Wirtualny Asystent Podróży

AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA
W KRAKOWIE

Autorzy
Hubert Asztabski asztabski@student.agh.edu.pl
Jakub Głowacki jglowacki@student.agh.edu.pl

26 czerwca 2024

Spis treści

1 Wstęp

Celem projektu Wirtualny Asystent Podróży jest stworzenie aplikacji, która pomoże użytkownikom w planowaniu i realizacji ich podróży. Asystent na podstawie dostarczonych preferencji użytkownika w postaci dokumentów tekstowych, wyszukuje potencjalnie najbardziej interesujące POI (Points of Interest) w wybranym mieście. Link do repozytorium: https://github.com/Sztaba/VirtualTravelAssistant

2 Wprowadzenie teoretyczne

W dzisiejszych czasach ilość dostępnych danych tekstowych rośnie w niezwykłym tempie. W związku z tym, analiza tekstu i wydobywanie z niego istotnych informacji stały się kluczowymi zadaniami w dziedzinie przetwarzania języka naturalnego (NLP). Jednym z najważniejszych aspektów tej analizy jest ocena znaczenia poszczególnych słów w dokumentach. Do tego celu często wykorzystuje się algorytmy takie jak TF-IDF i LSA, które pomagają zrozumieć kontekst i priorytetyzować informacje. W poniższej sekcji przedstawimy teoretyczne podstawy tych metod.

2.1 TF-IDF (Term Frequency-Inverse Document Frequency)

TF-IDF (Term Frequency-Inverse Document Frequency) jest jedną z najczęściej stosowanych metod do oceny znaczenia słów w dokumentach. Algorytm ten łączy dwie miary: częstość terminu (TF) oraz odwrotną częstość dokumentów (IDF).

Częstość Terminu (TF)

Częstość terminu (TF) mierzy, jak często dany termin pojawia się w dokumencie. Jest to miara lokalnej ważności słowa w dokumencie. Można ją obliczyć jako stosunek liczby wystąpień terminu do łącznej liczby słów w dokumencie:

$$\mathrm{TF}(t, d) = \frac{f_{t,d}}{n_d} \tag{1}$$

gdzie:

- $f_{t,d}$ oznacza liczbę wystąpień terminu t w dokumencie d,
- n_d oznacza łączną liczbę słów w dokumencie d.

Odwrotna Częstość Dokumentów (IDF)

Odwrotna częstość dokumentów (IDF) mierzy, jak unikalne jest słowo w całym korpusie dokumentów. Terminy, które pojawiają się w wielu dokumentach, są mniej znaczące. IDF można obliczyć jako logarytmicznie przeskalowaną odwrotność liczby dokumentów, w których pojawia się termin:

$$\mathrm{IDF}(t, D) = \log\left(\frac{N}{|\{d \in D : t \in d\}|}\right) \tag{2}$$

gdzie:

- N oznacza łączną liczbę dokumentów w korpusie,
- $|\{d \in D : t \in d\}|$ oznacza liczbę dokumentów zawierających termin t.

Wskaźnik TF-IDF

Wskaźnik TF-IDF jest iloczynem TF i IDF:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D) \tag{3}$$

Dzięki temu podejściu, TF-IDF pozwala na ocenę znaczenia słowa zarówno w kontekście pojedynczego dokumentu, jak i całego korpusu dokumentów.

2.2 Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA), znana również jako Latent Semantic Indexing (LSI), to technika redukcji wymiarowości używana do odkrywania ukrytych relacji pomiędzy słowami w dużych korpusach tekstowych. LSA opiera się na analizie współwystępowania słów, umożliwiając identyfikację wzorców tematycznych w dokumentach.

Proces LSA można podzielić na kilka kluczowych etapów:

- Tworzenie macierzy term-dokument: Na początku tworzona jest macierz term-dokument, w której wiersze reprezentują terminy, a kolumny dokumenty. Wartości w macierzy odpowiadają częstościom występowania terminów w dokumentach.
- Redukcja wymiarowości za pomocą SVD: Następnie na macierzy term-dokument wykonywana jest dekompozycja wartości osobliwych (SVD), która rozkłada macierz na trzy mniejsze macierze: U, Σ i V. Macierze te reprezentują terminy, wartości osobliwe oraz dokumenty w zredukowanej przestrzeni wymiarowej.
- Analiza tematów: W zredukowanej przestrzeni wymiarowej, terminy i dokumenty mogą być grupowane na podstawie ich podobieństw, co pozwala na identyfikację tematów ukrytych w korpusie.

Dzięki LSA możliwe jest wydobycie tematów z tekstu, które nie są bezpośrednio widoczne, co pozwala na bardziej zaawansowaną analizę treści.

3 Algorytm Wyznaczania POI

W naszym projekcie, Wirtualny Asystent Podróży wykorzystuje algorytmy TF-IDF oraz LSA do analizy dostarczonych plików testowych i identyfikacji punktów zainteresowania (POI). Proces ten składa się z kilku kluczowych etapów, które opisano poniżej.

1. **Przygotowanie danych wejściowych**: Po otrzymaniu plików testowych, dane są przetwarzane i tokenizowane. Tokenizacja polega na rozbiciu tekstu na poszczególne terminy (słowa kluczowe).

