Université Mohammed V

Faculté des Sciences - Rabat

# Mini projet

## Natural Language Processing

**Réalisé par :** Morjani zouhir

Mohamed Amine Zaher

**Encadré par :** Abdelhak Mahmoudi

**2019-2020**

## Introduction :

Les données sont au cœur de tout projet de science des données, mais nous tenons souvent pour acquis la disponibilité des données, surtout lorsqu'elles arrivent proprement dans une base de données SQL ou mieux encore dans notre boîte de réception. Cela dit, parfois, les données que vous recherchez ne sont pas facilement disponibles en raison de leur nature spécifique. Une solution possible à ce problème est l'idée de Web Scraping ou l'extraction d'informations à partir d'un site Web spécifique en lisant attentivement son HTML. Par exemple, supposons que vous planifiez des vacances et que vous soyez à l'affût de la date de mise en vente du billet d'avion. Oui, vous pouvez parcourir le même site de voyage chaque heure en espérant que le prix baisserait, mais une méthode plus efficace serait de gratter le site de voyage chaque heure et de disposer d'un fichier de sortie vous fournissant les prix les plus récents des billets. Avertissement De nombreux sites Web n'aiment pas que leurs données soient récupérées, en particulier si elles contiennent des informations utilisateur identifiables (par exemple, Facebook, Linkedin, etc.). Veuillez tenir compte des données que vous choisissez de récupérer et de leur fréquence.

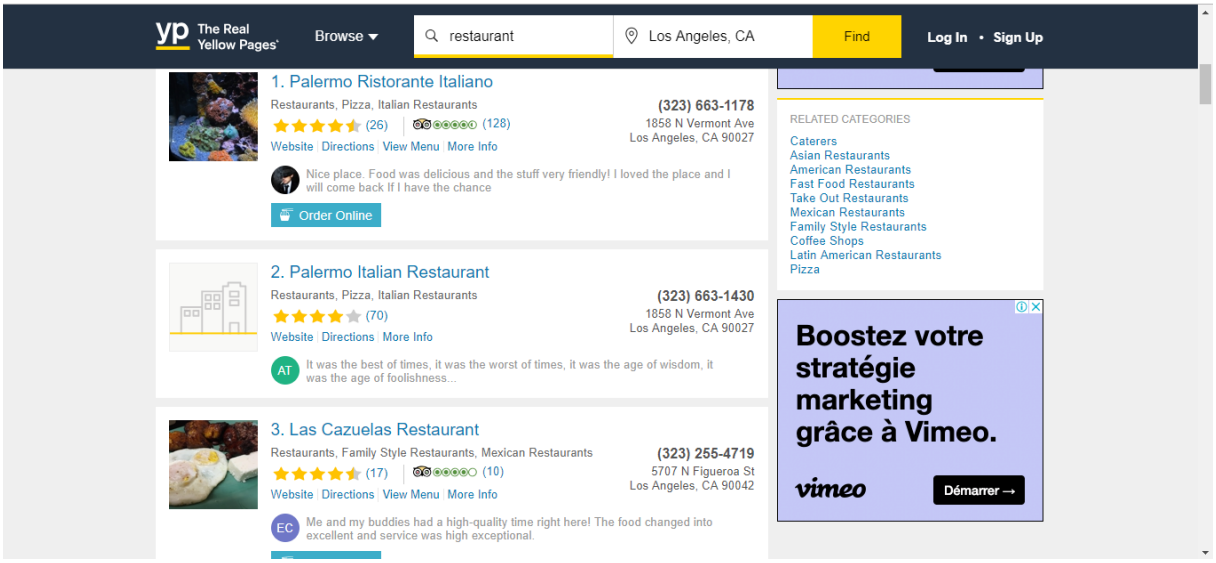
Dans ce projet, nous explorerons les techniques de récupération des données de site Web, de prétraitement et de préparation de nos données pour l'analyse, et enfin des traitement d'informations à partir de nos données NLP.

## Web Scraping :

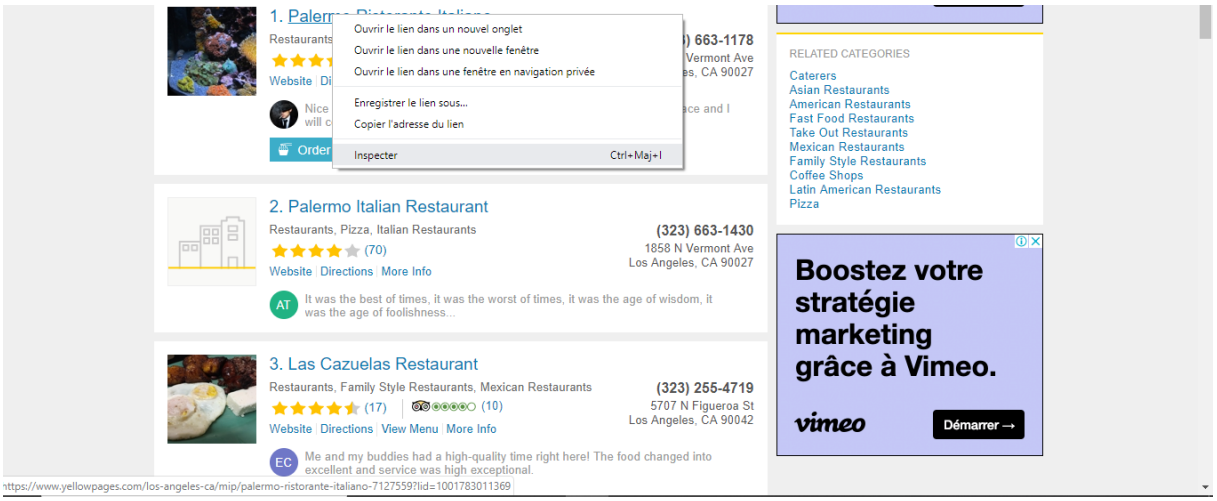
Dans ce projet, Nous essayons de faire 'Web Scraping' pour Yellowpage.com, mais plus particulièrement les avis sur les restaurants. Ciblons les notes des clients, examinons les titres, révisons les descriptions ainsi que commentaire des clients.

