**MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE**
**DIRECTION GENERALE DES ETUDES TECHNOLOGIQUES**

**Institut Supérieur des Études Technologiques de Djerba**
**Département : Technologies de l'Informatique**

# Projet de fin d'études

En vue d'obtention du diplôme de Mastère en Technologies de l'Information et Communication et Innovation Touristique(TICIT)

# Analyse des sentiments des clients pour améliorer les services hôteliers

Élaboré par : **Firas Bahri**

Encadré par : **Mr. Blaghgi Mejdi**
Organisme d'accueil :

Année Universitaire : 2021/2022

Sommaire

# TABLE DES MATIERES

# INTRODUCTION GENERALE

L’intelligence artificielle (IA) est le grand sujet à la mode. Les possibilités de son application dans le tourisme sont quasi illimitées – de l’hyperpersonnalisation des services au pilotage des équipements, en passant par le guidage ou la production automatique de contenus éditoriaux.

Si l’utilisation de l’IA est encore quasi expérimentale dans le tourisme, les acteurs du secteur doivent anticiper afin de ne pas passer à côté de ce qui s’annonce être une véritable révolution technologique.

Au fil des ans, l'influence de l'intelligence artificielle (IA) s'est étendue à presque tous les aspects de l'industrie du voyage et de l'hospitalité.

La prolifération de l'IA dans l'industrie du voyage et de l'accueil peut être attribuée à l'énorme quantité de données générées aujourd'hui. L'IA permet d'analyser des données provenant de sources évidentes, apporte une valeur ajoutée en assimilant des modèles d'images, de voix, de vidéos et de textes, et les transforme en informations significatives et exploitables pour la prise de décision.

Les tendances, les valeurs aberrantes et les modèles sont déterminés à l'aide d'algorithmes basés sur l'apprentissage machine qui aident à guider une entreprise de voyage ou d'hôtellerie à prendre des décisions éclairées.

Les remises, les programmes, les forfaits, les saisons et les voyageurs à cibler sont formulés à l'aide de ces données intelligentes combinées à la science du comportement et à l'attribution de médias sociaux pour connaître le comportement et les connaissances des clients.

C’est dans ce contexte qu’on se propose dans ce projet de réaliser une solution d’analyse des sentiments des voyageurs et touristes.

Ce rapport est composé de 3 chapitres :

- Le premier sera consacré à la presentation du context à savoir les avis des consommateurs en general et dans le domaine touristique en particulier.
- Le deuxième est destinée à avoir une idée sur les différentes notions technologiques en relation avec la solution propose
- Le dernier sera une presentation des interfaces et du code source utilize.

# CHAPITRE 1: ETUDE PREALABLE

# Introduction

Quand on parle réputation des marques et retours d'expérience des clients sur le web, on pense souvent *aux* réseaux sociaux. Or, pour de nombreux secteurs et produits, cela se passe aussi ou surtout ailleurs, sur les forums généralistes et spécialisés et sur les sites d'avis clients. 88% des consommateurs consultent des avis en ligne avant de passer à l'acte d'achat. En Europe, la majorité des clients déclarent que leurs décisions d'achat dépendent beaucoup des avis négatifs.

# 1. Importance des avis clients

La stratégie marketing doit répondre à certains enjeux essentiels comme :

- Mieux comprendre l'audience cible
- Analyser les perceptions des clients
- Piloter un lancement de produit
- Innover régulièrement
- Optimiser l'expérience client et les services
- Améliorer les fonctionnalités produits
- Alimenter le service de relation client

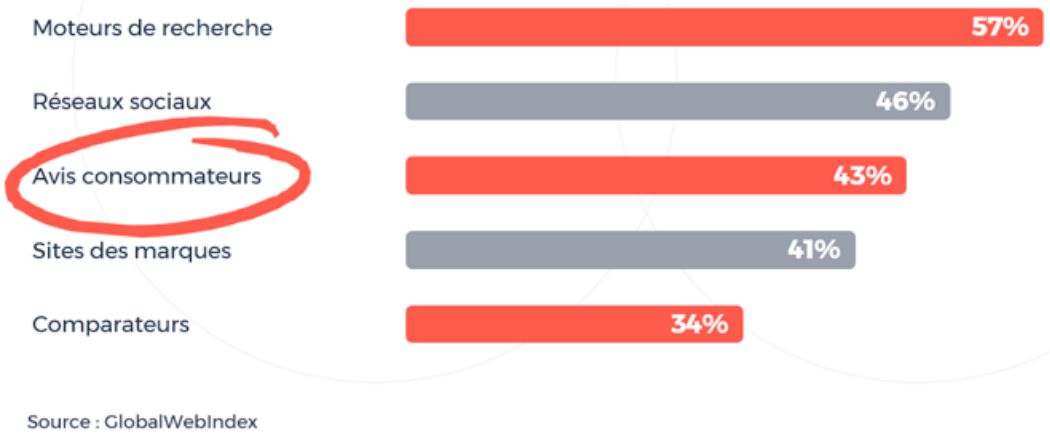
Combinée à la collecte des data sur les média sociaux, l'analyse des avis des consommateurs permet de mieux répondre à ces différents enjeux via une meilleure compréhension des attentes et comportements, des centres d'intérêts, des sentiments et émotions, des intentions d'achats, des freins à l'achat et bien sûr des feedback sur les produits, des plaintes et des commentaires.

Présents sur de nombreux espaces de conversation, les avis des consommateurs sont importants à capter et analyser car les clients et prospects sont très actifs autour de ces messages.

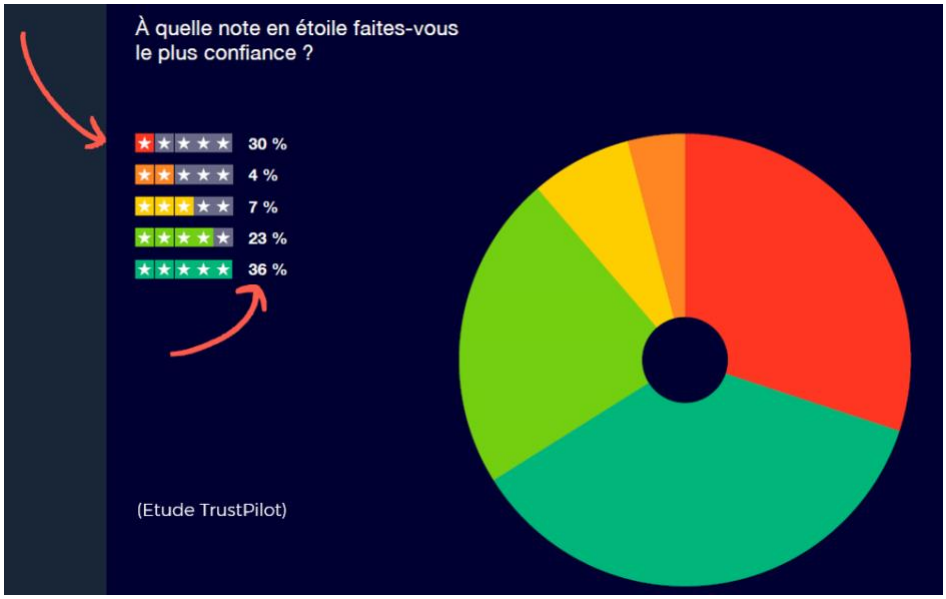
# 2. Statistiques sue les avis clients

Voici quelques statistiques :

- Les avis des consommateurs constituent en effet le 3ème canal de recherche d'information sur un produit ou service (43% des internautes) derrière les moteurs de recherche et les réseaux sociaux d'après Global Web Index.
- Dans le Top 5 des canaux de recherche d'information produit les avis consommateurs se trouvent en 3ème place



- 88% des consommateurs consultent des avis en ligne avant de passer à l'acte d'achat et 75% des consommateurs déclarent faire confiance aux appréciations déposées par les autres internautes. (Ifop-Reputation VIP).
- 55% des français ont utilisé Internet pour publier des avis sur les entreprises, marques ou dirigeants (Ifop pour Havas Paris - AD 2019). Par exemple dans l'univers Food, 50% des 18-24 ans partagent des photos de plats tandis que 39% des 18-35 ans donnent leur avis sur les marques et les produits alimentaires (Kantar TNS).
- Une étude réalisée sur 550 000 commentaires publiés sur 4 plateformes de e-commerce leaders en Europe (produits d'électronique, d'électroménager, de jardinage et de bricolage) révèle que les avis sont très souvent positifs (64% des commentaires atteignent un score de 5 étoiles), mais que les avis négatifs (1 étoile) sont les plus valorisés, c'est-à-dire signalés comme « intéressants » par les internautes (Linkproved 2019).
- Les notes considérées comme les plus fiables sont soit 5 étoiles (36 %), soit 1 étoile (30 %) c'est à dire les extrêmes (Trustpilot 2019)




- C'est ce que confirme aussi une autre étude : En Europe, la majorité des clients déclarent que leurs décisions d'achat dépendent beaucoup des avis négatifs. En fait, 71 % des clients considèrent ces avis négatifs (1 et 2 étoiles) plus utiles que les avis positifs (4 et 5 étoiles). En effet, les notes considérées comme les plus fiables sont soit 5 étoiles (36 %), soit 1 étoile (30 %) c'est à dire les extrêmes (Trustpilot 2019).

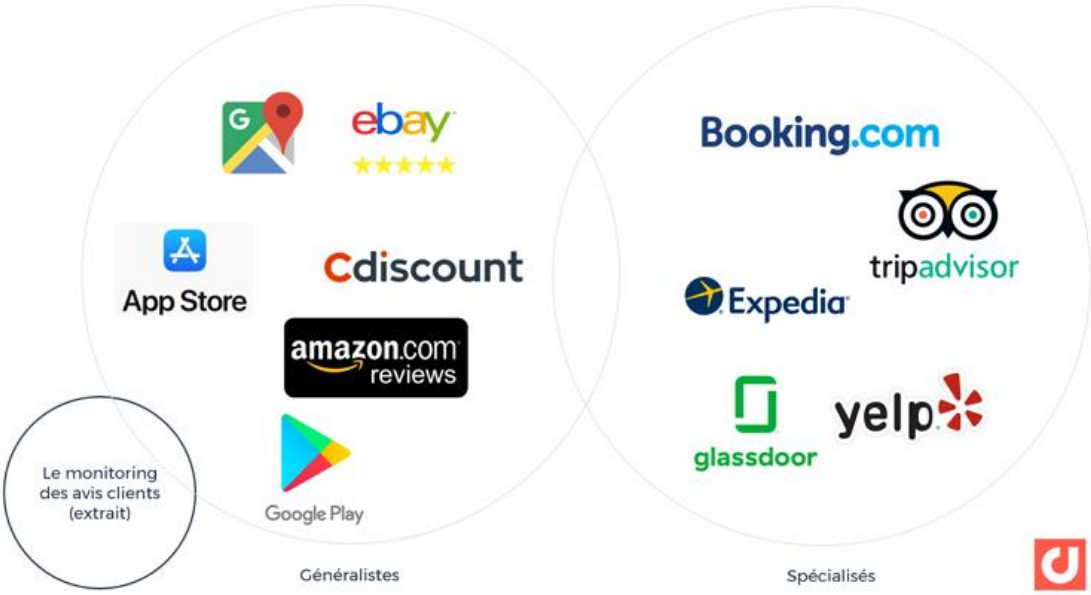
## 3. Avis à observer

Au-delà des réseaux sociaux généralistes tels que Facebook, Instagram, Twitter, WeChat, Pinterest ou encore LinkedIn, il existe des sites d'avis consommateurs clés comme Trip Advisor pour l'hôtellerie et la restauration, Yelp pour les commerces ou encore Glassdoor pour la marque employeur, etc.

En fait, ces sites spécialisés ne constituent qu'une partie des avis des consommateurs. Des centaines de milliers d'avis clients sont présents sur des sites et plateformes généralistes (comme Amazon, eBay, l'App Store, Google Play, Google Maps...etc.), autant de sites disponibles dans nos outils de social media listening.

Recueillir puis analyser les avis sur Google Maps est par exemple devenu essentiel pour les hôtels, les commerces, les restaurants et tout autre service de retail, du garage en passant par le coiffeur jusqu'à l'aéroport voire le musée.

Les avis des consommateurs représentent désormais un enjeu crucial pour les marques et un levier majeur d'avantages compétitifs au sein du processus d'achat. Ainsi, lorsque TripAdvisor met en place des options payantes - les cafés et restaurants peuvent payer pour mettre en avant les avis positifs de leur choix- cela fait débat. L'influence des avis clients est potentiellement énorme. En effet, il y a quelque temps, un responsable du groupe hôtelier Accor affirmait : "1 point sur TripAdvisor, c'est 10% de prix de chambre en plus pour un hôtel, c'est du vrai argent ces 10%, ce n'est pas juste une note pour faire beau, cela fait partie de la perception et de l'image de l'hôtel" (Vivek Badrinath, alors directeur général adjoint du groupe Accor).

