# UE 12 Extraction

MASSART DOLANT

# Ce qu’on a fait

1. Récupération de tweets avec TwitterSearch
2. Stockage, formattage, nettoyage
3. Entraînement de plusieurs modèles avec des *preprocessing* différents

```python
from string import punctuation
# Remove HTML special entities (e.g. &amp;)
tweet = re.sub(r'\&\w*;', '', tweet)
#Convert @username to AT_USER
tweet = re.sub(r'@[^\s]+','',tweet)
# remove numbers
tweet = re.sub(r'\d+', ' ', tweet)
tweet = re.sub(r'([a-z])([A-Z])','\\1 \\2', tweet)
# Remove tickers
tweet = re.sub(r'\$\w*', '', tweet)
# To lowercase
tweet = tweet.lower()
# Remove hyperlinks
tweet = re.sub(r'https:\/\/t.co\/.{9}', '', tweet)
# Remove hashtags
tweet = re.sub(r'#', ' ', tweet)
# Remove Punctuation and split 's, 't, 've with a space for filter
tweet = re.sub(r'[' + punctuation.replace('@', '') + ']+', ' ', tweet)
tweet = re.sub(r'[^\w\s]', ' ', tweet)
# Remove words with 2 or fewer letters
tweet = re.sub(r'\b\w{1,2}\b', '', tweet)
# Remove whitespace (including new line characters)
tweet = re.sub(r'\s\s+', ' ', tweet)
# Remove single space remaining at the front of the tweet.
tweet = tweet.lstrip(' ')
# Remove characters beyond Basic Multilingual Plane (BMP) of Unicode:
tweet = ''.join(c for c in tweet if c <= '\uFFFF')
return tweet
```

# Résultats

| | LTSM | Simple NN |
|---|---|---|
| Bag of words | 62% | 80% |
| Tokenisation | 56% | 80% |
| One Hot | X | 77.8% |

# Mention spéciale à Multinomial Naives Bayes

| Bag of words | Tokenization | One Hot |
|---|---|---|
| 41% | 40% | 37% |

# Simple NN

| |
|---|
| EMBEDDING |
| FLATTEN |
| DENSE |