## Bases du HTML :

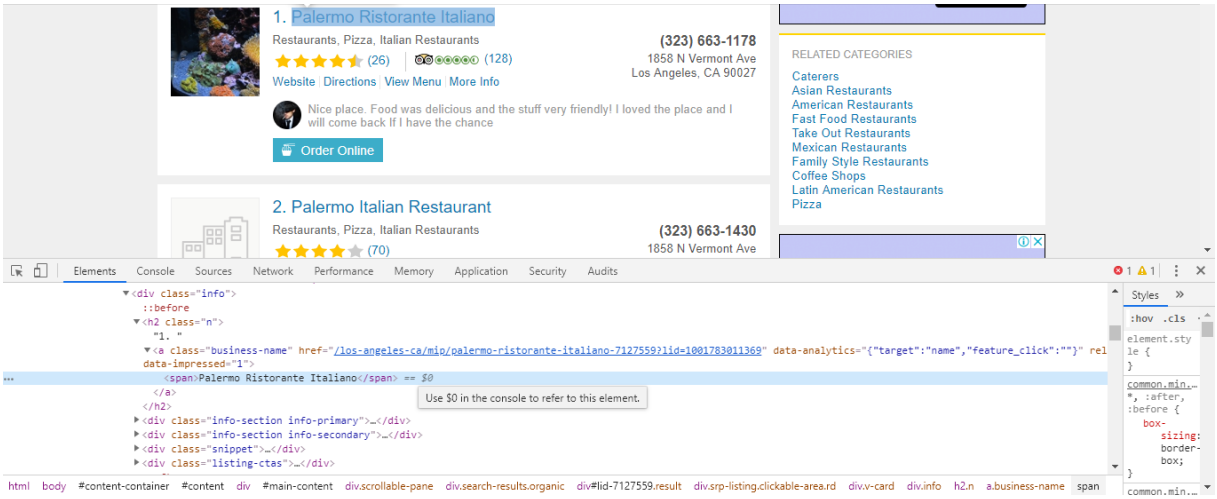
Avant de pouvoir réellement commencer à collecter des informations, nous devons nous familiariser avec la structure de base du HTML car nous utiliserons en fait des balises HTML pour identifier les informations que nous souhaitons collecter.

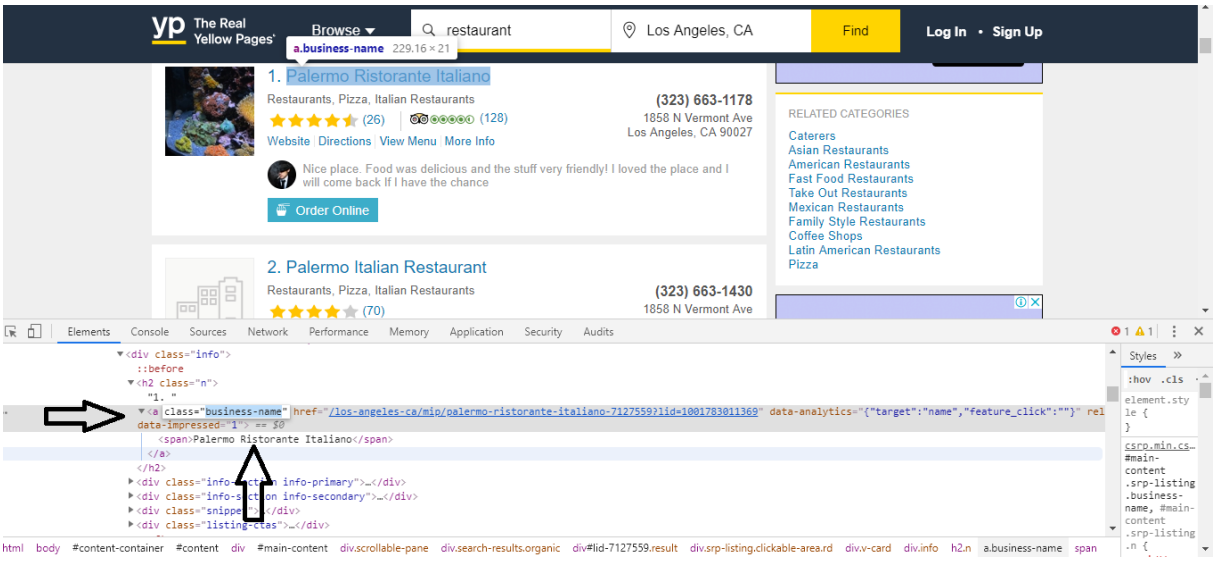



Nous pouvons accéder au code HTML d'un site Web en ouvrant les outils de développement dans votre navigateur actuel. Par exemple, Firefox (options → Développeur Web → Inspecteur). Toutes ces balises, flèches, classes et identifiants «div» bleus correspondent au code HTML du site Web sur lequel vous vous trouvez actuellement.




Avant d'examiner le code HTML pour yallowpage.com, passons en revue la structure de base à l'aide de l'exemple ci-dessous.



HTML décrit la structure sous-jacente d'un site Web. En d'autres termes, il identifie que le site Web aura un en-tête, plusieurs paragraphes, une vidéo intégrée, un pied de page de fermeture, etc. HTML ne décrira pas comment ces composants seront disposés, leur style, leur taille, leur couleur, etc.