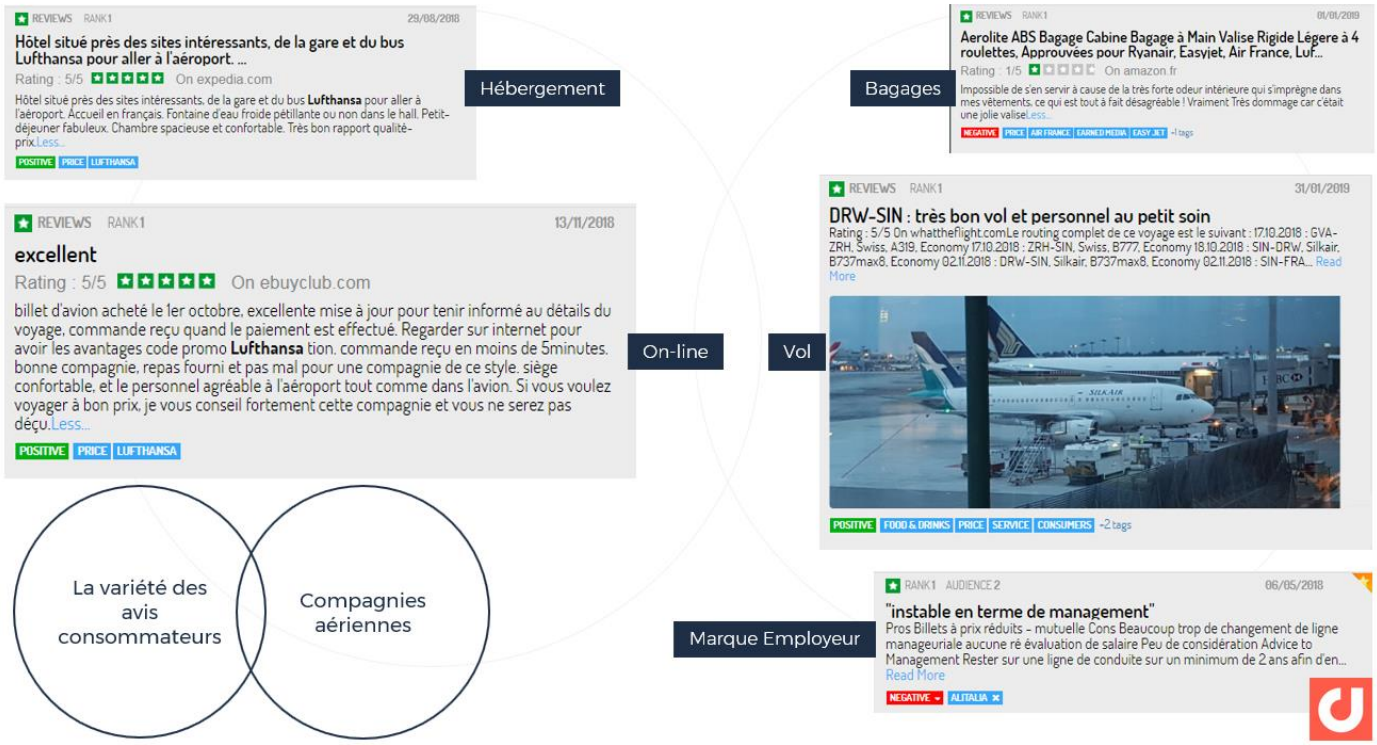



Le monitoring des avis clients via le social media listening (extrait)

Les avis des clients représentent donc un canal incontournable pour accéder à des retours d'expérience client, des suggestions et idées de consommateurs. Cela permet de collecter des insights précieux et très diversifiés sur les produits et services de vos marques et de pouvoir se comparer à ses concurrents.

Par exemple, dans le secteur des compagnies aériennes, les avis consommateurs concernent des sujets très diversifiés comme:

- L'expérience on-line (navigation, réservation, conseil, comparatifs, échanges)
- L'expérience d'hébergement
- L'expérience de vol
- Le choix des bagages
- La marque employeur
- ...



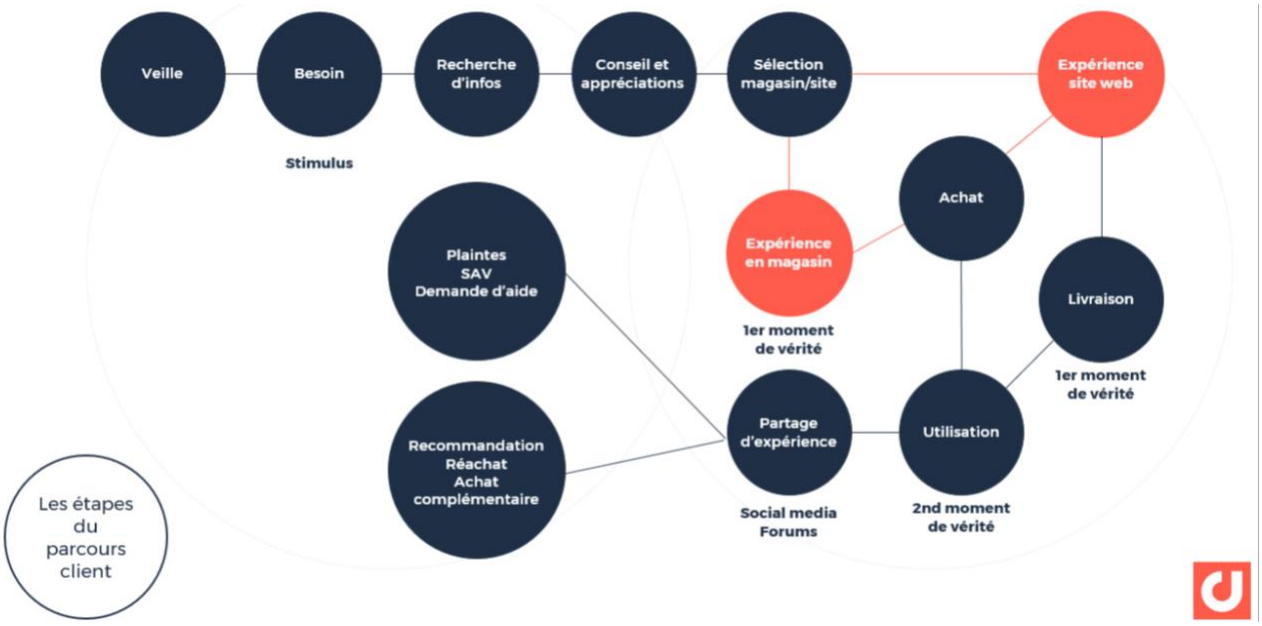

La variété des avis consommateurs : secteur des compagnies aériennes

En fait, des centaines de sites spécialisés recueillent les feedbacks des clients. Ainsi, pour tout ce qui touche à un voyage en avion (du choix des bagages jusqu'au vol en passant par l'hôtel), les espaces hébergeant des avis clients sont très nombreux. En voici un extrait : airlinequality.com, amazon.fr, booking.com, cdiscount.com, expedia.com, …

## 4. Analyser des avis des consommateurs

Il conviendra de mettre en place une analyse en continu des avis consommateurs et des avis clients détectés. Cette analyse doit permettre notamment :

- De compléter, confirmer ou infirmer les insights issus d'études consommateurs ou de focus group.
- De perfectionner l'information de votre service client en identifiant les actuelles ou futures thématiques émergentes de satisfaction, plaintes et points de friction.
- De collecter des informations spontanées en temps réel catégorisée par thématiques
- D'améliorer l'expérience utilisateur via la qualification et le renseignement des différentes étapes du parcours client.
- De multiplier la collecte de data pour l'optimisation des campagnes, le développement produit ou l'adaptation de votre offre.

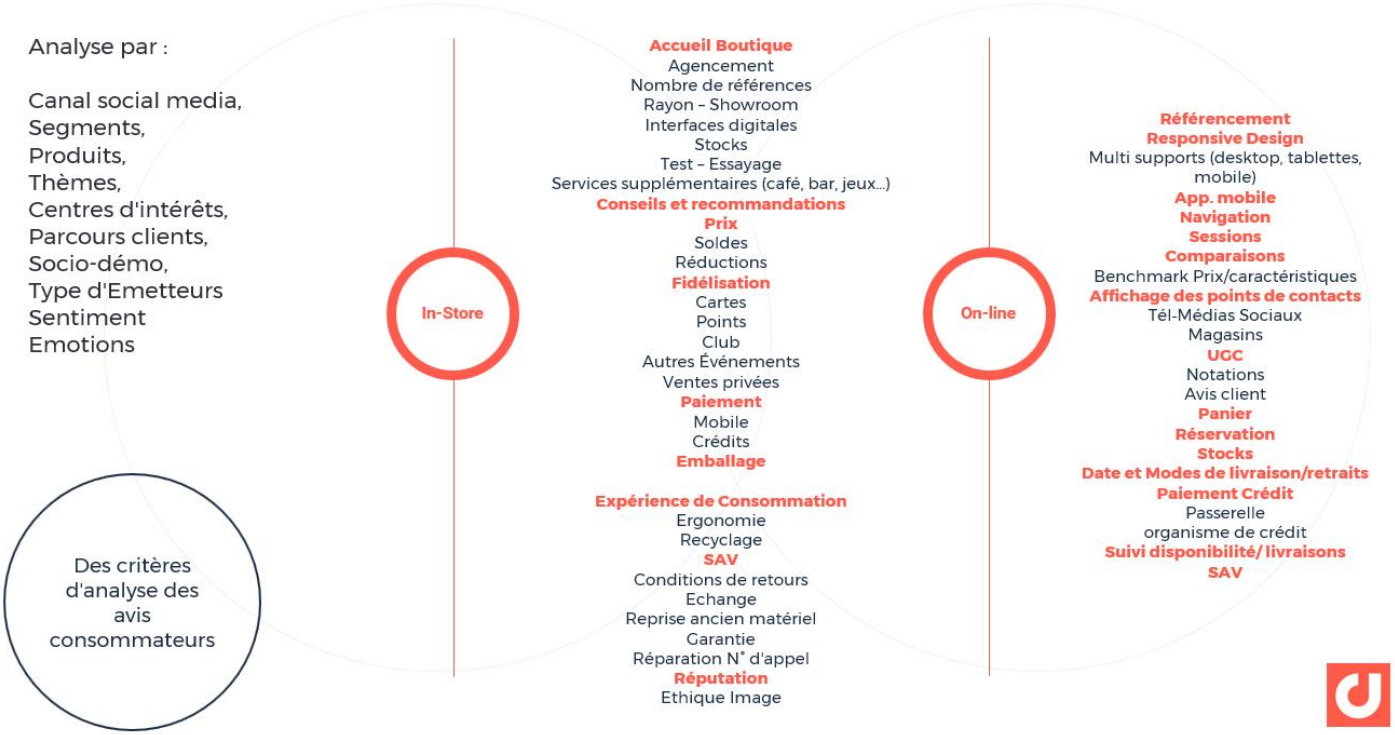



Les étapes du parcours client : un axe essentiel pour l'analyse des avis clients

## 4.2 Critères d'analyse

Pour faciliter la transformation de ces données brutes d'avis consommateurs en insights exploitables, il faut faciliter l'analyse en les segmentant par étapes du parcours client puis en thématiques spécifiques à votre secteur.

Il convient de croiser les étapes du parcours client avec les thèmes du secteur souhaité et les analyses plus classiques par canal/lieux, par satisfaction (sentiments, émotions) et critères socio-démographiques (Sexe, Age, Centre d'intérêts).




Croiser les critères d'analyses avec les thèmes liés au parcours client : l'exemple du Retail

Le parcours d'achat des consommateurs est devenu de plus en plus complexe avec le développement d'internet. Ils se renseignent en ligne puis achètent en magasin ou inversement, donnent leurs avis sur les réseaux sociaux et utilisent leur ordinateur ou leur smartphone pour se renseigner ou acheter en ligne.

Ceci entraine une explosion de la quantité de données à traiter, caractérisée par le terme Big Data. Ces données, en plus d'être massives, sont de plus en plus non structurées (images, avis clients, posts sur les réseaux sociaux …). Contrairement aux données structurées, pouvant être classées par variables (âge, chiffre d'affaires …), cela n'est pas possible pour celles non structurées rendant plus difficile leur traitement et leur analyse.

Ces deux tendances, l'augmentation du nombre de données à traiter et leur nature de moins en moins structurée, font qu'il est de plus en plus difficile pour un service marketing de les analyser sans utiliser des outils adaptés.

# 5. Solution proposée

Dans les dernières années, de nombreuses entreprises ont investi massivement dans le domaine de l'intelligence artificielle telles que Facebook, Microsoft et Google pour ne mentionner que ceux-ci. Ce concept qui a vu naître durant le 20e siècle au même moment que l'ordinateur est maintenant, en 2021, de plus en plus utilisé dans le domaine du marketing. En effet, l'IA fait partie intégrante des stratégies de marketing dès aujourd'hui. La question est de savoir comment l'utilise-t-on.

Dans le contexte du marketing digital, l'abondance de donnée à disposition des entreprises permet au marketeur d'utiliser l'intelligence afin de personnaliser sur une très grande échelle l'offre et de l'expérience client. Effectivement, ce genre de tache peut facilement être effectué par une machine comparativement à un des individus.

D'ailleurs l'IA permet au marketeur de développer des modèles prédictifs de plus en plus précis et de faire des simulations poussées de stratégie marketing à un moindre coup que dans le passé. Google Analytics est un outil connu de tout le monde dans le domaine et fonctionne justement en partie grâce à l'IA.

En ce qui concerne l'impact, IA permettra d'accroître considérablement l'efficacité au niveau du ciblage d'audience et les recommandations de produit e l'analyse des sentiments. De plus, d'un point de vue stratégique il sera difficile pour l'intelligence artificiel de remplacer le rôle des marketeurs toutefois son apport permettra a ceux-ci de se concentrer davantage sur l'aspect créatif du métier.

Enfin, ce virage dans le secteur du marketing est inévitable et prendra définitivement de l'ampleur dans le futur.

# Conclusion

La solution proposée se basera essentiellement sur la technologie de l'IA en vue de mieux connaitre les comportements des consommateurs et plus particulièrement l'aspect sentimental qui sera présentée dans le chapitres suivants.

# CHAPITRE 2: ETUDE DE LA SOLUTION PROPOSÉE

# Introduction

En tant que navigateur sur Internet, on a accès à beaucoup d'informations qui portent sur des clients, des offres, des cours d'actions, des phénomènes physiques, etc. Ces données peuvent être lues par d'utilisateurs, mais on aimerait pouvoir les exploiter en les transformant dans un format opérationnel pour enfin les analyser et en tirer profit.

# 1. Le web scraping

### 1.1. Définition

Le web scraping est la technique qui permet de retirer ces informations en un format exploitable par les programmes informatiques.

