if word not in␣
 ↪(stopWords)]))

messagesTwitter.head(10)
```

[46]:
```
                                               tweet  existence  \
0  global warming report urges governments act|br…          1
1        fighting poverty global warming africa link          1
2  carbon offsets vatican forest failed reduce gl…          1
3  carbon offsets vatican forest failed reduce gl…          1
4  uruguay tools needed vulnerable climate change…          1
5  rt user rt user ocean saltiness shows global w…          1
6  global warming evidence around us|a message gl…          1
7  migratory birds new climate change strategy st…          1
8  southern africa competing limpopo water climat…          1
9  global warming impact wheat rice production in…          1

   existence.confidence
0                1.0000
1                1.0000
2                0.8786
3                1.0000
4                0.8087
5                1.0000
6                1.0000
7                1.0000
8                1.0000
9                1.0000
```

1.1.4 La stemmisation

[47]:
```
from nltk.stem import SnowballStemmer
from nltk.tokenize import word_tokenize

# Téléchargez les données de stemming si ce n'est pas déjà fait
nltk.download('snowball_data')

# Définir la langue pour la stemmatisation (par exemple, 'english')
stemmer = SnowballStemmer('english')

# Définir la fonction lambda pour la stemmatisation des mots
messagesTwitter["tweet"] = messagesTwitter["tweet"].apply(
  lambda message: ' '.join([stemmer.stem(word) for word in message.split()]))


# Afficher le résultat
messagesTwitter.head(10)
```

```
[nltk_data] Downloading package snowball_data to /root/nltk_data…
[nltk_data]   Package snowball_data is already up-to-date!
```

[47]:
```
                                               tweet  existence  \
0  global warm report urg govern act|brussel belg…          1
1              fight poverti global warm africa link          1
2  carbon offset vatican forest fail reduc global…          1
3  carbon offset vatican forest fail reduc global…          1
4         uruguay tool need vulner climat chang link          1
5  rt user rt user ocean salti show global warm i…          1
6  global warm evid around us|a messag global war…          1
7  migratori bird new climat chang strategi stay …          1
8  southern africa compet limpopo water climat ch…          1
9  global warm impact wheat rice product india|lu…          1

   existence.confidence
0                1.0000
1                1.0000
2                0.8786
3                1.0000
4                0.8087
5                1.0000
6                1.0000
7                1.0000
8                1.0000
9                1.0000
```

1.1.5 La lemmatisation

[49]:
```
from nltk.stem import WordNetLemmatizer
nltk.download('wordnet')

lemmatizer = WordNetLemmatizer()
messagesTwitter['tweet'] = messagesTwitter['tweet'].apply(
    lambda message: ' '.join([lemmatizer.lemmatize(word) for word in message.
 ↪split()]))
```

```
[nltk_data] Downloading package wordnet to /root/nltk_data…
```

1.2 2. Phases d'apprentissage et de prédiction

1.2.1 Découpage en jeux de tests et d'apprentissage

[51]:
```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(messagesTwitter['tweet' ].
 ↪values,
messagesTwitter['existence'].values, test_size=0.2)
```

1.2.2 Création d'un pipeline d'apprentissage

[52]:
```
from sklearn.pipeline import Pipeline
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.naive_bayes import MultinomialNB

etapes_apprentissage = Pipeline([('frequence',
CountVectorizer ()),

('tfidf', TfidfTransformer ()),
('algorithme',

MultinomialNB () ) ])
```

1.2.3 Apprentissage et analyse des résultats

[55]:
```
modele = etapes_apprentissage.fit(X_train, y_train)

from sklearn.metrics import classification_report
print (classification_report(y_test, modele.predict (X_test),
digits=4))
```

```
              precision    recall  f1-score   support

           0     0.9105    1.0000    0.9532      1109
           1     0.0000    0.0000    0.0000       109

    accuracy                         0.9105      1218
   macro avg     0.4553    0.5000    0.4766      1218
weighted avg     0.8290    0.9105    0.8679      1218

/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1344:
UndefinedMetricWarning: Precision and F-score are ill-defined and being set to
0.0 in labels with no predicted samples. Use `zero_division` parameter to
control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1344:
UndefinedMetricWarning: Precision and F-score are ill-defined and being set to
0.0 in labels with no predicted samples. Use `zero_division` parameter to
control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1344:
UndefinedMetricWarning: Precision and F-score are ill-defined and being set to
0.0 in labels with no predicted samples. Use `zero_division` parameter to
control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
```

1.2.4 Classification d'un nouveau message

[56]:
```
phrase = "Why should trust scientists with global warming if they didnt know␣
 ↪Pluto wasnt a planet"
print (phrase)

#Normalisation
phrase = normalisation (phrase)

#Suppression des stops words
phrase = ' '.join([mot for mot in phrase.split() if mot not in
(stopWords) ])

#Stemmatisation
phrase = ' '.join([stemmer.stem(mot) for mot in phrase.split(' ')])
```

```
Why should trust scientists with global warming if they didnt know Pluto wasnt a
planet
```

[57]:
```
#Lemmatisation
phrase = ' '.join([lemmatizer. lemmatize(mot) for mot in
phrase.split(' ')])
print(phrase)

prediction = modele.predict([phrase])
print (prediction)
if[prediction[0] == 0]:
  print (">> Ne croit pas au rechauffement climatique ... ")
else:
  print (">> Croit au rechauffement climatique ... ")
```

```
trust scientist global warm didnt know pluto wasnt planet
[0]
>> Ne croit pas au rechauffement climatique …
```