Le code HTML est de nature hiérarchique et les balises indentées (c'est-à-dire , <ul>, etc.) identifient chaque nouveau niveau de la hiérarchie. Par exemple, la balise «<html>» se trouve en haut de la hiérarchie (également appelée conteneur) et contient tous les autres éléments. La balise body est particulièrement importante car elle contient la grande majorité des informations visibles sur un site Web. Tournons notre attention sur le code python qui explore cette structure hiérarchique de balises pour extraire le texte.

## Charger les librairies nécessaires :

Tout d'abord, nous devons importer les bibliothèques requises.

In [3]:
```
# Charger les librairie nécessaire .
from bs4 import BeautifulSoup
import lxml
import requests
import pandas as pd
import numpy as np
import nltk
import string
import fasttext
import contractions
from matplotlib import pyplot as plt
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords, wordnet
from nltk.stem import WordNetLemmatizer
#plt.xticks(rotation=70)
#pd.options.mode.chained_assignment = None
#pd.set_option('display.max_colwidth', 100)
#%matplotlib inline
#pip install fasttext_win
#pip install contractions
#nltk.download('averaged_perceptron_tagger')
```

La bibliothèque "request" importée a une fonction get () qui demandera au serveur yellowpage.com le contenu de l'URL et stockera la réponse du serveur dans la variable "base_url". Si nous imprimons la variable "base_url", nous verrons en fait l'intégralité du code HTML de la page.

Commençons par définir une fonction nommée "parse" nécessitant un argument qui sera l'URL réelle de la page que nous tentons d'analyser / scrape. Ensuite, nous utiliserons le créateur de classe BeautifulSoup pour analyser le contenu (code HTML) du site Web fourni. Nous utiliserons l'analyseur "lxml" au cas où le HTML ne serait pas parfaitement formé. Pour plus d'informations sur les différents analyseurs disponibles avec BeautifulSoup, visitez ce lien.

```
In [4]: import re
def parse(full_url):


    df = pd.DataFrame(columns =
         ['businessname', 'rating', 'ratnumber', 'comment'])

    headers = {'User-Agent':'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_2) AppleWebKit/601.3.9 (KHTML, like Gecko) Version/9.0.2 Safar
i/601.3.9'}
    response=requests.get('https://www.yellowpages.com/los-angeles-ca/restaurant',headers=headers)
    soup=BeautifulSoup(response.content,'lxml')
    for item in soup.select('.v-card'):
        try:
            businessname=item.select('.business-name')[0].get_text()
        except:
            businessname=None
        try:

            rating= item.select('.rating div')[0].get('class')
        except:
            rating=None

        try:

            ratnumber= item.select('.rating div span')[0].get_text()
        except:
            rating=None
        try:
            comment = item.find('p', {'class': 'body with-avatar'}).text.replace('\n', '')
        except:
            comment=None
        if (not  rating) or  (not comment) :
            rating=None
        else :
            df = df.append({'businessname': businessname, 'rating': rating, 'ratnumber': ratnumber, 'comment': comment}, ignore_index=Tru
e)
    return df
```

Par conséquent, nous utiliserons la méthode "findall ()" pour extraire tous les conteneurs "div" qui ont un attribut de classe "**v-card**". Ensuite, nous allons créer un dataframe pandas vide nommé "df" auquel nous ajouterons les données récupérées.

Maintenant que nous avons identifié le conteneur pour toutes les données que nous souhaitons extraire, plongeons un peu plus dans le HTML pour identifier les éléments qui hébergent les données réelles. Tout d'abord, la note d'évaluation est à nouveau hébergée dans le conteneur «**rating**», mais en explorant quelques couches, nous trouvons la balise «<div» avec un attribut de classe « **rating div span**» qui stocke en fait le« 5.0 " évaluation. Prenons note de la balise et de l'attribut de classe qui stocke ces données car nous aurons besoin de ces informations pour notre script python ci-dessous. Nous répétons le processus pour toutes les données restantes qui souhaitent extraire.

```
Entrée [6]: base_url = 'https://www.yellowpages.com/los-angeles-ca/restaurant'
            all_reviews_df = pd.DataFrame(columns = ['businessname', 'rating', 'ratnumber', 'comment'])
            num_reviews = 1

            while num_reviews < 4:

                full_url = base_url + str(num_reviews)

                get_url = requests.get(full_url, timeout=5)

                partial_reviews_df = parse(get_url)
                all_reviews_df = all_reviews_df.append(
                                 partial_reviews_df, ignore_index=True)

                num_reviews += 1
```

Nous n'avons pas encore terminé car si vous exécutez la fonction "parse ()", vous obtiendrez un dataframe avec seulement 20 enregistrements et cela est dû au fait que nous n'avons scrapper qu'une seule page.

Ensuite, nous demandons et obtenons à nouveau le code HTML complet de la page sur laquelle la boucle while est itérée. Nous appliquons notre fonction parse () sur la page et

ajoutons les commentaires récemment récupérés dans notre dataframe de données. Enfin, nous augmentons le compteur de 1  pour que la boucle while soit itérée sur la page suivante.

## Sauvegarder les résultats dans un fichier CSV :

```
In [8]: all_reviews_df.to_csv('yellowpage_scrape.csv')
```

## Pré-traitement des données texte à l'aide de Python :

La partie précendente a exploré le concept de récupération d'informations textuelles à partir de sites Web à l'aide d'une bibliothèque python nommée BeautifulSoup. Nous avons rapidement pu récupérer les notes des restaurant  sur yellowpage.com et exporter les données dans un fichier CSV local. Le scrape des données n'est que la première étape pour traiter des informations utiles à partir de nos données textuelles nouvellement acquises. Le but de cet article est de passer à l'étape suivante et d'appliquer quelques étapes de pré-traitement standard afin de préparer les données pour l'analyse.