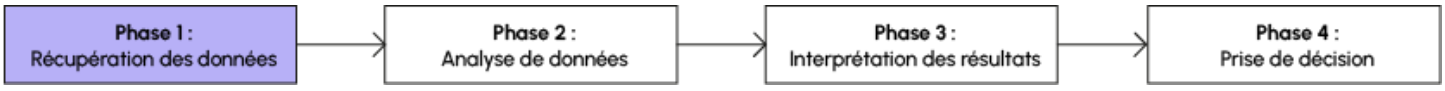
Par exemple,

- on peut avoir envie d'avoir accès à tous les avis d'un Pack de Cartouches d'encre noire HP sur Amazon pour pouvoir faire de l'analyse syntaxique, sémantique et sentimentale et se faire son propre avis.
- le web scraping à partir d'un localisateur de magasins (carte par exemple) permet de créer une liste d'emplacements commerciaux.
- on peut également obtenir les cours d'actions afin de prendre des meilleures décisions d'investissement.

En ce qui concerne la partie analyse de données il y a des techniques spécifiques pour chaque type de données et chaque objectif.

### 1.2. Etapes du web scraping

Dans le schéma suivant, on peut voir le processus « logistique » qui permet d'aboutir à une prise de décision en connaissance de cause :




Si on est dans la phase récupération des données, on souhaiterait avoir accès à toutes les informations présentes sur une page web pour pouvoir faire ensuite l'étude souhaitée.

Pour cela on a la possibilité de les copier « à la main » dans un autre document. Mais c'est un travail fastidieux puisque cela peut prendre beaucoup de temps, sans compter les erreurs de frappe qui pourraient se produire lors de la saisie. Comme dit dans l'introduction, le web scraping permet d'avoir accès à ces informations dans un format exploitable.

Pour la deuxième phase, on fait appel à des compétences techniques des data analystes, data engineers ou data scientists pour mettre en place des algorithmes et des études statistiques pertinentes. Par exemple, dans le cas de l'analyse des commentaires sur un produit on peut utiliser un algorithme NLP, qui permet aux machines de comprendre le langage humain.

L'interprétation des données est souvent faite au sein d'une équipe en tenant compte de l'avis des spécialistes du domaine (par exemple, tenir compte d'avis du médecin si on travaille sur un projet avec des données médicales), pour enfin arriver à une prise de décision optimale.

Dans ce qui suit, nous allons nous intéresser à une librairie qui permet de faire du web scraping disponible en langage python et qui constitue un excellent outil pour extraire des informations de données non structurées. Il s'agit de Beautiful Soup:

# 2. La librairie Beautiful Soup

La librairie Beautiful Soup permet d'extraire du contenu et le transforme en une liste, tableau ou dictionnaire Python.Cette librairie est très populaire parce qu'elle a une documentation complète et ses fonctionnalités sont bien structurées. De plus, il y a une grande communauté qui propose diverses solutions concernant l'utilisation de cette librairie.

### 2.1. Origine du nom de la librairie

Les sites web sont écrits avec les langages informatiques HTML et CSS qui permettent de mettre en page des pages web. Pour gérer et organiser le contenu on utilise le HTML. La partie gestion de l'apparence de la page web (couleurs, taille du texte, etc.) est gérée par le langage CSS.

Dans le domaine du développement web, la « tag soup » (soupe aux balises) est un terme dépréciatif désignant l'écriture du HTML syntaxiquement ou structurellement incorrecte écrite pour une page web.

Un exemple de web scraping avec Beautiful soup

### 2.2. Exemple de scrapping

L'exemple suivant a été pris sur Kaggle et le but est de scraper des données sur la population dans le monde. Les données sont disponibles sur le site Worldometer, une open source géré par une équipe internationale de développeurs et chercheurs bénévoles, dont l'objectif est de mettre les statistiques mondiales à disposition d'un large public dans le monde entier.

Voici un aperçu de la page à scraper

## Countries in the world by population (2022)

This list includes both **countries** and **dependent territories**. Data based on the latest *United Nations Population Division* estimates.
Click on the name of the country or dependency for current estimates (live population clock), historical data, and projected figures.
See also: World Population

Search:

| # | Country (or dependency) | Population (2020) | Yearly Change | Net Change | Density (P/Km²) | Land Area (Km²) | Migrants (net) | Fert. Rate | Med. Age | Urban Pop % | World Share |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | China | 1,439,323,776 | 0.39 % | 5,540,090 | 153 | 9,388,211 | -348,399 | 1.7 | 38 | 61 % | 18.47 % |
| 2 | India | 1,380,004,385 | 0.99 % | 13,586,631 | 464 | 2,973,190 | -532,687 | 2.2 | 28 | 35 % | 17.70 % |
| 3 | United States | 331,002,651 | 0.59 % | 1,937,734 | 36 | 9,147,420 | 954,806 | 1.8 | 38 | 83 % | 4.25 % |
| 4 | Indonesia | 273,523,615 | 1.07 % | 2,898,047 | 151 | 1,811,570 | -98,955 | 2.3 | 30 | 56 % | 3.51 % |
| 5 | Pakistan | 220,892,340 | 2.00 % | 4,327,022 | 287 | 770,880 | -233,379 | 3.6 | 23 | 35 % | 2.83 % |
| 6 | Brazil | 212,559,417 | 0.72 % | 1,509,890 | 25 | 8,358,140 | 21,200 | 1.7 | 33 | 88 % | 2.73 % |
| 7 | Nigeria | 206,139,589 | 2.58 % | 5,175,990 | 226 | 910,770 | -60,000 | 5.4 | 18 | 52 % | 2.64 % |
| 8 | Bangladesh | 164,689,383 | 1.01 % | 1,643,222 | 1,265 | 130,170 | -369,501 | 2.1 | 28 | 39 % | 2.11 % |
| 9 | Russia | 145,934,462 | 0.04 % | 62,206 | 9 | 16,376,870 | 182,456 | 1.8 | 40 | 74 % | 1.87 % |
| 10 | Mexico | 128,932,753 | 1.06 % | 1,357,224 | 66 | 1,943,950 | -60,000 | 2.1 | 29 | 84 % | 1.65 % |

Notre but est de récupérer cette table et la transformer en un DataFrame sans avoir à copier « à la main » toutes ces données.

On commence d'abord par importer les librairies nécessaires.

```
import csv
import requests
from bs4 import BeautifulSoup
import pandas as pd
```

Puis on crée une variable url en format string (du texte) qui contient le lien de la page en question.

```
url = "https://www.worldometers.info/world-population/populati
```

Pour préparer les données on utilise la fonction requests.get() :

```
req = requests.get(url)
```

Maintenant que les données sont préparées, la fonction BeautifulSoup() permet d'extraire le code HTML de cette page.

Dans l'argument de cette fonction on va sélectionner l'objet text.

```
soup = BeautifulSoup(req.text)
```

Dans la variable data on stocke le code HTML on cherche le mot clé « table » avec la fonction .find_all() :

```
data = soup.find_all("table")[0]
```

On utilise la commande .read_html(str()) pour que la machine lise le code HTML et puis on récupère le premier et unique élément de cet objet (le tableau).

```
df_population = pd.read_html(str(data))[0]
```

À présent on affiche les premiers éléments, commande .head() du DataFrame :

```
df_population.head()
```

On peut aussi exporter la base de données en format csv avec la commande suivante :

```
export_csv = df_population.to_csv(r"le chemin\nom du fichier.c
```

Cette commande crée un fichier en format csv localisé au chemin indiqué.

Pour rendre plus faciles les manipulations sur le DataFrame, on peut penser à changer les noms de colonnes ou en éliminer quelques-unes si on ne les utilise pas.

Maintenant que nous avons pu obtenir les données qui ont été mises à disposition sur le site de Worldometer et qu'elles sont bien dans un format DataFrame, on peut passer aux autres étapes et faire des études. On peut donc se lancer dans la suite du processus (Phase 2, 3 et 4, voir le schéma).

Selon la nature des données et en fonction d'objectifs à atteindre on peut faire différentes études : analyse exploratoire, proposer un modèle de machine learning, modélisation des séries temporelles, etc.

Nous venons de voir un exemple qui permet de récupérer les données stockées dans table, mais il faut retenir qu'en fonction de la structure de la page web sur laquelle on veut scraper les données on utilise des librairies et des fonctions différentes.

Pour résumer, le web scraping permet de naviguer « intelligemment » sur Internet et donc constitue une ressource riche pour tout domaine de recherche ou d'intérêt personnel.

# 3. Analyse de Sentiment

L'analyse de sentiment permet de connaître la satisfaction des clients et ainsi guider les stratégies commerciales.

Donc mettre en place un monitoring de la satisfaction permet alors de

- La suivre dans le temps et d'en suivre les tendances :
- Connaître l'évolution de l'avis des internautes sur la marque/les produits/les concurrents (moins coûteux que des enquêtes d'opinions),
- Identifier les influenceurs sur l'activité de la marque,
- Identifier les sources de frustration.

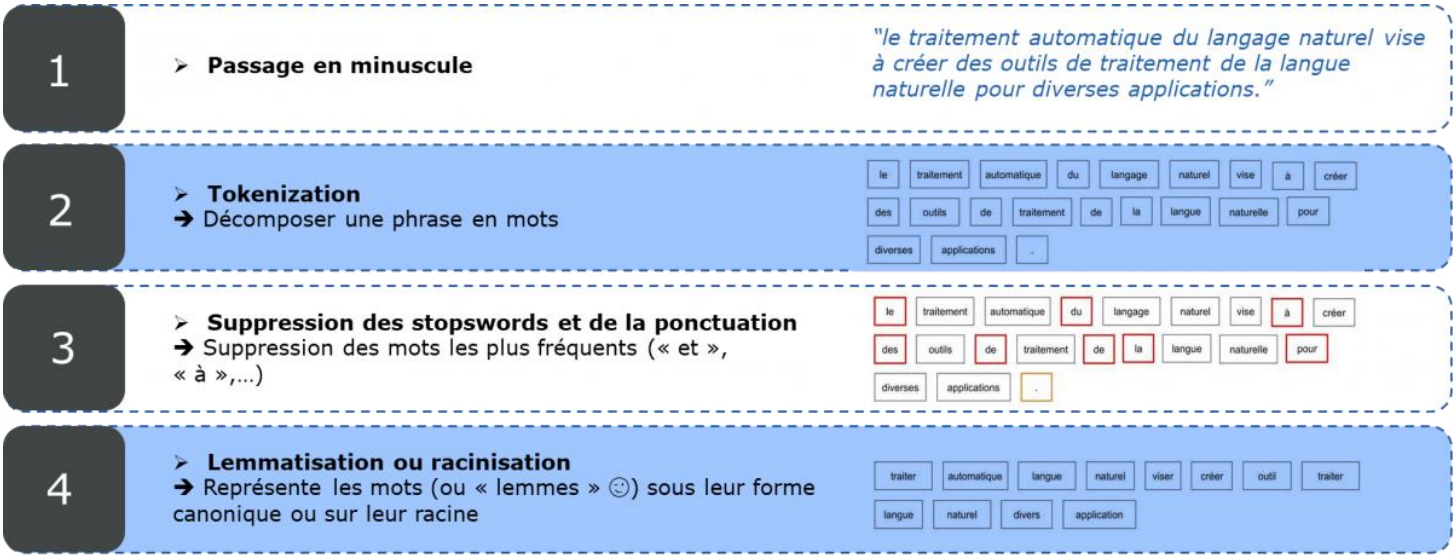
Pour avoir une communication adaptée à chaque groupe ou prioriser les réponses à donner en urgence il vaut mieux segmenter la base des utilisateurs en plusieurs catégories.

Un message ayant un sentiment négatif peut provenir d'un client particulièrement insatisfait auquel il sera préférable de répondre rapidement.

Enfin, il peut aussi servir d'indicateur dans les prédictions, par exemple, avec le lancement d'un nouveau produit ou d'une nouvelle campagne marketing.

Du côté pratique, quel que soit le cas d'usage, les algorithmes de Data Science ne savent pas manipuler du texte brut. C'est pourquoi une étape de préparation des données est nécessaire.

Le premier objectif de cette préparation est de réduire le nombre de mots pour ne conserver que ceux donnant son sens au message. Il est possible d'effectuer les étapes suivantes :




Ces étapes sont courantes mais pas systématiques. Tout dépend de notre objectif. Par exemple, pour détecter des spams, l'utilisation des majuscules donnera un signal important.

Ensuite, habituellement, chaque texte est transformé en des vecteurs de mots qui mènent à créer une matrice document-terme qui sera l'entrée de l'analyse. Cela facilite l'usage de certains algorithmes qui ne prennent en entrée que des nombres.

| | | Termes (mots) | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | anxiété | association | bien | résister | vitamines | contre | dépression |
| Documents (phrases) | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| | 2 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| | 5 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| | 6 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| | 8 | 0 | 1 | 0 | 0 | 0 | 1 | 1 |
| | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

Côté méthodologie, plusieurs approches s'offrent à nous :

- Approche par dictionnaire : il faut disposer d'un référentiel où chaque mot est associé à un score de sentiment. Le score d'un message est obtenu à partir des scores des mots qui le composent. L'avantage de cette méthode est qu'elle est simple à comprendre, à expliquer et à implémenter. Par contre, elle ne tient pas compte du contexte dans lequel le mot est employé et ne gère pas du tout les sarcasmes, l'ironie, etc.
- Approche supervisée : entraîner un modèle de Machine Learning à différencier les messages positifs de ceux négatifs à partir de données labélisées. Exemples de modèles : SVM, régression logistique, XGBoost. L'utilisation de classification naïve bayésienne est aussi une approche supervisée possible. Elle calcule pour un message la probabilité de chaque classe de sentiment (positive, négative ou neutre) sachant les mots qui le composent. Ces méthodes sont souvent plus performantes que l'approche par dictionnaire mais

sont un peu moins facilement explicables à des interlocuteurs métier. Elles nécessitent en outre un corpus de messages dont on connait déjà le sentiment.

