

td2a_sentiment_analysis

August 12, 2022

1 2A.ml - Analyse de sentiments

C'est désormais un problème classique de machine learning. D'un côté, du texte, de l'autre une appréciation, le plus souvent binaire, positive ou négative mais qui pourrait être graduelle.

```
[1]: %matplotlib inline
```

```
[2]: from jupyter_helper import add_notebook_menu
      add_notebook_menu()
```

```
[2]: <IPython.core.display.HTML object>
```

1.1 Les données

On récupère les données depuis le site [UCI Sentiment Labelled Sentences Data Set](#) où on utilise la fonction `load_sentiment_dataset`.

```
[3]: from ensae_teaching_cs.data import load_sentiment_dataset
      df = load_sentiment_dataset()
      df.head()
```

```
[3]:
```

	source	sentance	sentiment
0	amazon_cells_labelled	So there is no way for me to plug it in here i...	0
1	amazon_cells_labelled	Good case, Excellent value.	1
2	amazon_cells_labelled	Great for the jawbone.	1
3	amazon_cells_labelled	Tied to charger for conversations lasting more...	0
4	amazon_cells_labelled	The mic is great.	1

1.2 Exercice 1 : approche td-idf

La cible est la colonne `sentiment`, les deux autres colonnes sont les features. Il faudra utiliser les prétraitements `LabelEncoder`, `OneHotEncoder`, `TF-IDF`. L'un d'entre eux n'est pas nécessaire depuis la version `0.20.0` de `scikit-learn`.

```
[4]:
```

1.3 Exercice 2 : word2vec

On utilise l'approche `word2vec` du module `gensim` ou `spacy`.

[5] :

1.4 Exercice 3 : comparer les deux approches

Avec une courbe `ROC` par exemple.

[6] :