Les méthodes de prétraitement que nous choisissons d'exécuter dépendront de nos données, du résultat souhaité et / ou de la manière dont nous choisissons  d'analyser nos données. Cela dit, les méthodes de prétraitement répertoriées ci-dessous font partie des méthodes les plus couramment utilisées.

## Les méthodes de prétraitement :

- Importing Libraries along with our Data
- Expanding Contractions
- Tokenization
- Converting all Characters to Lowercase
- Removing Punctuations
- Removing Stopwords
- Parts of Speech Tagging
- Lemmatization

```
In [9]: with open('yellowpage_scrape.csv') as f:
            df = pd.read_csv(f)
        f.close()
```

## Importer nos données :

Nous allons importer les notes de notation des clients récupérées dans notre code précédent et examiner rapidement les données.

```
Entrée [10]: for col in df.columns:
                 print(col, df[col].isnull().sum())

Unnamed: 0 0
businessname 0
rating 0
ratnumber 0
comment 0
```

Ensuite, examinons si nous avons des valeurs manquantes. Il semble que «rating» et «comment» ne contiennent aucune valeur manquante.

```
Entrée [11]: rws = df.loc[:, ['rating', 'comment']]
```

On a choisi de travailler sur les deux colonnes rating et coment

```
Entrée [12]: df.drop('Unnamed: 0', axis=1, inplace=True)
```

Commençons par supprimer la colonne "Sans nom: 0", car elle duplique simplement l'index.

## Expanding Contractions :

Les contractions sont ces petits raccourcis littéraires que nous prenons là où au lieu de «Devrait avoir», nous préférons «Devrait» ou où «Ne pas» devient rapidement «Ne pas». Nous allons ajouter une nouvelle colonne à notre dataframe appelée «no_contract» et appliquer une fonction lambda au champ «comment» qui élargira les contractions. Soyez conscient du fait que les contractions élargies seront effectivement symbolisées ensemble. En d'autres termes, «j'ai» = «j'ai» au lieu de «je», «j'ai».

Entrée [13]:
```
rws['no_contract'] = rws['comment'].apply(lambda x: [contractions.fix(word) for word in x.split()])
rws.head()
```

Out[13]:

| | rating | comment | no_contract |
|---|---|---|---|
| 0 | ['result-rating', 'four', 'half'] | Nice place. Food was delicious and the stuff very friendly! I loved the place and I will come ba... | [Nice, place., Food, was, delicious, and, the, stuff, very, friendly!, I, loved, the, place, and... |
| 1 | ['result-rating', 'four'] | It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of... | [It, was, the, best, of, times,, it, was, the, worst, of, times,, it, was, the, age, of, wisdom,... |
| 2 | ['result-rating', 'four', 'half'] | Me and my buddies had a high-quality time right here! The food changed into excellent and servic... | [Me, and, my, buddies, had, a, high-quality, time, right, here!, The, food, changed, into, excel... |
| 3 | ['result-rating', 'four', 'half'] | It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of... | [It, was, the, best, of, times,, it, was, the, worst, of, times,, it, was, the, age, of, wisdom,... |
| 4 | ['result-rating', 'four', 'half'] | The staff were very welcoming and friendly. The traditional Middle Eastern cuisine was excellent... | [The, staff, were, very, welcoming, and, friendly., The, traditional, Middle, Eastern, cuisine, ... |

En fin de compte, nous voudrions que les contractions étendues soient segmentées séparément en «I», «have», par conséquent, convertissons les listes de la colonne «no_contract» en chaînes.

Entrée [15]:
```
rws['rating_description_str'] = [' '.join(map(str, l)) for l in rws['no_contract']]
rws.head()
```

Out[15]:

| | rating | comment | no_contract | rating_description_str |
|---|---|---|---|---|
| 0 | ['result-rating', 'four', 'half'] | Nice place. Food was delicious and the stuff very friendly! I loved the place and I will come ba... | [Nice, place., Food, was, delicious, and, the, stuff, very, friendly!, I, loved, the, place, and... | Nice place. Food was delicious and the stuff very friendly! I loved the place and I will come ba... |
| 1 | ['result-rating', 'four'] | It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of... | [It, was, the, best, of, times,, it, was, the, worst, of, times,, it, was, the, age, of, wisdom,... | It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of... |
| 2 | ['result-rating', 'four', 'half'] | Me and my buddies had a high-quality time right here! The food changed into excellent and servic... | [Me, and, my, buddies, had, a, high-quality, time, right, here!, The, food, changed, into, excel... | Me and my buddies had a high-quality time right here! The food changed into excellent and servic... |
| 3 | ['result-rating', 'four', 'half'] | It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of... | [It, was, the, best, of, times,, it, was, the, worst, of, times,, it, was, the, age, of, wisdom,... | It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of... |
| 4 | ['result-rating', 'four', 'half'] | The staff were very welcoming and friendly. The traditional Middle Eastern cuisine was excellent... | [The, staff, were, very, welcoming, and, friendly., The, traditional, Middle, Eastern, cuisine, ... | The staff were very welcoming and friendly. The traditional Middle Eastern cuisine was excellent... |

## Tokenization :

Maintenant que nous avons supprimé toutes les critiques non anglaises, appliquons notre tokenizer afin de diviser chaque mot en un jeton. Nous appliquerons la fonction NLTK.word_tokenize () à la colonne «rating_description_str» et créerons une nouvelle colonne nommée «tokenized».