- Word Embedding : utiliser un réseau de neurones résumant un texte en un vecteur de nombres avant d'en prédire le sentiment. Ces vecteurs peuvent être pré-calculés et utilisés en entrée des modèles supervisés présentés juste avant, ou alors être découverts en entraînant un réseau de neurones adapté à la problématique.
  Cette méthode est très performante car c'est la méthode prenant le mieux en compte le contexte dans lequel le mot est utilisé. Cependant, comme tout réseau de neurones, il est difficile à expliquer et à interpréter.
- API : il est également possible d'utiliser des API, déjà très performantes, telles que l'API Natural Language de Google ou l'API Cognitive Services d'Azure.

Toutes ces méthodes conduisent au même résultat qui donne pour chaque message un score de satisfaction permettant de quantifier la satisfaction ou l'insatisfaction du client.

# 4. NLP (Natural Language Processing)

L'analyse et surtout la compréhension efficace des retours-clients sont devenues des chantiers marketing fondamentaux pour toute entreprise souhaitant optimiser son expérience d'achat, améliorer un produit ou évaluer l'impact d'une décision stratégique.

Que pensent les clients de mon produit ? Quelles sont les attentes de mes clients ? Quels sont mes points forts et mes points faibles dans le tunnel d'achat ? Est-ce que le nouvel agencement d'un de mes magasins plaît ?

Pour répondre à ces questions, bon nombre de commerçants et e-commerçants se sont dotés d'outils qui prennent en compte les retours suite à des achats de produits et services ou à l'abandon d'un processus en cours. Mais ces questionnaires traditionnels montrent leurs limites quant à la précision et la pertinence des insights délivrés. Ils viennent compléter l'ensemble des informations laissées par les internautes sur Google, TripAdvisor, Twitter, Booking, …

Bien sûr, il est possible de parcourir l'ensemble des commentaires des questionnaires de satisfaction et de se faire une idée, mais cette analyse n'est pas assez robuste pour prendre des décisions stratégiques.

Au détriment du traditionnel questionnaire de satisfaction, les nouvelles générations utilisent des canaux divers et variés pour émettre avis et commentaires sur leur expérience client. La multiplicité des canaux et l'important volume de données rendent donc impossible un traitement manuel.

## 4.1. Importance du NLP pour optimiser l'expérience client

Le traitement automatique du langage dit naturel, appelé aussi Natural Language Processing (NLP) en anglais, est une technologie permettant à des machines d'analyser le langage humain grâce à l'intelligence artificielle (IA). L'ordinateur peut alors comprendre et synthétiser les retours clients.

Plusieurs techniques avancées de NLP sont utilisées pour extraire des informations précieuses à partir des commentaires écrits par les clients.

Les algorithmes de Topic Modeling nous permettent de détecter des similarités entre groupes de commentaires et ainsi identifier des sujets-clés. Nous pouvons alors suivre l'évolution de ces thèmes principaux et surveiller en temps réel l'efficacité d'une décision stratégique.

Une fois les thématiques identifiées, nous utilisons des techniques avancées de machine learning afin de mesurer les émotions positives ou négatives associées.

Couplé avec du web scrapping, nous sommes capables d'automatiser cette analyse sur l'ensemble des canaux de contacts : questionnaires de satisfaction, forums de consommateurs, réseaux sociaux etc

Le NLP est donc un outil puissant du Data Scientist – à vocation marketing- qui permet aux entreprises de prendre de la hauteur mais également d'identifier des leviers actionnables rapidement sur le terrain.

L'ère du numérique a modifié profondément le rôle du client dans le commerce. La compréhension de leurs attentes, tant pour la direction des ventes que pour le marketing et la relation client, est désormais un atout stratégique de premier plan. Ces dernières années, les entreprises se sont attachées à collecter une multitude de données-clients autant quantitatives que qualitatives. Le défi consiste maintenant à les exploiter de manière efficiente et opérationnelle.

La Data Science accompagne les entreprises pour passer d'une simple connaissance à une véritable intelligence-client.

### 4.2. Techniques autour du NLP

Il existe de nombreuses techniques autour du NLP, telles que :

- L'analyse de sentiment,
- La traduction automatique
- La détection de thèmes (nouveaux trends twitter, nouveaux sujets abordés dans les médias, trouver les différents aspects d'un produit abordés par des commentaires afin de pouvoir plus facilement l'améliorer à partir de feedbacks utilisateurs, …),
- La classification de messages (exemple : spams),
- La reconnaissance vocale,
- Les assistants personnels tels que Apple Siri, Microsoft Cortana, Amazon Alexa,
- Les chatbots,
- La génération automatique de texte,
- …

Nous ne nous focaliserons ici que sur les techniques de traitement de textes qui nous permettent de remplir notre objectif d'écoute des clients : l'analyse de sentiment, la détection automatique de thèmes, appelée topic modeling, ou encore la classification de messages dans des thèmes.

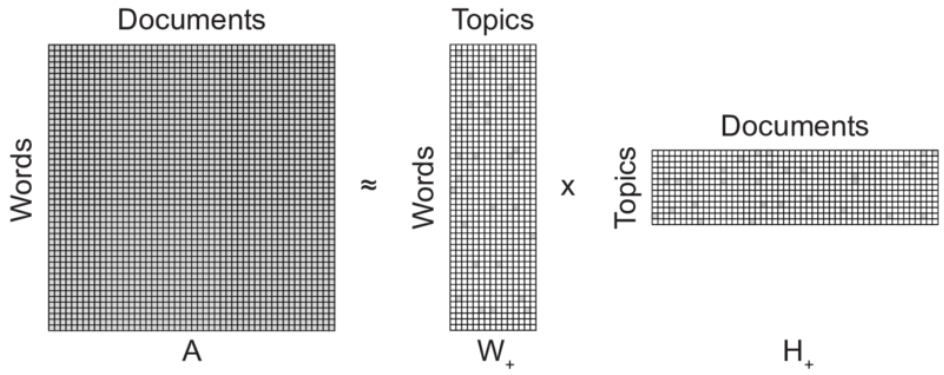
### 4.3. Le Topic Modeling:

La deuxième fonctionnalité principale du NLP est l'extraction de thèmes plus communément appelée topic modeling. Le topic modeling peut s'appliquer à toute forme de texte : mails, tickets, feedbacks, etc. pour avoir une vision globale des préoccupations des clients.

Les principaux modèles de topic modeling sont non-supervisés. C'est-à-dire qu'ils n'apprennent pas à lier des messages à un thème donné, ils découvrent eux-mêmes les thèmes.

Mais avant d'être analysés, les messages doivent passer par la même préparation que pour l'analyse de sentiment. Ensuite, il existe, là encore, plusieurs méthodologies possibles :

- Latent Dirichlet Allocation (LDA) : modèle probabiliste et algorithmique parcourant les messages pour former des groupes de mots qui co-occurrent souvent et ainsi découvrir des thèmes.
- Latent Semantic Analysis (LSA) : modèle d'algèbre linéaire décomposant le lien « terme-document » en un lien « terme-thème » + « thème-document ». Il est basé sur la même intuition que la matrix factorization pour la recommandation de produits.
- Non-negative Matrix Factorization (NMF) : modèle d'algèbre linéaire réalisant le même travail que la LSA pour découvrir des variables latentes, les thèmes. Les deux modèles se différencient par leur méthode de décomposition mais la LSA est plus fréquemment utilisée notamment par son caractère unique et son interprétation un peu plus aisée (grâce à l'importance des thèmes).



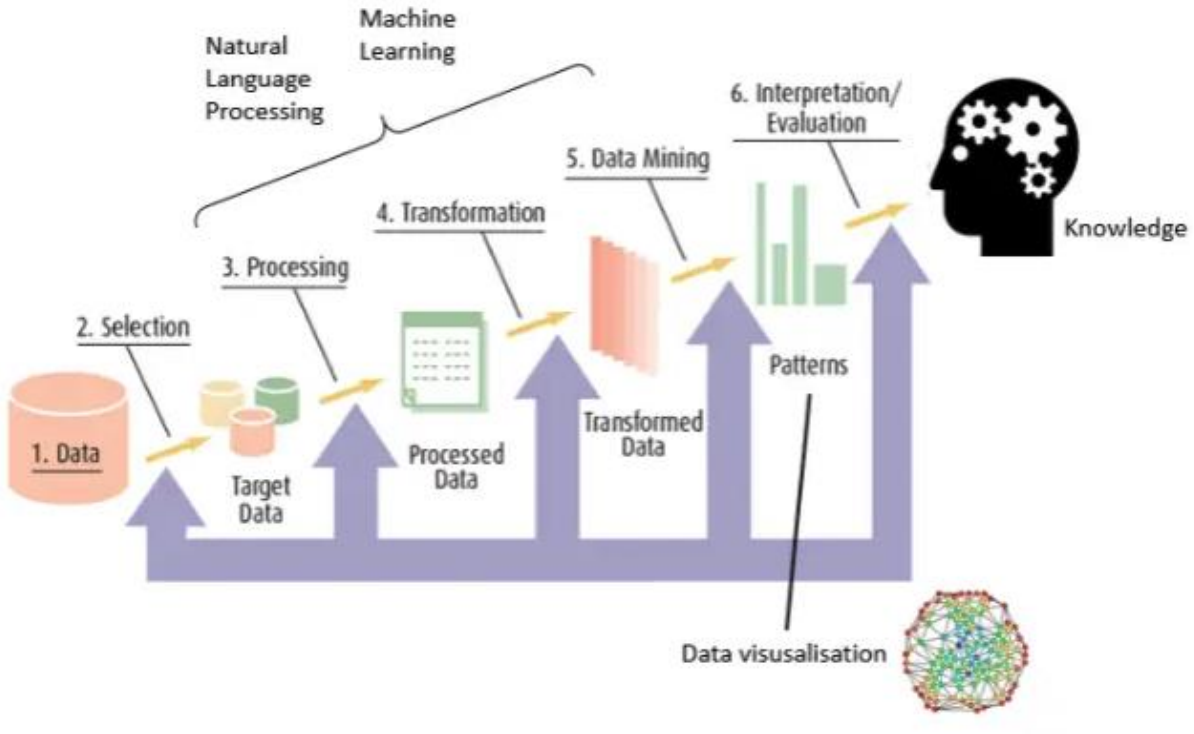

Ces 3 approches demandent de donner en paramètre le nombre de thèmes. Il existe des critères statistiques pour donner une indication sur le nombre de thèmes optimal mais il reste nécessaire, pour choisir sa méthodologie ou son nombre de thèmes, de s'assurer de la pertinence de l'interprétation des topics.

Les modèles nous fournissent le numéro du thème auquel le message est associé, voire même l'importance relative de chaque thème pour le message.

Quoiqu'il en soit, le thème ne sera qu'un listing de mots auquel vous devrez associer vous-mêmes un nom.

| Thèmes | Mots associés |
|---|---|
| 1 | Voiture, parking, vélo, transport, bus, … |
| 2 | Alternance, stage, candidature, CV, … |

## 4.4. Les Phases NLP




### 4.4.1 La phase de prétraitement : du texte aux données :

Supposons que nous voulons être capables de déterminer si un mail est un spam ou non, uniquement à partir de son contenu.

À cette fin, il est indispensable de transformer les données brutes (le texte du mail) en des données exploitables.

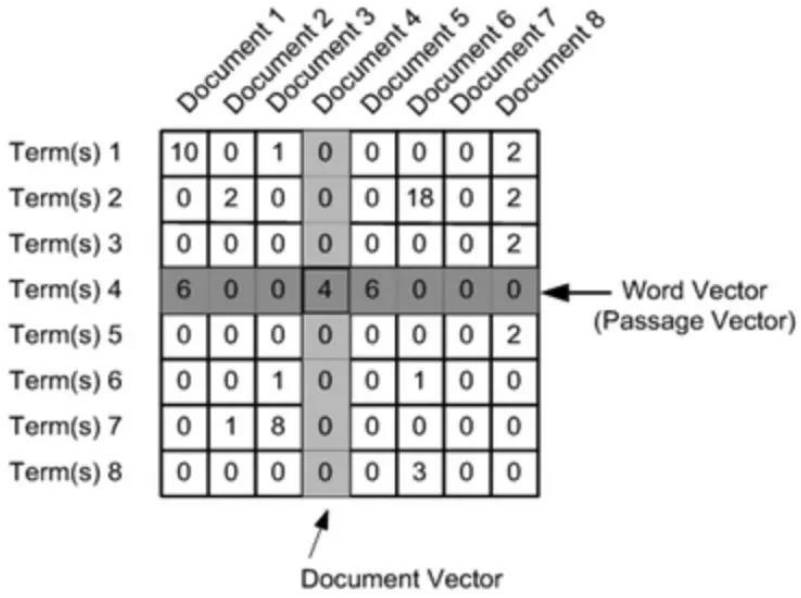
Parmi les principales étapes, on retrouve :

- Nettoyage : Variable selon la source des données, cette phase consiste à réaliser des tâches telles que la suppression d'urls, d'emoji, etc.
- Normalisation des données
- Tokenisation, ou découpage du texte en plusieurs pièces appelés tokens. Exemple : « Vous trouverez en pièce jointe le document en question » ; « Vous », « trouverez », « en pièce jointe », « le document », « en question ».
- Stemming : un même mot peut se retrouver sous différentes formes en fonction du genre (masculin féminin), du nombre (singulier, pluriel), la personne (moi, toi, eux…) etc. Le stemming désigne généralement le processus heuristique brut qui consiste à découper la fin des mots dans afin de ne conserver que la racine du mot. Exemple : « trouverez » -> « trouv »
- Lemmatisation : cela consiste à réaliser la même tâche mais en utilisant un vocabulaire et une analyse fine de la construction des mots. La lemmatisation permet donc de supprimer uniquement les terminaisons inflexibles et donc à isoler la forme canonique du mot, connue sous le nom de lemme. Exemple : « trouverez » -> trouvez
- Autres opérations : suppression des chiffres, ponctuation, symboles et stopwords, passage en minuscule.

Afin de pouvoir appliquer les méthodes de Machine Learning aux problèmes relatifs au langage naturel, il est indispensable de transformer les données textuelles en données numériques.

