07/01/2025

Timothy MARCIA, Elias AIT HASSOU

Master 1 Informatique - Université Lumière Lyon II

Rapport de projet

Développement d’un moteur de recherche en Python

Table des matières

[Présentation du projet 2](#_Toc187160186)

[Partie conceptuelle 3](#_Toc187160187)

[L’environnement de travail 3](#_Toc187160188)

[Les données 3](#_Toc187160189)

[Diagramme des classes 3](#_Toc187160190)

[Mise en place du projet 4](#_Toc187160191)

[Répartition des tâches 4](#_Toc187160192)

[Algorithmes spécifiques 4](#_Toc187160193)

[Les difficultés rencontrées 4](#_Toc187160194)

[Exemple d’utilisation 5](#_Toc187160195)

[Validation du logiciel 6](#_Toc187160196)

[Tests unitaires 6](#_Toc187160197)

[Tests globaux 6](#_Toc187160198)

[Conclusion et perspectives 7](#_Toc187160199)

# Présentation du projet

Ce projet a pour objectif de développer un moteur de recherche textuel sur des données touristiques en utilisant le langage de programmation Python. Le moteur de recherche permet aux utilisateurs de rechercher des activités qu'ils souhaitent réaliser et retourne des descriptions de villes où ces activités sont disponibles. Ce logiciel repose sur la collecte, le traitement, le stockage, l’analyse et la restitution des données.

Pour constituer le corpus de données, nous avons exploité l'API WikiVoyage. Cette API fournit des descriptions des activités touristiques proposées dans différentes villes à travers le monde. Ces données textuelles représentent la base de notre moteur de recherche.

L’objectif principal est de proposer une interface permettant :

* À l’utilisateur de saisir une activité spécifique qu’il aimerait réaliser.
* D'analyser les descriptions des villes contenues dans le corpus
* De restituer les résultats sous forme d’une liste des villes pertinentes accompagnées de leurs descriptions.

# Partie conceptuelle

## L’environnement de travail

Pour le développement de notre moteur de recherche, nous avons choisi un environnement de travail adapté aux exigences de collecte, traitement et d’analyse des données textuelles.

Le projet a été implémenté en Python, étant le langage de programmation étudié lors des séances de cours. Sa simplicité, et ses bibliothèques dédiées au traitement des données textuelles ont été essentielles pour mener à bien ce projet.

Les différentes bibliothèques utilisées ont des rôles spécifiques décrit ci-après :

* **Requests** pour interagir avec l’API WikiVoyage
* **Pandas**
* **JSON** pour lire et écrire des fichiers JSON tels que les réponses d’API ou nos fichiers de stockage des données.
* **NLTK** pour récupérer une liste des termes à faible valeur ajoutée dans la langue anglaise nous permettant de nettoyer le corpus.
* **NumPy**
* **Regex** pour mettre en place des expressions régulières nous permettant de nettoyer nos données textuelles en retirant les nombres et les caractères spéciaux.
* **Tkinter** pour réaliser l’interface graphique finale permettant d’effectuer une rechercher et afficher les résultats.

## Les données

## Diagramme des classes

# Mise en place du projet

## Répartition des tâches

## Algorithmes spécifiques

## Les difficultés rencontrées

L’une des difficultés rencontrées lors de ce projet a été la qualité des données textuelles. En effet, nous après avoir collecté les données et les avoir stockées dans un fichier JSON, nous avons implémenté une méthode permettant d’afficher des statistiques sur le corpus. Celle-ci nous a permis d’avoir un aperçu des données que nous avions à notre disposition.

Ayant récoltés des données sur des activités relatives à des villes dans le monde entier, certaines descriptions comportaient des caractères provenant d’autres alphabets et qui étaient donc inutile dans le cadre de notre projet. Afin de retirer toutes ces données inutiles, nous avons mis en place des expressions régulières afin de conserver uniquement les caractères de notre alphabet. Cette solution nous a permis également d’éviter d’allonger le temps de traitement des méthodes d’analyse du corpus.

Une fois que nous avions retiré ces caractères, nous avons décidé de mettre en place une méthode nous permettant de visualiser quelles étaient les dix termes les plus utilisés dans le corpus.

Une image contenant texte, capture d’écran, Police

Description générée automatiquementUne image contenant texte, capture d’écran, Police

Description générée automatiquement

Comme nous pouvons l’observer, avant le nettoyage des données, les dix termes les plus utilisés étaient des mots de liaison. Ces derniers n’étant pas pertinents dans le cadre de notre analyse, nous avons donc décidé de les retirer en utilisant la liste de ‘stopwords’ mise en place par la bibliothèque NLTK et en l’enrichissant avec d’autres termes spécifiques à notre corpus.

## Exemple d’utilisation

# Validation du logiciel

### Tests unitaires

### Tests globaux

# Conclusion et perspectives