Entrée [16]:
```
rws['tokenized'] = rws['rating_description_str'].apply(word_tokenize)
rws.head()
```

Out[16]:

| | rating | comment | no_contract | rating_description_str | tokenized |
|---|---|---|---|---|---|
| 0 | ['result-rating', 'four', 'half'] | Nice place. Food was delicious and the stuff very friendly! I loved the place and I will come ba... | [Nice, place., Food, was, delicious, and, the, stuff, very, friendly!, I, loved, the, place, and... | Nice place. Food was delicious and the stuff very friendly! I loved the place and I will come ba... | [Nice, place, ., Food, was, delicious, and, the, stuff, very, friendly, !, I, loved, the, place,... |
| 1 | ['result-rating', 'four'] | It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of... | [It, was, the, best, of, times,, it, was, the, worst, of, times,, it, was, the, age, of, wisdom,... | It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of... | [It, was, the, best, of, times, ,, it, was, the, worst, of, times, ,, it, was, the, age, of, wis... |
| 2 | ['result-rating', 'four', 'half'] | Me and my buddies had a high-quality time right here! The food changed into excellent and servic... | [Me, and, my, buddies, had, a, high-quality, time, right, here!, The, food, changed, into, excel... | Me and my buddies had a high-quality time right here! The food changed into excellent and servic... | [Me, and, my, buddies, had, a, high-quality, time, right, here, !, The, food, changed, into, exc... |
| 3 | ['result-rating', 'four', 'half'] | It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of... | [It, was, the, best, of, times,, it, was, the, worst, of, times,, it, was, the, age, of, wisdom,... | It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of... | [It, was, the, best, of, times, ,, it, was, the, worst, of, times, ,, it, was, the, age, of, wis... |
| 4 | ['result-rating', 'four', 'half'] | The staff were very welcoming and friendly. The traditional Middle Eastern cuisine was excellent... | [The, staff, were, very, welcoming, and, friendly., The, traditional, Middle, Eastern, cuisine, ... | The staff were very welcoming and friendly. The traditional Middle Eastern cuisine was excellent... | [The, staff, were, very, welcoming, and, friendly, ., The, traditional, Middle, Eastern, cuisine... |

## Converting all Characters to Lowercase :

Transformer tous les mots en minuscules est également une étape de pré-traitement très courante. Dans ce cas, nous ajouterons une nouvelle fois une nouvelle colonne nommée «inférieure» au dataframe qui transformera tous les mots tokenisés en minuscules. Cependant, comme nous devons itérer sur plusieurs mots, nous utiliserons une simple boucle for dans une fonction lambda pour appliquer la fonction «inférieure» à chaque mot.

Entrée [17]:
```
rws['lower'] = rws['tokenized'].apply(lambda x: [word.lower() for word in x])
rws.head()
```

Out[17]:

| | rating | comment | no_contract | rating_description_str | tokenized | lower |
|---|---|---|---|---|---|---|
| 0 | ['result-rating', 'four', 'half'] | Nice place. Food was delicious and the stuff very friendly! I loved the place and I will come ba... | [Nice, place., Food, was, delicious, and, the, stuff, very, friendly!, I, loved, the, place, and... | Nice place. Food was delicious and the stuff very friendly! I loved the place and I will come ba... | [Nice, place, ., Food, was, delicious, and, the, stuff, very, friendly, !, I, loved, the, place,... | [nice, place, ., food, was, delicious, and, the, stuff, very, friendly, !, i, loved, the, place,... |
| 1 | ['result-rating', 'four'] | It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of... | [It, was, the, best, of, times,, it, was, the, worst, of, times,, it, was, the, age, of, wisdom,... | It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of... | [It, was, the, best, of, times, ,, it, was, the, worst, of, times, ,, it, was, the, age, of, wis... | [it, was, the, best, of, times, ,, it, was, the, worst, of, times, ,, it, was, the, age, of, wis... |
| 2 | ['result-rating', 'four', 'half'] | Me and my buddies had a high-quality time right here! The food changed into excellent and servic... | [Me, and, my, buddies, had, a, high-quality, time, right, here!, The, food, changed, into, excel... | Me and my buddies had a high-quality time right here! The food changed into excellent and servic... | [Me, and, my, buddies, had, a, high-quality, time, right, here, !, The, food, changed, into, exc... | [me, and, my, buddies, had, a, high-quality, time, right, here, !, the, food, changed, into, exc... |
| 3 | ['result-rating', 'four', 'half'] | It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of... | [It, was, the, best, of, times,, it, was, the, worst, of, times,, it, was, the, age, of, wisdom,... | It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of... | [It, was, the, best, of, times, ,, it, was, the, worst, of, times, ,, it, was, the, age, of, wis... | [it, was, the, best, of, times, ,, it, was, the, worst, of, times, ,, it, was, the, age, of, wis... |
| 4 | ['result-rating', 'four', 'half'] | The staff were very welcoming and friendly. The traditional Middle Eastern cuisine was excellent... | [The, staff, were, very, welcoming, and, friendly., The, traditional, Middle, Eastern, cuisine, ... | The staff were very welcoming and friendly. The traditional Middle Eastern cuisine was excellent... | [The, staff, were, very, welcoming, and, friendly, ., The, traditional, Middle, Eastern, cuisine... | [the, staff, were, very, welcoming, and, friendly, ., the, traditional, middle, eastern, cuisine... |

## Removing Punctuations :

La ponctuation est souvent supprimée de notre corpus car elle n'a que peu de valeur une fois que nous commençons à analyser nos données. En continuant le modèle précédent, nous allons créer une nouvelle colonne dont la ponctuation est supprimée. Nous utiliserons à nouveau une boucle for dans une fonction lambda pour parcourir les jetons, mais cette fois en utilisant une condition IF pour afficher uniquement les caractères alpha. Cela peut être un peu difficile à voir, mais la «période» symbolique dans la colonne «inférieure» a été supprimée.