Il existe plusieurs approches dont les principales sont les suivantes :

Term-Frequency (TF) : cette méthode consiste à compter le nombre d'occurrences des tokens présents dans le corpus pour chaque texte. Chaque texte est alors représenté par un vecteur d'occurrences. On parle généralement de Bag-Of-Word, ou sac de mots en français.




Représentation des vecteurs issues de la méthode Term-Frequency (TF)

Néanmoins, cette approche présente un inconvénient majeur : certains mots sont par nature plus utilisés que d'autres, ce qui peut conduire le modèle à des résultats erronés.

Term Frequency-Inverse Document Frequency (TF-IDF) : cette méthode consiste à compter le nombre d'occurrences des tokens présents dans le corpus pour chaque texte, que l'on divise ensuite par le nombre d'occurrences total de ces même tokens dans tout le corpus.

Pour le terme x présent dans le document y, on peut définir son poids par la relation suivante :

$$w_{x,y} = tf_{x,y} \cdot log(\frac{N}{df_x})$$

Où :

- t$f$x,y est la fréquence du terme x dans y ;
- d$f$x est le nombre de documents contenant x ;
- N est le total de documents.

Cette approche permet donc d'obtenir pour chaque texte une représentation vectorielle qui comporte des vecteurs de poids et non plus d'occurrences.

L'efficacité de ces méthodes diffère selon le cas d'application. Toutefois, elles présentent deux principales limites : Plus le vocabulaire du corpus est riche, plus la taille des vecteurs est grande, ce qui peut représenter un problème pour les modèles d'apprentissage utilisées dans l'étape suivante.

Le comptage d'occurrence des mots ne permet pas de rendre compte de leur agencement et donc du sens des phrases.

Il existe une autre approche qui permet de remédier à ces problèmes : Word Embedding. Elle consiste à construire des vecteurs de taille fixe qui prennent en compte le contexte dans lequel se trouvent les mots.

Ainsi, deux mots présents dans des contextes similaires auront des vecteurs plus proches (en terme de distance vectorielle). Cela permet alors de capturer à la fois similarités sémantiques, syntaxiques ou thématiques des mots.

### 4.4.2. La phase d'apprentissage : des données au modèle

De manière globale, on peut distinguer 3 principales approches NLP : les méthodes basées sur des règles, modèles classiques de Machine Learning et modèles de Deep Learning.

- Méthodes basées sur des règles : Les méthodes fondées sur des règles reposent en grande partie sur l'élaboration de règles spécifiques à un domaine (par exemple, les expressions régulières). Elles peuvent être utilisées pour résoudre des problèmes simples tels que l'extraction de données structurées à partir de données non structurées (par exemple, les pages web).
  Dans le cas de la détection de spams, cela pourrait consister à considérer comme e-mails indésirables, ceux qui comportent des buzz words tels que « promotion », « offre limitée », etc. Néanmoins, ces méthodes simples peuvent être rapidement dépassées par la complexité du langage naturel et s'avérer être inefficaces.
- Modèles classiques de Machine Learning : Les approches classiques d'apprentissage automatique peuvent être utilisées pour résoudre des problèmes plus difficiles.
  Contrairement aux méthodes fondées sur des règles prédéfinies, elles reposent sur des méthodes qui portent réellement sur la compréhension du langage.
  Elles exploitent les données obtenues à partir des textes bruts prétraités via une des méthodes décrites en haut par exemple. Elles peuvent également utiliser des données relatives à la longueur des phrases, à l'occurrence de mots spécifiques, etc.
  Elles mettent généralement en œuvre un modèle statistique d'apprentissage automatique tels que ceux de Naive Bayes, de Régression Logistique, etc.
- Modèles de Deep Learning : L'utilisation de modèles d'apprentissage en profondeur pour les problématiques NLP fait l'objet de nombreuses recherches actuellement.
  Ces modèles se généralisent encore mieux que les approches classiques d'apprentissage car ils nécessitent une phase de prétraitement du texte moins sophistiquée : les couches de neurones peuvent être perçues comme des extracteurs automatiques de features.
  Cela permet alors de construire des modèles de bout en bout avec peu de prétraitement des données. En dehors de la partie feature engineering, lescapacités d'apprentissage des algorithmes de Deep Learning sont généralement plus puissantes que celles de Machine Learning classique, ce qui permet d'obtenir de meilleurs scores sur différentes tâches complexes de NLP dures telles que la traduction.

# Conclusion

Les méthodes vues peuvent être combinées entre elles ou avec d'autres pour extraire d'autres informations.

Par exemple, le topic modeling, utilisé en parallèle d'une analyse de sentiment, permet de mettre en lumière les sujets de mécontentement des utilisateurs. Il permet de savoir si les clients sont satisfaits ou non d'un service en particulier. Il indique ainsi les services à améliorer en priorité.

Le NLP peut aussi être élargi aux données de centres d'appels ou d'assistants personnels. Des méthodes de text-to-speech permettent de transcrire les données vocales en texte sur lequel toutes les méthodes vues précédemment peuvent s'appliquer.

Les données textuelles, ou plus généralement de langage, sont omniprésentes et souvent sous-exploitées bien qu'elles contiennent des informations clés. Utiliser des méthodes de NLP comme l'analyse de sentiment, le topic modeling et la classification permet d'être plus à l'écoute de vos clients et ainsi améliorer la prise de décisions stratégiques. La diversification des types de données, l'augmentation du volume de données (sous forme de texte ou de son) vont s'accélérer dans les prochaines années. C'est donc une source d'information capitale pour la relation client et la stratégie de l'entreprise

# CHAPITRE 3: RÉALISATION DE LA SOLUTION

# Introduction

L'hôtellerie est certainement l'un des secteurs à la croissance la plus rapide dans le secteur du tourisme. Le tourisme est également une opportunité d'emploi potentiellement importante et les hôtels constituent une partie importante de ce secteur hôtelier. L'industrie hôtelière a contribué activement à la croissance économique du pays.

Cette tendance devrait croître progressivement et, à son tour, stimuler ou ajouter du sens au tourisme de n'importe quel endroit.

Dans ce projet, nous allons prendre les mesures suivantes pour aider l'hôtel à améliorer la satisfaction de ses clients :

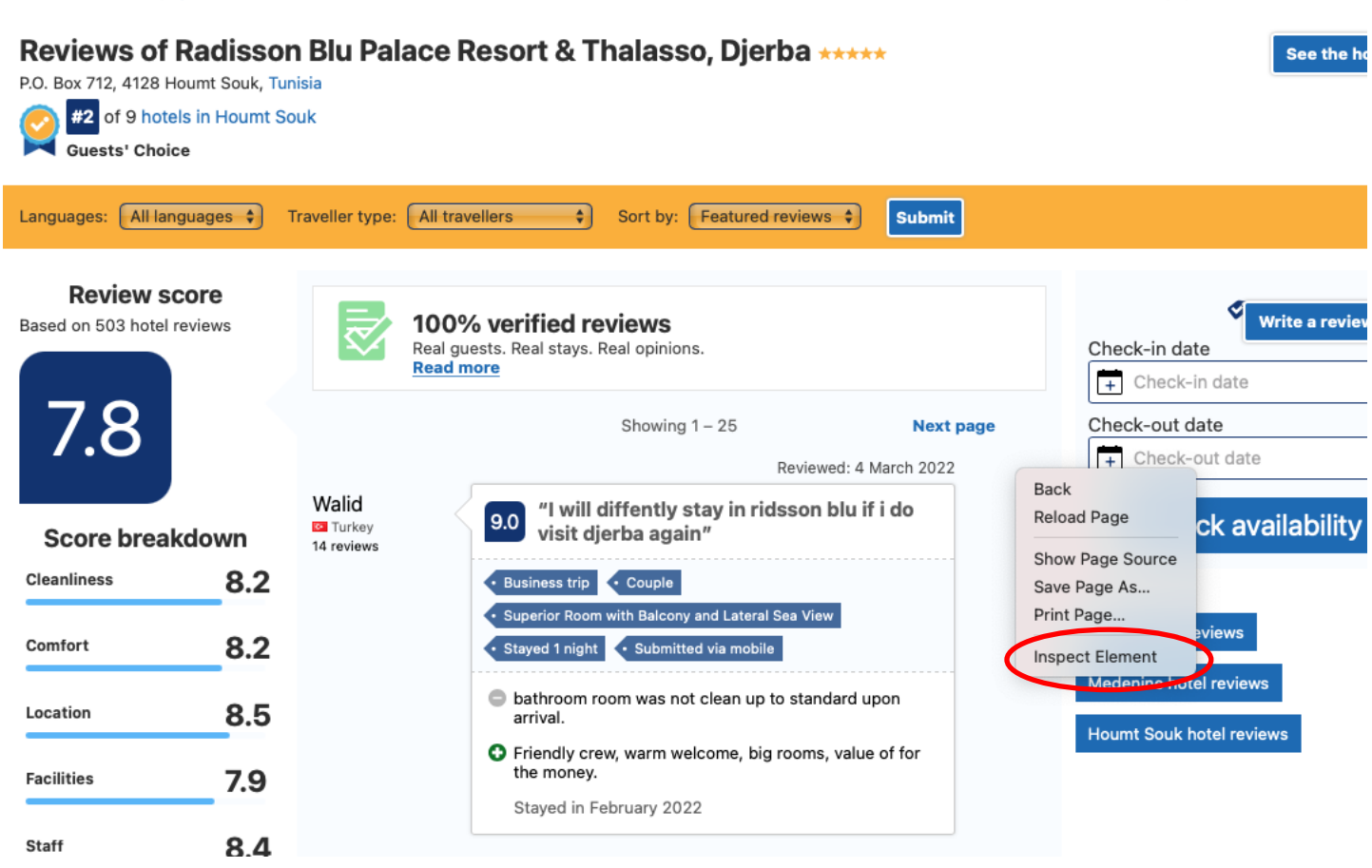
- Extraire les avis des hôtels du site Web « Booking.com » via Web Scrapping.
- Analyse exploratoire des données pour obtenir des informations significatives à partir des données.
- Analyse des sentiments pour comprendre les sentiments du client envers l'hôtel.
- Sujet Modélisation pour comprendre les principaux facteurs entraînant un sentiment négatif des clients.

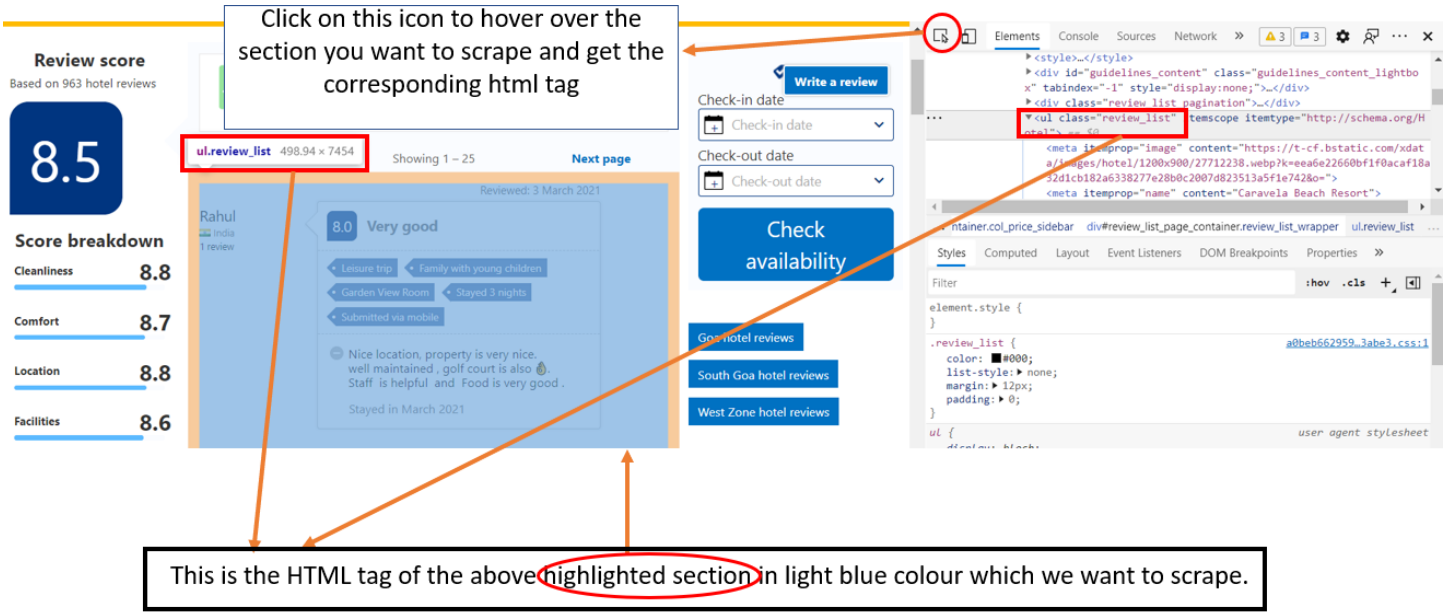
# 1.Web Scrapping

Dans ce projet, nous allons gratter les critiques de « Hotel Radisson Blu Djerba » situé à la Zone Touristique Houmt Souk Djerba. Les avis ont été extraits du site booking.com avec le lien ci-dessous:

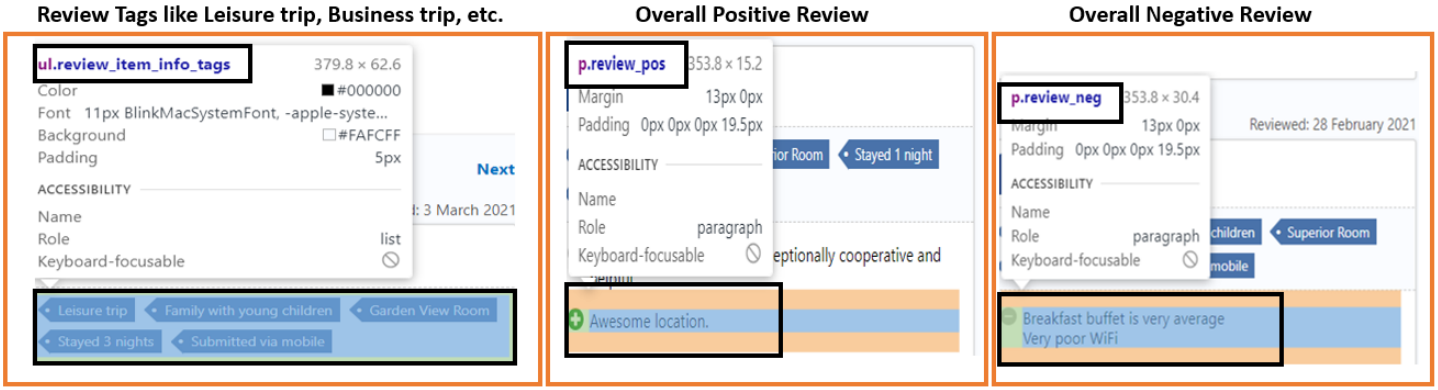
https://www.booking.com/reviews/tn/hotel/radisson-sas-resort-thalasso-djerba.html?r_lang=all&page=

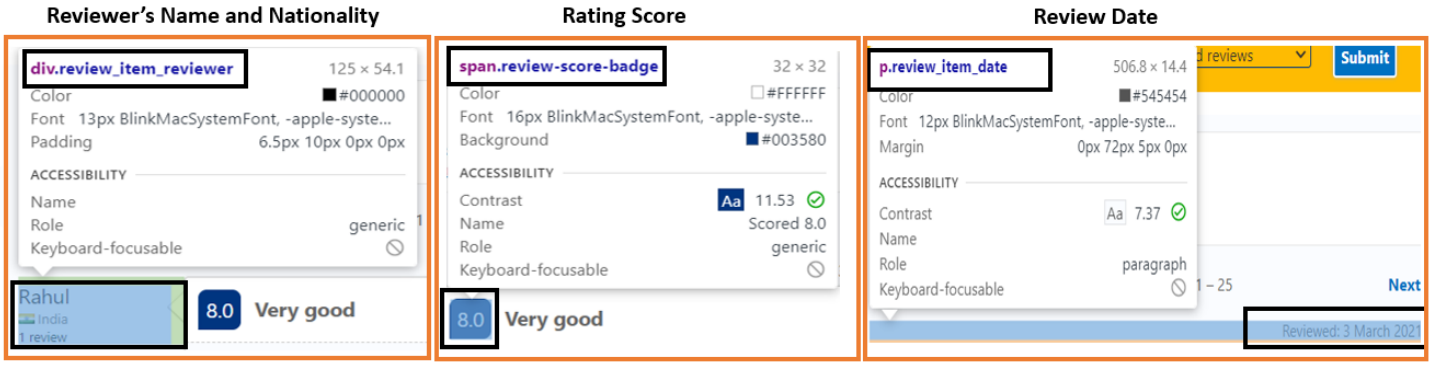
Pour commencer le processus de Scrapping, on doit "Inspecter" pour trouver les balises HTML associée aux informations que nous voulons gratter de cette page Web comme le montre la capture d'écran ci-dessous.






Sous cette balise « ul.review_list », nous grattons bien diverses informations comme indiqué ci-dessous, ainsi que la capture d'écran des balises HTML correspondantes :



## 2. Importation des bibliothèques

```
# importing packages
import numpy as np
import pandas as pd
import seaborn as sns
import plotly as px
import matplotlib.pyplot as plt
#%matplotlib inline
import re
from bs4 import BeautifulSoup as bs
import requests
import string
import nltk
from nltk.stem import WordNetLemmatizer
from nltk.corpus import stopwords,wordnet
#from wordcloud import WordCloud
from textblob import TextBlob
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.model_selection import GridSearchCV
from sklearn.decomposition import LatentDirichletAllocation
import pyLDAvis
import pyLDAvis.sklearn
pyLDAvis.enable_notebook()
import warnings
warnings.filterwarnings("ignore")

```

Maintenant, j'ai créé une fonction ci-dessous nommée « scrape_reviews » qui a 2 arguments :

- hotel_linkname = Mentionnez le nom de tout hôtel que vous souhaitez gratter comme mentionné dans le lien booking.com
- total_pages = Mentionnez le nombre total de pages de révision que vous voulez gratter.

Comprenons la vue d'ensemble de la fonction « scrape_reviews » :

- Mentionner l'URL de la page web de la revue.
- Récupérer les données du serveur.
- Dans ce projet, nous utiliserons la bibliothèque Beautiful Soup pour gratter la page HTML.
- Nettoyer le texte en utilisant les méthodes .strip() et .replace()
- Utiliser une boucle While pour gratter toutes les pages.

Cette fonction nous donnera les 3 sorties ci-dessous sous forme de trames de données :

- reviewer_info : une base de données qui inclut les informations de base des évaluateurs
- pos_reviews : une base de données qui inclut toutes les critiques positives
- neg_reviews : une base de données qui inclut toutes les critiques négatives

```python
def scrape_reviews(hotel_linkname,total_pages ):
    #Create empty lists to put in reviewers' information as well as all of the positive & negative reviews
    info = []
    positive = []
    negative = []

    #bookings.com reviews link
    url = "https://www.booking.com/reviews/tn/hotel/"+hotel_linkname+".html?r_lang=all&page="
    page_number = 1
    #Use a while loop to scrape all the pages
    print('connecting to'+url)
    while page_number <= total_pages:
        print(page_number)

        page = requests.get(url + str(page_number)) #retrieve data from server
        soup = bs(page.text, "html.parser") # initiate a beautifulsoup object using the html source and Python's html.parser
        review_box = soup.find('ul',{'class':'review_list'})
        #ratings
        ratings = [i.text.strip() for i in review_box.find_all('span',{'class':'review-score-badge'})]

        #reviewer_info
        print('reviews')
        reviewer_info = [i.text.strip() for i in review_box.find_all('span',{'itemprop':'name'})]
        reviewer_name = reviewer_info[0::3]
        reviewer_country = reviewer_info[1::3]
        general_review = reviewer_info[2::3]
        # reviewer_review_times
```

```python
    # reviewer_review_times
    review_times = [i.text.strip() for i in review_box.find_all('div',{'class':'review_item_user_review_count'})]
    # review_date
    review_date = [i.text.strip().strip('Reviewed: ') for i in review_box.find_all('p',{'class':'review_item_date'})]
    # reviewer_tag
    reviewer_tag = [i.text.strip().replace('\n\n\n','').replace('•',',').lstrip(', ') for i in review_box.find_all('ul',{'class':'review_item_info_tags'
    # positive_review
    positive_review = [i.text.strip('妆').strip() for i in review_box.find_all('p',{'class':'review_pos'})]
    # negative_review
    negative_review = [i.text.strip('庄').strip() for i in review_box.find_all('p',{'class':'review_neg'})]
    # append all reviewers' info into one list
    print('append')
    for i in range(len(reviewer_name)):
        info.append([ratings[i],reviewer_name[i],reviewer_country[i],general_review[i],
        review_times[i],review_date[i],reviewer_tag[i]])
    # build positive review list
    for i in range(len(positive_review)):
        positive.append(positive_review[i])
    # build negative review list
    for i in range(len(negative_review)):
        negative.append(negative_review[i])
    # page change
    page_number +=1
#Reviewer_info df
print('faming')
reviewer_info = pd.DataFrame(info,
columns = ['Rating','Name','Country','Overall_review','Review_times','Review_date','Review_tags'])
reviewer_info['Rating'] = pd.to_numeric(reviewer_info['Rating'] )
reviewer_info['Review_times'] = pd.to_numeric(reviewer_info['Review_times'].apply(lambda x:re.findall("\d+", x)[0]))
reviewer_info['Review_date'] = pd.to_datetime(reviewer_info['Review_date'])

#positive & negative reviews dfs
pos_reviews = pd.DataFrame(positive,columns = ['positive_reviews'])
neg_reviews = pd.DataFrame(negative,columns = ['negative_reviews'])

return reviewer_info, pos_reviews, neg_reviews
```

La fonction ci-dessous "show_data" affichera la longueur des dataframes, le total des valeurs manquantes, ainsi que les cinq premières lignes d'une dataframe.

```
show_data(df):
print("The length of the dataframe is: {}".format(len(df)))
print("Total NAs: {}".format(reviewer_info.isnull().sum().sum()))
return df.head()
```

Maintenant, après avoir créé la fonction, nous allons mentionner le nom de l'hôtel et le nombre total de pages que nous voulons gratter aussi la fonction ‘show_data’ pour la vérification des données grattées.

```
reviewer_info, pos_reviews, neg_reviews = scrape_reviews('radisson-sas-resort-thalasso-djerba',total_pages = 2 )

show_data(reviewer_info) #reviewers' basic information
show_data(pos_reviews)   #Positive reviews
show_data(neg_reviews)   #Negative reviews
```

| Name | Type | Size | Value |
|---|---|---|---|
| neg_reviews | DataFrame | (20, 1) | Column names: negative_reviews |
| pos_reviews | DataFrame | (24, 1) | Column names: positive_reviews |
| reviewer_info | DataFrame | (24, 7) | Column names: Rating, Name, Country, Overa… |

| Index | negative_reviews |
|---|---|
| 0 | bathroom room was not clean up to standard upon arriva… |
| 1 | no info TV channel or information in the rooms, so fin… |
| 2 | car parking |
| 3 | car parking |
| 4 | Facilities i.e. hair dryer and such like had to be req… |
| 5 | the noise coming from the employees either finishing t… |
| 6 | Wifi didn't work in room.  Food in italian restaurant … |
| 7 | the pool is closing early at 8PM it doesn't suit someo… |
| 8 | We asked about the dinner in the asian restaurant and … |
| 9 | Garden a bit messy. Location far from city attractions… |
| 10 | Room cleaning wasn't so good , oneday the sheets haven… |
| 11 | Expensive room service with little food items available |
| 12 | unfortunately due to the covid times, the business was… |
| 13 | The price |
| 14 | Bathroom wasn't very nice, there was construction that… |
| 15 | Ménage des parties communes |
| 16 | Ongoing works in other rooms.<br>Did not hear noise but would smell the dust when walki… |
| 17 | Outdated rooms and poor quality of breakfast. Too expe… |
| 18 | half the hotel was in darkness. poor restaurant food (… |

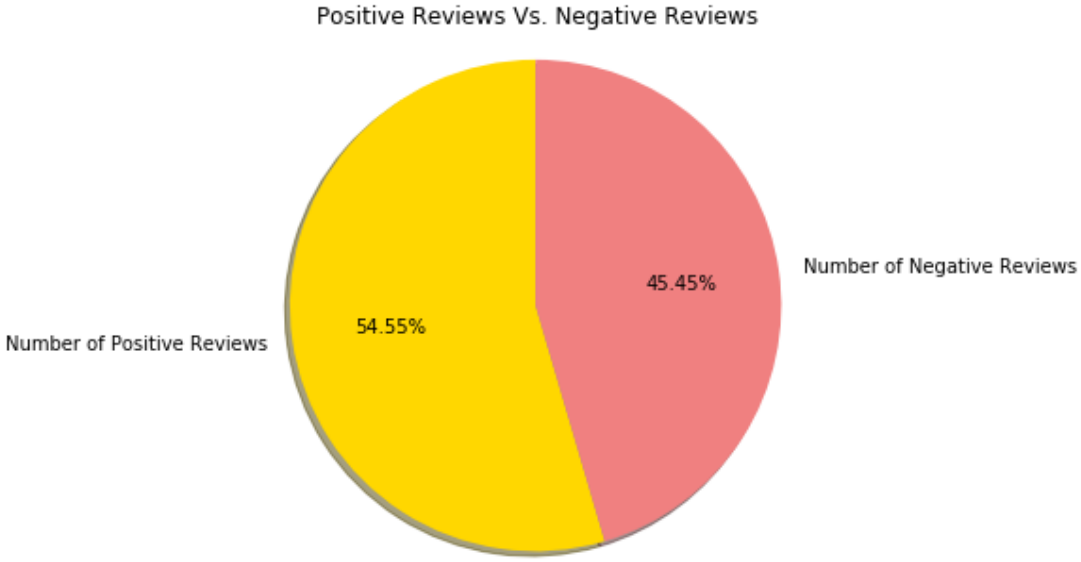
| Index | positive_reviews |
|---|---|
| 0 | Friendly crew, warm welcome, big rooms, value of for t… |
| 1 | very striking buildings, lovely gardens, clean and wel… |
| 2 | Shisha cafe, swimming pool, Room, GYM |
| 3 | Swimming pool, GYM, Shisha cafe, Room |
| 4 | The hotel as whole, the decor, good size rooms, the br… |
| 5 | the location- the rooms renovated - the access to the … |
| 6 | Beautiful grounds, very good breakfast and buffet dinn… |
| 7 | the pool and beach are excellent , the facilities in t… |
| 8 | Great place , very clean and nice facilities . |
| 9 | Scandinavian-Tunisian style. Buffet, view. Close to be… |
| 10 | The hotel design is fantastic , the view of the see an… |
| 11 | Good service at the reception as well.<br>Lots of choice at breakfast buffet<br>Good facilities - pool , gym and sauna - |
| 12 | the hotel is great and spacious. The rooms are very we… |
| 13 | The food is really delicious and all the staff are ama… |
| 14 | Beach view in the second floor |
| 15 | The staff in the restaurants were great, especially th… |
| 16 | The sea view is just amazing and the architecture of t… |
| 17 | Beautiful |
| 18 | My room was great : newly refurbished, spacious , sea … |
| 19 | Luxurious and comfortable |
| 20 | The beach and the pool. |
| 21 | parking. easy to find |
| 22 | we liked everything |