Entrée [18]:
```
punc = string.punctuation
rws['no_punc'] = rws['lower'].apply(lambda x: [word for word in x if word not in punc])
rws.head()
```

Out[18]:

| | rating | comment | no_contract | rating_description_str | tokenized | lower | no_punc |
|---|---|---|---|---|---|---|---|
| 0 | ['result-rating', 'four', 'half'] | Nice place. Food was delicious and the stuff very friendly! I loved the place and I will come ba... | [Nice, place., Food, was, delicious, and, the, stuff, very, friendly!, I, loved, the, place, and... | Nice place. Food was delicious and the stuff very friendly! I loved the place and I will come ba... | [Nice, place, ., Food, was, delicious, and, the, stuff, very, friendly, !, I, loved, the, place,... | [nice, place, ., food, was, delicious, and, the, stuff, very, friendly, !, i, loved, the, place,... | [nice, place, food, was, delicious, and, the, stuff, very, friendly, i, loved, the, place, and, ... |
| 1 | ['result-rating', 'four'] | It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of... | [It, was, the, best, of, times,, it, was, the, worst, of, times,, it, was, the, age, of, wisdom,... | It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of... | [It, was, the, best, of, times, ,, it, was, the, worst, of, times, ,, it, was, the, age, of, wis... | [it, was, the, best, of, times, ,, it, was, the, worst, of, times, ,, it, was, the, age, of, wis... | [it, was, the, best, of, times, it, was, the, worst, of, times, it, was, the, age, of, wisdom, i... |
| 2 | ['result-rating', 'four', 'half'] | Me and my buddies had a high-quality time right here! The food changed into excellent and servic... | [Me, and, my, buddies, had, a, high-quality, time, right, here!, The, food, changed, into, excel... | Me and my buddies had a high-quality time right here! The food changed into excellent and servic... | [Me, and, my, buddies, had, a, high-quality, time, right, here, !, The, food, changed, into, exc... | [me, and, my, buddies, had, a, high-quality, time, right, here, !, the, food, changed, into, exc... | [me, and, my, buddies, had, a, high-quality, time, right, here, the, food, changed, into, excell... |
| 3 | ['result-rating', 'four', 'half'] | It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of... | [It, was, the, best, of, times,, it, was, the, worst, of, times,, it, was, the, age, of, wisdom,... | It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of... | [It, was, the, best, of, times, ,, it, was, the, worst, of, times, ,, it, was, the, age, of, wis... | [it, was, the, best, of, times, ,, it, was, the, worst, of, times, ,, it, was, the, age, of, wis... | [it, was, the, best, of, times, it, was, the, worst, of, times, it, was, the, age, of, wisdom, i... |
| 4 | ['result-rating', 'four', 'half'] | The staff were very welcoming and friendly. The traditional Middle Eastern cuisine was excellent... | [The, staff, were, very, welcoming, and, friendly., The, traditional, Middle, Eastern, cuisine, ... | The staff were very welcoming and friendly. The traditional Middle Eastern cuisine was excellent... | [The, staff, were, very, welcoming, and, friendly, ., The, traditional, Middle, Eastern, cuisine... | [the, staff, were, very, welcoming, and, friendly, ., the, traditional, middle, eastern, cuisine... | [the, staff, were, very, welcoming, and, friendly, the, traditional, middle, eastern, cuisine, w... |

## Removing Stopwords :

Stopwords  sont généralement des mots inutiles et n'ajoutent pas beaucoup de sens à une phrase. En anglais, les mots vides courants incluent «“you, he, she, in, a, has, are, etc.”Tout d'abord, nous devons importer la bibliothèque de mots vides NLTK et définir nos mots vides sur «anglais». Nous allons ajouter une nouvelle colonne «no_stopwords» qui supprimera les mots vides de la colonne «no_punc» car elle a été tokenisée, a été convertie en minuscules et la ponctuation a été supprimée. Une fois de plus, une boucle for dans une fonction lambda effectuera une itération sur les jetons de «no_punc» et ne retournera que les jetons qui n'existent pas dans notre variable «stop_words».

Entrée [19]:
```
stop_words = set(stopwords.words('english'))
rws['stopwords_removed'] = rws['no_punc'].apply(lambda x: [word for word in x if word not in stop_words])
rws.head()
```

Out[19]:

| | rating | comment | no_contract | rating_description_str | tokenized | lower | no_punc | stopwords_removed |
|---|---|---|---|---|---|---|---|---|
| 0 | ['result-rating', 'four', 'half'] | Nice place. Food was delicious and the stuff very friendly! I loved the place and I will come ba... | [Nice, place., Food, was, delicious, and, the, stuff, very, friendly!, I, loved, the, place, and... | Nice place. Food was delicious and the stuff very friendly! I loved the place and I will come ba... | [Nice, place, ., Food, was, delicious, and, the, stuff, very, friendly, !, I, loved, the, place,... | [nice, place, ., food, was, delicious, and, the, stuff, very, friendly, !, i, loved, the, place,... | [nice, place, food, was, delicious, and, the, stuff, very, friendly, i, loved, the, place, and, ... | [nice, place, food, delicious, stuff, friendly, loved, place, come, back, chance] |
| 1 | ['result-rating', 'four'] | It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of... | [It, was, the, best, of, times,, it, was, the, worst, of, times,, it, was, the, age, of, wisdom,... | It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of... | [It, was, the, best, of, times, ,, it, was, the, worst, of, times, ,, it, was, the, age, of, wis... | [it, was, the, best, of, times, ,, it, was, the, worst, of, times, ,, it, was, the, age, of, wis... | [it, was, the, best, of, times, it, was, the, worst, of, times, it, was, the, age, of, wisdom, i... | [best, times, worst, times, age, wisdom, age, foolishness, ...] |
| 2 | ['result-rating', 'four', 'half'] | Me and my buddies had a high-quality time right here! The food changed into excellent and servic... | [Me, and, my, buddies, had, a, high-quality, time, right, here!, The, food, changed, into, excel... | Me and my buddies had a high-quality time right here! The food changed into excellent and servic... | [Me, and, my, buddies, had, a, high-quality, time, right, here, !, The, food, changed, into, exc... | [me, and, my, buddies, had, a, high-quality, time, right, here, !, the, food, changed, into, exc... | [me, and, my, buddies, had, a, high-quality, time, right, here, the, food, changed, into, excell... | [buddies, high-quality, time, right, food, changed, excellent, service, high, exceptional] |
| 3 | ['result-rating', 'four', 'half'] | It was the best of times, it was the worst of times, it was the age of wisdom, it was the | [It, was, the, best, of, times,, it, was, the, worst, of, times,, it, was, the, age, of, wisdom,... | It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of... | [It, was, the, best, of, times, ,, it, was, the, worst, of, times, ,, it, was, the, age, of, wis... | [it, was, the, best, of, times, ,, it, was, the, worst, of, times, ,, it, was, the, age, of, | [it, was, the, best, of, times, it, was, the, worst, of, times, it, was, the, age, of, wisdom, | [best, times, worst, times, age, wisdom, age, foolishness, ...] |