# 3. Analyse exploratoire des données ( EDA)

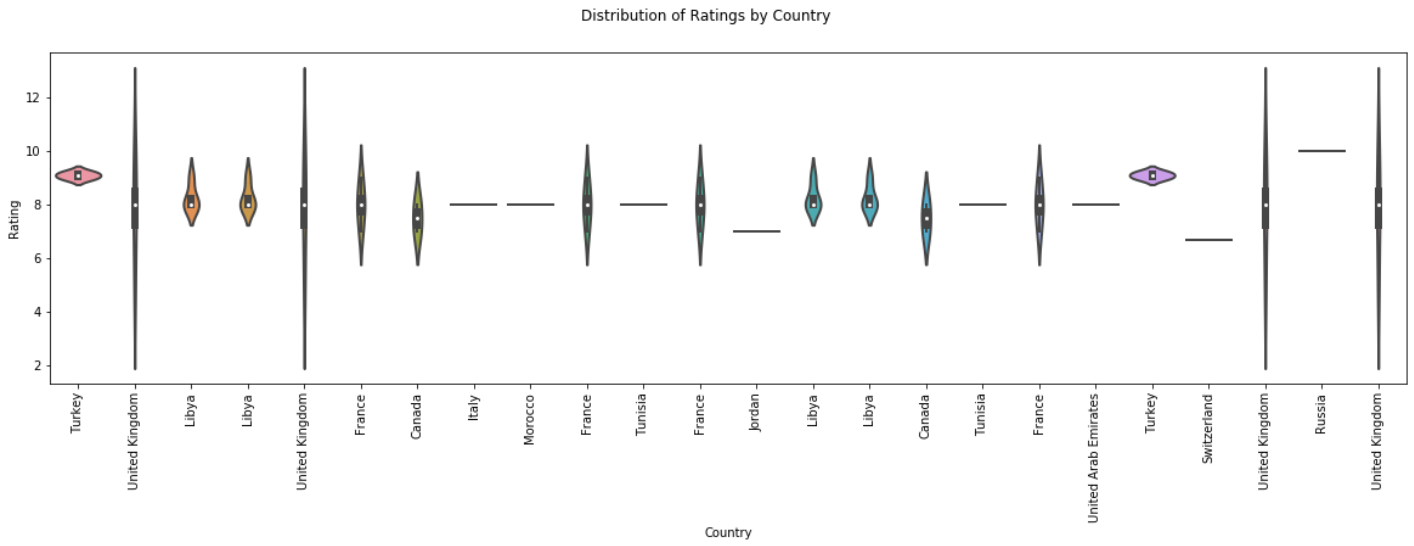
## 3.1.Distribution des avis positifs vs négatif

```
# DATA ANALYSIS
   # Pos vs Neg

fig = plt.figure()
ax = fig.add_axes([0,0,1,1])
values = [len(pos_reviews), len(neg_reviews)]
ax.pie(values, labels = ['Number of Positive Reviews', 'Number of Negative Reviews'],colors=['gold', 'lightcoral'],
 shadow=True,
 startangle=90,
 autopct='%1.2f%%')
ax.axis('equal')
plt.title('Positive Reviews Vs. Negative Reviews');
```




## 3.2. Graphique de violon des évaluations des clients pour le pays d'origine des 10 meilleurs évaluateurs



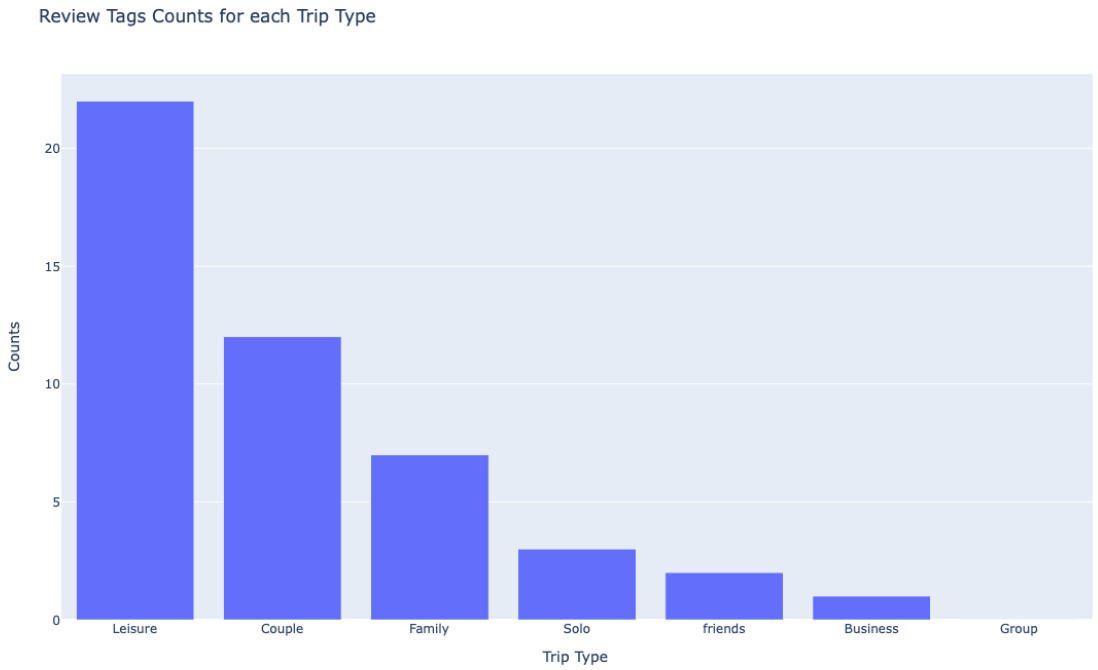
Le graphique ci-dessus est affiché dans l'ordre des comptes de révision de chaque pays. Il montre la relation entre les notations et le pays d'origine des évaluateurs.

À partir des éléments de graphe, nous voyons que la note médiane donnée par les évaluateurs de Libye, de Turkey et canada est un peu plus élevée que celle du reste des évaluateurs d'autres pays, tandis que la note médiane donnée par les évaluateurs du Royaume-Uni, la France est la plus faible.

La plupart des formes des distributions (maigres à chaque extrémité et larges au milieu) indiquent que les pondérations des notes données par les évaluateurs sont fortement concentrées autour de la médiane, qui est d'environ 8 à 9.

Cependant, nous avons probablement besoin de plus de données pour avoir une meilleure idée des distributions.

### 3.3. Répartition des étiquettes d'évaluation Nombre pour chaque type de voyage



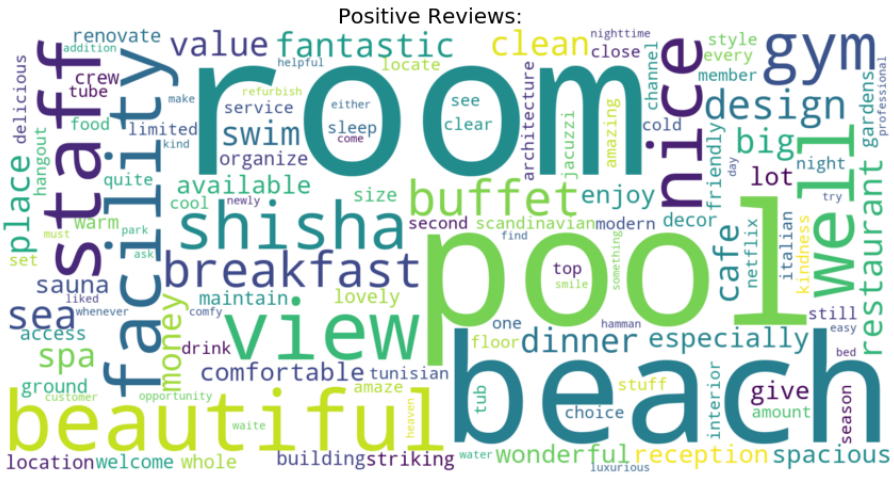

la plupart des gens venaient au Radisson Djerba pour les loisirs, que ce soit en famille ou en couple.

### 3.4. Créer Word Cloud pour les avis positifs et négatifs

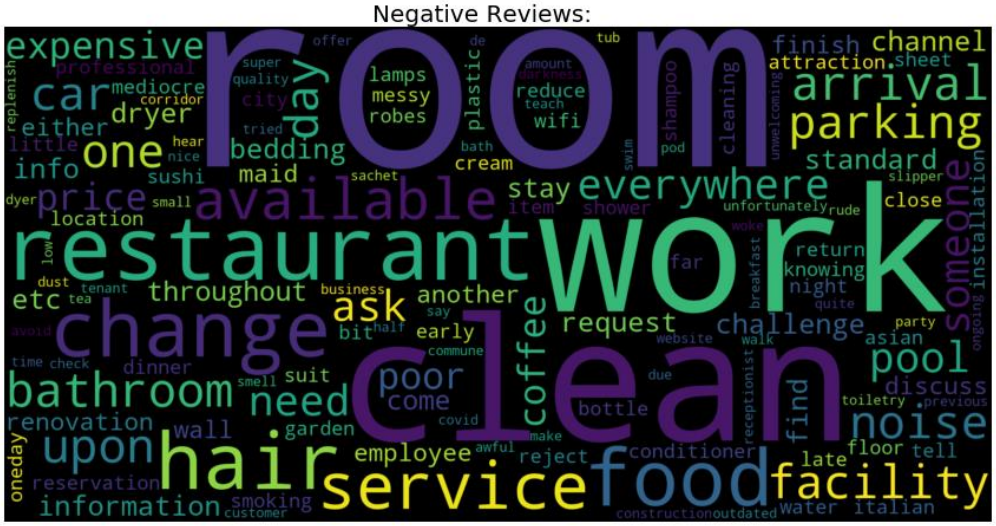
```
Wordcloud visuazliation
'''
text = review_df[review_colname].tolist()
text_str = ' '.join(lemmatized_tokens(' '.join(text))) #call function "lemmatized_tokens"
wordcloud = WordCloud(collocations = False, background_color = color, width=1600, height=800, margin=2,min_font_size
plt.figure(figsize = (15, 10))
   plt.imshow(wordcloud, interpolation = 'bilinear')
   plt.axis("off")
   plt.figtext(.5,.8,title,fontsize = 20, ha='center')
   plt.show()

      # Wordcoulds for Positive Reviews
 wordcloud(pos_reviews,'positive_reviews', 'white','Positive Reviews: ')
      # # WordCoulds for Negative Reviews
 wordcloud(neg_reviews,'negative_reviews', 'black', 'Negative Reviews:')
```

Positive Reviews:




Negative Reviews:



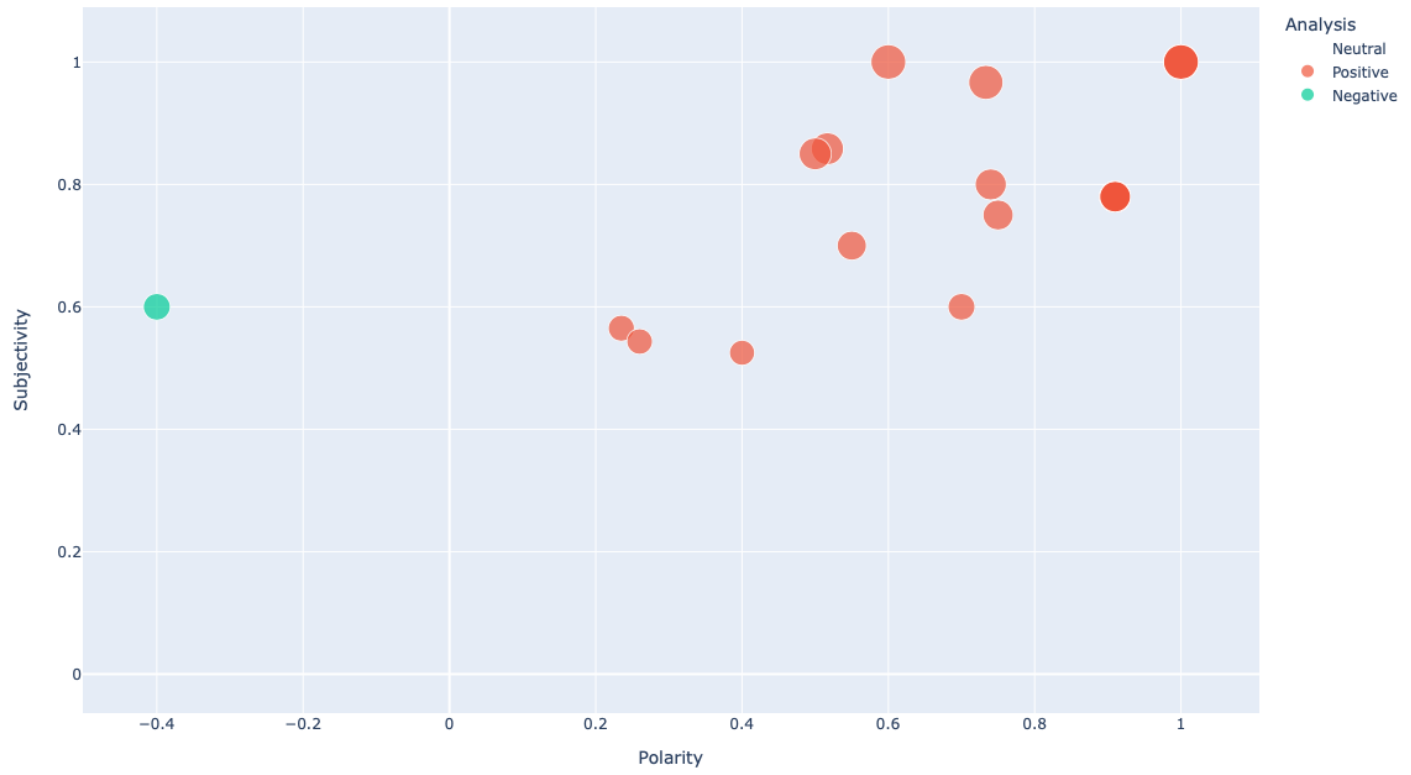

Dans les commentaires positifs, la plupart des gens sont très satisfaits de l'emplacement de la propriété près de la plage, personnel amical.

Dans les commentaires négatifs, nous avons également pu voir le mot Personnel, ce qui signifie qu'il y a aussi peu de membres du personnel dans la propriété dont le comportement n'était pas bon, les gens ne sont pas non plus satisfaits de la chambre et de la nourriture que l'entreprise devrait se concentrer sur résoudre.