## Parts of Speech Tagging :

```
Entrée [20]: rws['pos_tags'] = rws['stopwords_removed'].apply(nltk.tag.pos_tag)
             rws.head()
```

Out[20]:

| | rating | comment | no_contract | rating_description_str | tokenized | lower | no_punc | stopwords_removed | pos_tags |
|---|---|---|---|---|---|---|---|---|---|
| 0 | ['result-rating', 'four', 'half'] | Nice place. Food was delicious and the stuff very friendly! I loved the place and I will come ba... | [Nice, place., Food, was, delicious, and, the, stuff, very, friendly!, I, loved, the, place, and... | Nice place. Food was delicious and the stuff very friendly! I loved the place and I will come ba... | [Nice, place, ., Food, was, delicious, and, the, stuff, very, friendly, !, I, loved, the, place,... | [nice, place, ., food, was, delicious, and, the, stuff, very, friendly, !, i, loved, the, place,... | [nice, place, food, was, delicious, and, the, stuff, very, friendly, i, loved, the, place, and, ... | [nice, place, food, delicious, stuff, friendly, loved, place, come, back, chance] | [(nice, JJ), (place, NN), (food, NN), (delicious, JJ), (stuff, NN), (friendly, RB), (loved, VBD)... |
| 1 | ['result-rating', 'four'] | It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of... | [It, was, the, best, of, times,, it, was, the, worst, of, times,, it, was, the, age, of, wisdom,... | It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of... | [It, was, the, best, of, times, ,, it, was, the, worst, of, times, ,, it, was, the, age, of, wis... | [it, was, the, best, of, times, ,, it, was, the, worst, of, times, ,, it, was, the, age, of, wis... | [it, was, the, best, of, times, it, was, the, worst, of, times, it, was, the, age, of, wisdom, i... | [best, times, worst, times, age, wisdom, age, foolishness, ...] | [(best, RBS), (times, NNS), (worst, RB), (times, NNS), (age, NN), (wisdom, NN), (age, NN), (fool... |
| 2 | ['result-rating', 'four', 'half'] | Me and my buddies had a high-quality time right here! The food changed into excellent and servic... | [Me, and, my, buddies, had, a, high-quality, time, right, here!, The, food, changed, into, excel... | Me and my buddies had a high-quality time right here! The food changed into excellent and servic... | [Me, and, my, buddies, had, a, high-quality, time, right, here, !, The, food, changed, into, exc... | [me, and, my, buddies, had, a, high-quality, time, right, here, !, the, food, changed, into, exc... | [me, and, my, buddies, had, a, high-quality, time, right, here, the, food, changed, into, excell... | [buddies, high-quality, time, right, food, changed, excellent, service, high, exceptional] | [(buddies, NNS), (high-quality, JJ), (time, NN), (right, JJ), (food, NN), (changed, VBD), (excel... |

L'idée de la racine est de réduire les différentes formes d'utilisation du mot dans son mot racine. Par exemple, “drive”, “drove”, “driving”, “driven”, “driver” sont des dérivés du mot «drive» et très souvent les chercheurs souhaitent supprimer cette variabilité de leur corpus. Par rapport à la lemmatisation, la radicalisation est certainement la méthode la moins compliquée, mais elle ne produit souvent pas de racine morphologique spécifique au dictionnaire du mot. En d'autres termes, la racine du mot «pies» produira souvent une racine de «pi» alors que la lemmatisation trouvera la racine morphologique de «pies».

Au lieu de prendre la solution la plus simple avec la racine, appliquons la lemmatisation à nos données, mais cela nécessite quelques étapes supplémentaires par rapport à la racine.

Premièrement, nous devons appliquer des parties de balises vocales, en d'autres termes, déterminer la partie du discours (ie. noun, verb, adverb, etc.) pour chaque mot.

## Stemming vs Lemmatization :

Nous allons utiliser le Lemmatization de mots de NLTK qui nécessite que les parties des balises vocales soient converties au format wordnet. Nous allons écrire une fonction qui effectue la conversion appropriée, puis utiliser la fonction dans une compréhension de liste pour appliquer la conversion. Enfin, nous appliquons le mot lemmatizer de NLTK.

Entrée [21]:

```
def get_wordnet_pos(tag):
    if tag.startswith('J'):
        return wordnet.ADJ
    elif tag.startswith('V'):
        return wordnet.VERB
    elif tag.startswith('N'):
        return wordnet.NOUN
    elif tag.startswith('R'):
        return wordnet.ADV
    else:
        return wordnet.NOUN
rws['wordnet_pos'] = rws['pos_tags'].apply(lambda x: [(word, get_wordnet_pos(pos_tag)) for (word, pos_tag) in x])
rws.head()
```

Out[21]:

| | rating | comment | no_contract | rating_description_str | tokenized | lower | no_punc | stopwords_removed | pos_tags | wordnet_pos |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | ['result-rating', 'four', 'half'] | Nice place. Food was delicious and the stuff very friendly! I loved the place and I will come ba... | [Nice, place., Food, was, delicious, and, the, stuff, very, friendly!, I, loved, the, place, and... | Nice place. Food was delicious and the stuff very friendly! I loved the place and I will come ba... | [Nice, place, ., Food, was, delicious, and, the, stuff, very, friendly, !, I, loved, the, place,... | [nice, place, ., food, was, delicious, and, the, stuff, very, friendly, !, i, loved, the, place,... | [nice, place, food, was, delicious, and, the, stuff, very, friendly, i, loved, the, place, and, ... | [nice, place, food, delicious, stuff, friendly, loved, place, come, back, chance] | [(nice, JJ), (place, NN), (food, NN), (delicious, JJ), (stuff, NN), (friendly, RB), (loved, VBD)... | [(nice, a), (place, n), (food, n), (delicious, a), (stuff, n), (friendly, r), (loved, v), (place... |
| 1 | ['result-rating', 'four'] | It was the best of times, it was the worst of times, it was the age of wisdom, it was the age | [It, was, the, best, of, times,, it, was, the, worst, of, times,, it, was, the, age, of, wisdom.... | It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of... | [It, was, the, best, of, times, ,, it, was, the, worst, of, times, ,, it, was, the, | [it, was, the, best, of, times, ,, it, was, the, worst, of, times, ,, it, was, the, | [it, was, the, best, of, times, it, was, the, worst, of, times, it, was, the, age, of, wisdom. i... | [best, times, worst, times, age, wisdom, age, foolishness, ...] | [(best, RBS), (times, NNS), (worst, RB), (times, NNS), (age, NN), (wisdom, NN), (age, NN), | [(best, r), (times, n), (worst, r), (times, n), (age, n), (wisdom, n), (age, n), (foolishness, |

Nous pouvons maintenant appliquer le Lemmatization de mots de NLTK dans notre compréhension de liste fidèle. Remarquez que la fonction lemmatizer nécessite deux paramètres le mot et sa balise (sous forme wordnet).

Entrée [22]:

```
wnl = WordNetLemmatizer()
rws['lemmatized'] = rws['wordnet_pos'].apply(lambda x: [wnl.lemmatize(word, tag) for word, tag in x])
rws.head()
```

Out[22]:

| | rating | comment | no_contract | rating_description_str | tokenized | lower | no_punc | stopwords_removed | pos_tags | wordnet_pos | lemmatized |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | ['result-rating', 'four', 'half'] | Nice place. Food was delicious and the stuff very friendly! I loved the place and I will come ba... | [Nice, place., Food, was, delicious, and, the, stuff, very, friendly!, I, loved, the, place, and... | Nice place. Food was delicious and the stuff very friendly! I loved the place and I will come ba... | [Nice, place, ., Food, was, delicious, and, the, stuff, very, friendly, !, I, loved, the, place,... | [nice, place, ., food, was, delicious, and, the, stuff, very, friendly, !, i, loved, the, place,... | [nice, place, food, was, delicious, and, the, stuff, very, friendly, i, loved, the, place, and, ... | [nice, place, food, delicious, stuff, friendly, loved, place, come, back, chance] | [(nice, JJ), (place, NN), (food, NN), (delicious, JJ), (stuff, NN), (friendly, RB), (loved, VBD)... | [(nice, a), (place, n), (food, n), (delicious, a), (stuff, n), (friendly, r), (loved, v), (place... | [nice, place, food, delicious, stuff, friendly, love, place, come, back, chance] |
| 1 | ['result-rating', 'four'] | It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of... | [It, was, the, best, of, times,, it, was, the, worst, of, times,, it, was, the, age, of, wisdom,... | It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of... | [It, was, the, best, of, times, ,, it, was, the, worst, of, times, ,, it, was, the, age, of, wis... | [it, was, the, best, of, times, ,, it, was, the, worst, of, times, ,, it, was, the, age, of, wis... | [it, was, the, best, of, times, it, was, the, worst, of, times, it, was, the, age, of, wisdom, i... | [best, times, worst, times, age, wisdom, age, foolishness, ...] | [(best, RBS), (times, NNS), (worst, RB), (times, NNS), (age, NN), (wisdom, NN), (age, NN), (fool... | [(best, r), (times, n), (worst, r), (times, n), (age, n), (wisdom, n), (age, n), (foolishness, n... | [best, time, worst, time, age, wisdom, age, foolishness, ... |
| 2 | ['result-rating', 'four', 'half'] | Me and my buddies had a high-quality time right here! The food changed | [Me, and, my, buddies, had, a, high-quality, time, right, here!, The, food, | Me and my buddies had a high-quality time right here! The food changed into excellent and | [Me, and, my, buddies, had, a, high-quality, time, right, here, !, | [me, and, my, buddies, had, a, high-quality, time, right, here, !, the, | [me, and, my, buddies, had, a, high-quality, time, right, here, the, | [buddies, high-quality, time, right, food, changed, excellent, service, high, | [(buddies, NNS), (high-quality, JJ), (time, NN), (right, JJ), (food, NN), | [(buddies, n), (high-quality, a), (time, n), (right, a), (food, n), (changed, v), | [buddy, high-quality, time, right, food, change, excellent, service, |

## Exporter le résultat final dans un CSV

Entrée [23]:

```
rws.to_csv('yellowpage_scrape_clean.csv')
```