## 4. Analyse des sentiments

L'analyse du sentiment est effectuée sur la colonne « Overall_review » de la base de données nommée « reviewer_info. ».

```
   #Create a function to get the subjectivity
def subjectivity(text):
 return TextBlob(text).sentiment.subjectivity
#Create a function to get the polarity
def polarity(text):
 return TextBlob(text).sentiment.polarity
#Create two new columns
reviewer_info['Subjectivity'] = reviewer_info['Overall_review'].apply(subjectivity)
reviewer_info['Polarity'] = reviewer_info['Overall_review'].apply(polarity)
###############################################################################
#Create a function to compute the negative, neutral and positive analysis
def getAnalysis(score):
 if score <0:
    return 'Negative'
 elif score == 0:
  return 'Neutral'
 else:
  return 'Positive'
reviewer_info['Analysis'] = reviewer_info['Polarity'].apply(getAnalysis)
###############################################################################
# plot the polarity and subjectivity
fig = px.scatter(reviewer_info, x='Polarity', y='Subjectivity', color = 'Analysis',size='Subjectivi
#add a vertical line at x=0 for Netural Reviews
#fig.update_layout(title='Sentiment Analysis',shapes=[dict(type= 'line',yref= 'paper', y0= 0, y1= 1
fig.show()
```




L'axe des X est la polarité et l'axe des Y est la subjectivité.

La polarité représente le degré de positif, de négatif ou de neutralité des avis et la subjectivité sont des opinions qui décrivent les sentiments des gens envers un sujet ou un sujet spécifique.

Plus la subjectivité est élevée, mieux elle décrit les sentiments des gens envers un sujet. Des points plus gros indiquent plus de subjectivité.

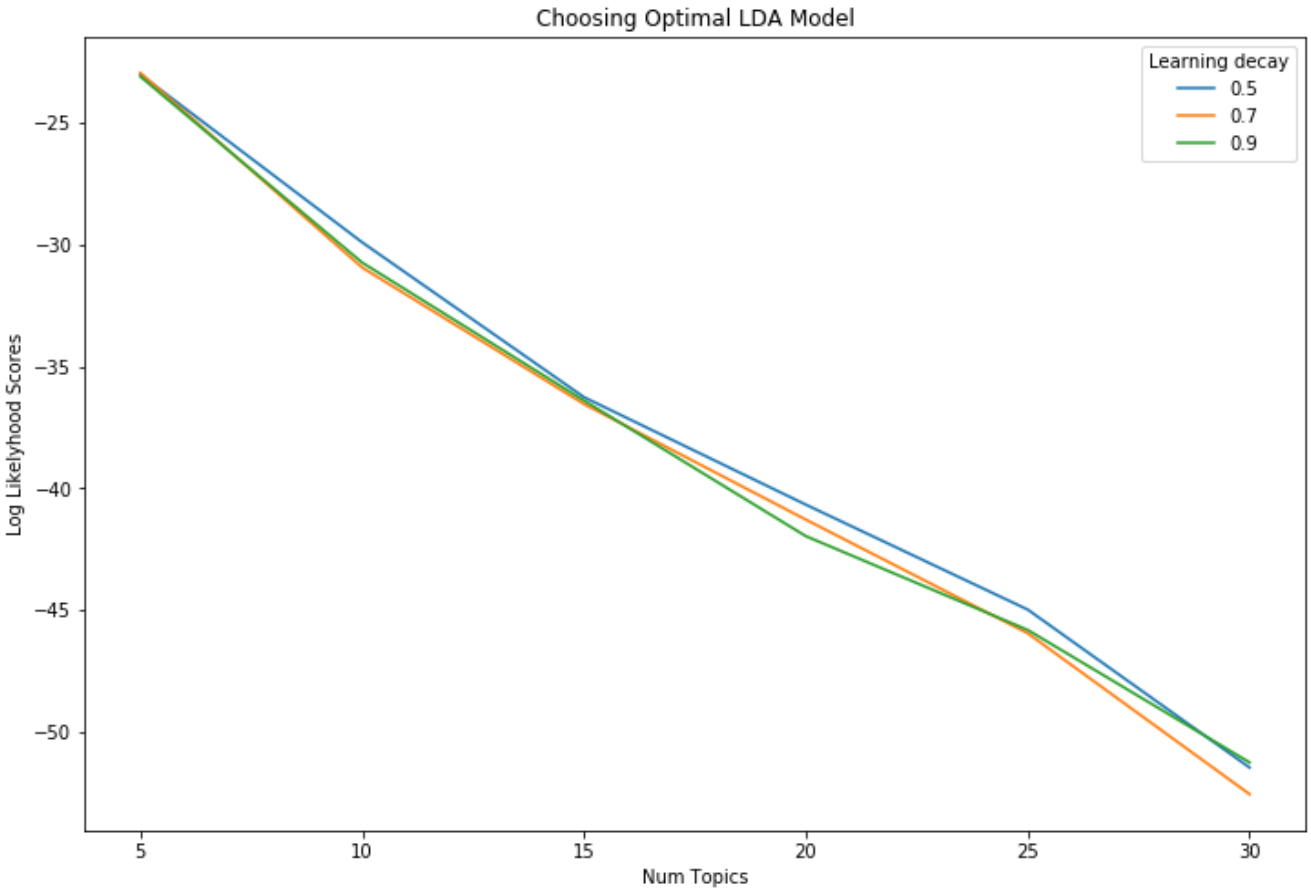
Nous avons pu voir que les avis positifs sont plus que des avis négatifs, mais nous devons savoir sur quel sujet les gens ont un sentiment négatif envers l'hôtel, nous allons donc faire de la modélisation thématique LDA.

Nous appliquerons le modèle LDA pour trouver la distribution du sujet et la forte probabilité de mot dans chaque sujet.

Ici, l'objectif de la modélisation des sujets LDA est de regarder les commentaires négatifs pour savoir quels sujets l'hôtel devrait se concentrer pour l'amélioration de la satisfaction de la clientèle.

Les étapes suivantes ont été effectuées pour effectuer la modélisation thématique LDA :

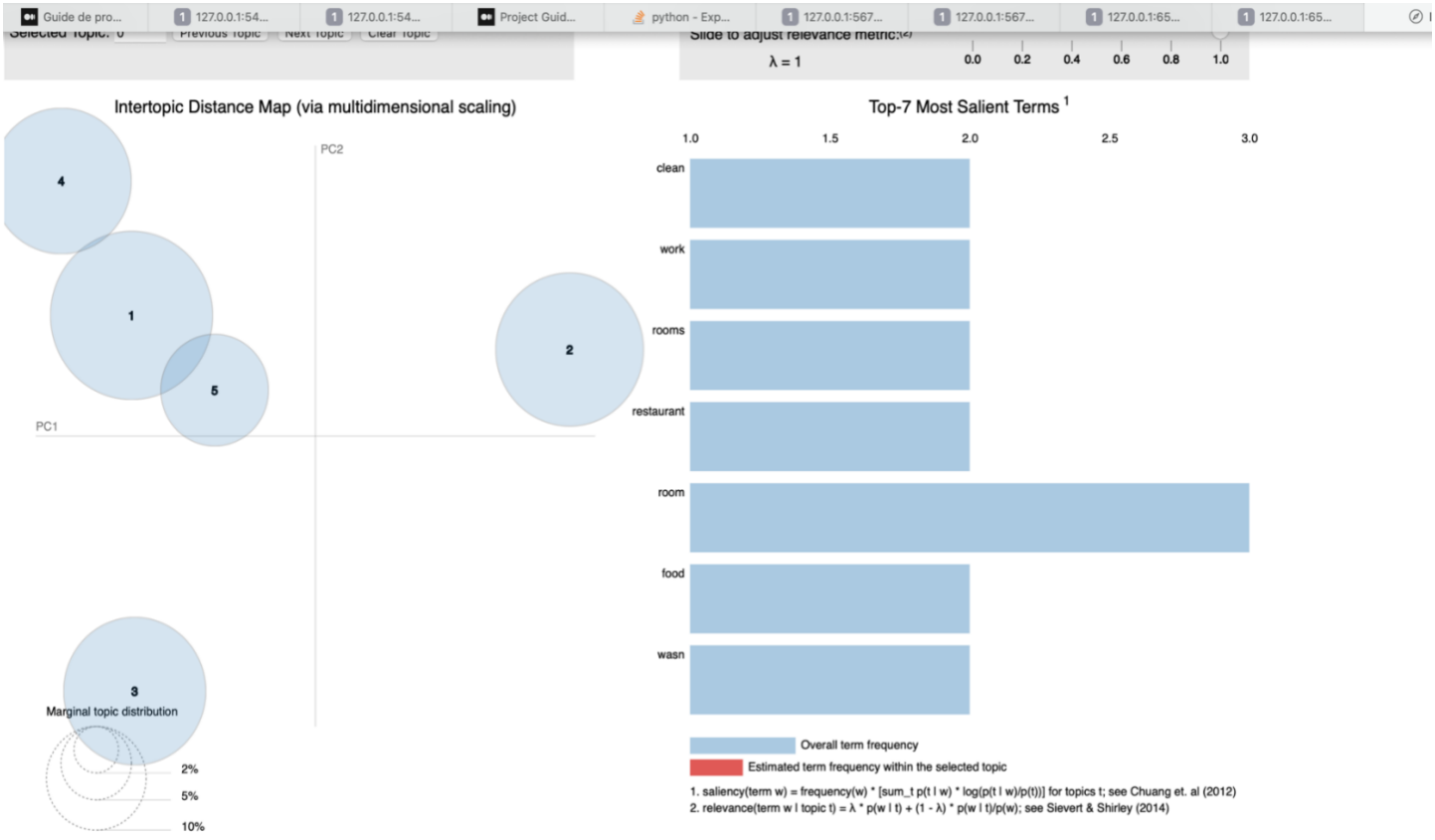
- Les commentaires ont été convertis en matrice document-term
- Trouver le modèle LDA optimal à l'aide de GridSearch et du réglage des paramètres
- Comparer les scores de performance du modèle LDA




D'après le graphique, nous voyons qu'il y a peu d'impact à choisir une décadence d'apprentissage différente avant 15 sujets, cependant, 5 sujets produiraient le meilleur modèle.Maintenant, produisons les mots dans les rubriques que nous venons de créer.

topics_data - DataFrame

| Index | Topic 1 words | Topic 1 weights | Topic 2 words | Topic 2 weights | Topic 3 words | Topic 3 weights | Topic 4 words | Topic 4 weights | Topic 5 words | Topic 5 weights |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | wasn | 2.3 | work | 2.5 | food | 1.0 | rooms | 3.2 | restaurant | 2.4 |
| 1 | clean | 1.4 | room | 1.2 | room | 0.8 | room | 2.3 | food | 1.4 |
| 2 | room | 0.2 | food | 0.2 | clean | 0.2 | clean | 1.2 | room | 0.2 |
| 3 | rooms | 0.2 | restaurant | 0.2 | rooms | 0.2 | wasn | 0.2 | work | 0.2 |
| 4 | food | 0.2 | clean | 0.2 | wasn | 0.2 | food | 0.2 | clean | 0.2 |
| 5 | work | 0.2 | rooms | 0.2 | work | 0.2 | work | 0.2 | rooms | 0.2 |
| 6 | restaurant | 0.2 | wasn | 0.2 | restaurant | 0.2 | restaurant | 0.2 | wasn | 0.2 |




Sur le côté gauche de la visualisation, chaque sujet est représenté par une bulle. Plus la bulle est grande, plus ce sujet est répandu où le numéro 1 est le sujet le plus populaire, et le numéro 5 étant le sujet le moins populaire. La distance entre deux bulles représente la similitude du sujet.

Le côté droit montre les 7 termes les plus pertinents pour le sujet que vous sélectionnez à gauche. La barre bleue représente la fréquence globale des termes, et la barre rouge indique la fréquence estimée des termes dans le sujet sélectionné. Donc, si vous voyez une barre avec à la fois du rouge et du bleu, cela signifie que le terme apparaît également sur d'autres sujets. Vous pouvez survoler le terme pour voir dans quel(s) sujet(s) le terme est également inclus.

Ainsi, par exemple, dans la sortie ci-dessus, nous avons pu voir que le sujet 3 consiste en un mot nommé « food», de sorte que l'entreprise devrait se concentrer sur l'amélioration de nourriture chambres.

# Conclusion

La solution réalisée a donc pu générer un ensemble d’indicateurs permettant de bien comprendre les avis des internautes et de se focaliser sur l’aspect sentimental souvent ignoré et difficile à faire sortir sans le recours à des outils comme l’IA.

# CONCLUSION GENERALE

Les résultats de ce projet peuvent aider la direction de l'hôtel à comprendre quelles mesures doivent être prises pour les touristes afin d'augmenter un plus grand nombre de touristes visitant l'hôtel et d'améliorer la satisfaction touristique, de sorte que ce projet puisse être réalisé par n'importe quel hôtel pour améliorer la satisfaction des clients. Ce projet peut également être utilisé pour obtenir des informations pour que les concurrents de l'hôtel comprennent la perception des clients envers leurs concurrents.

Connexion internet irréprochable, cloud, algorithmes, outils intelligents et applications digitales, la technologie représente un investissement nécessaire qu'il peut être intéressant pour les sociétés hôtelières d'intégrer dès aujourd'hui. Elle permet d'offrir aux clients une interaction optimisée tout en gagnant en productivité. Les applications digitales et outils technologiques représentent un gain de temps pour le personnel, des économies au niveau de la maintenance et se traduit également par un plus fort taux de remplissage, de conversion et de satisfaction client.

L'intelligence artificielle est en train de révolution le secteur de l'hôtellerie et d'offrir aux sociétés ayant déjà adopté des outils intelligents, notamment d'analyse de données, un avantage concurrentiel durable. Il s'agit d'un nouveau cycle technologique demandant aux hôtels de se moderniser pour pouvoir prendre le virage et gagner en efficacité opérationnelle.