2. **Obliczanie TF-IDF**: Dla każdego terminu w dokumentach obliczane są wartości TF-IDF.

3. **Przeprowadzanie LSA**: Następnie, na podstawie współwystępowania terminów w dokumentach, wykonywana jest analiza latentnych zmiennych semantycznych (LSA).

4. **Selekcja POI**: W wyniku analizy TF-IDF oraz LSA powstają dwa zestawy wyników w postaci tabelarycznej (DataFrame). Wyniki te są następnie przekazywane do funkcji selekcjonującej POI. Każdy punkt zainteresowania (POI) jest analizowany pod kątem zgodności z tematami i terminami zidentyfikowanymi w krokach wcześniejszych na podstawie przypisanych do punktu tagów:

   (a) Sprawdzane jest, czy tagi POI występują w wynikach LSA.

   (b) Obliczana jest sumaryczna punktacja TF-IDF dla tagów POI.

   (c) POI jest wybierane, jeśli jego punktacja TF-IDF jest większa od zera lub jeśli występuje zgodność z tematami LSA.

5. **Dodawanie informacji do wybranych POI**: Wybrane POI są wzbogacane o dodatkowe informacje, takie jak sumaryczna punktacja TF-IDF oraz zgodność z tematami LSA.

6. **Sortowanie POI**: Na końcu, wyselekcjonowane POI są sortowane według następujących kryteriów:

   (a) Długość najdłuższej listy tematów LSA przypisanej do POI.

   (b) Liczba tematów LSA przypisanych do POI.

   (c) Sumaryczna punktacja TF-IDF.

   (d) Ocena punktu zainteresowania.

Dzięki temu podejściu, nasz Wirtualny Asystent Podróży uzyskuje posortowaną listę punktów od najważniejszych do najmniej ważnych.

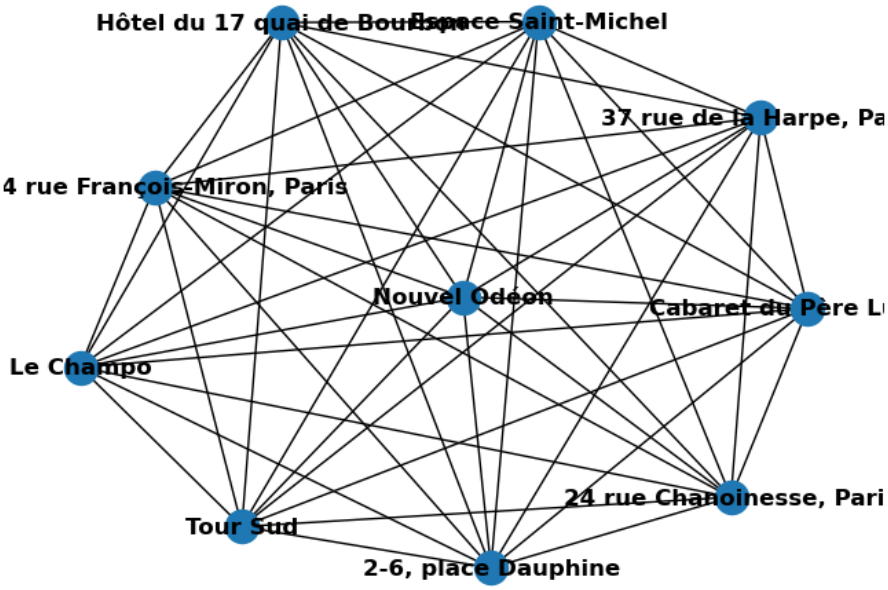
4 Wyznaczanie Trasy

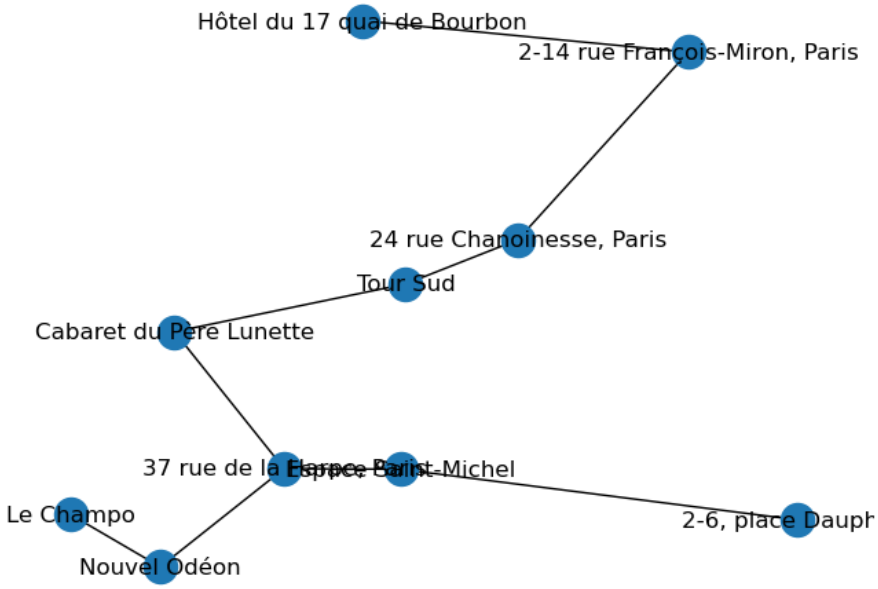
Algorytm Wyznaczania trasy składa się z:

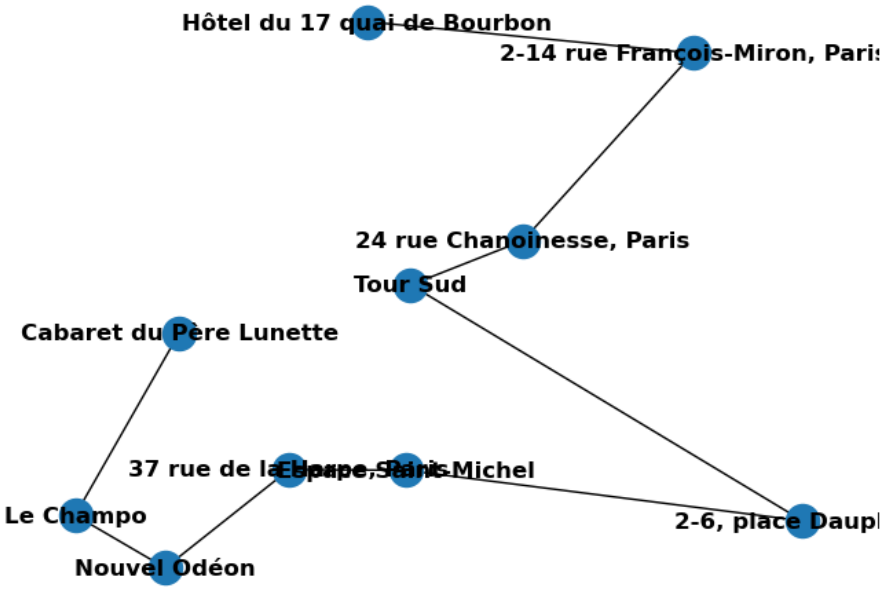
- Pobranie POI i zaprezentowanie grafu pełnego,
- Wyznaczenie minimalnego drzewa rozpinającego
- Znalezienie ścieżki Eulera
- Wyznaczenie wstępnej najkrótszej drogi.
- Ulepszenie najkrótszej drogi metodą two-opt.

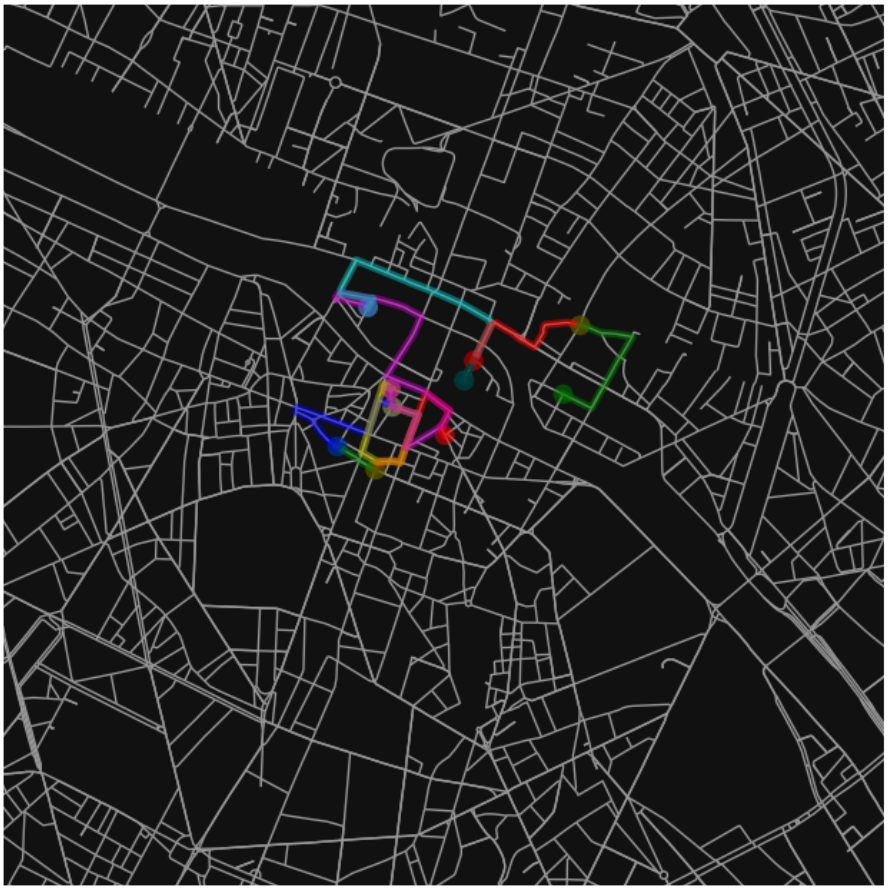
5 Wyniki jakie udało się uzyskać

5.1 Paryż

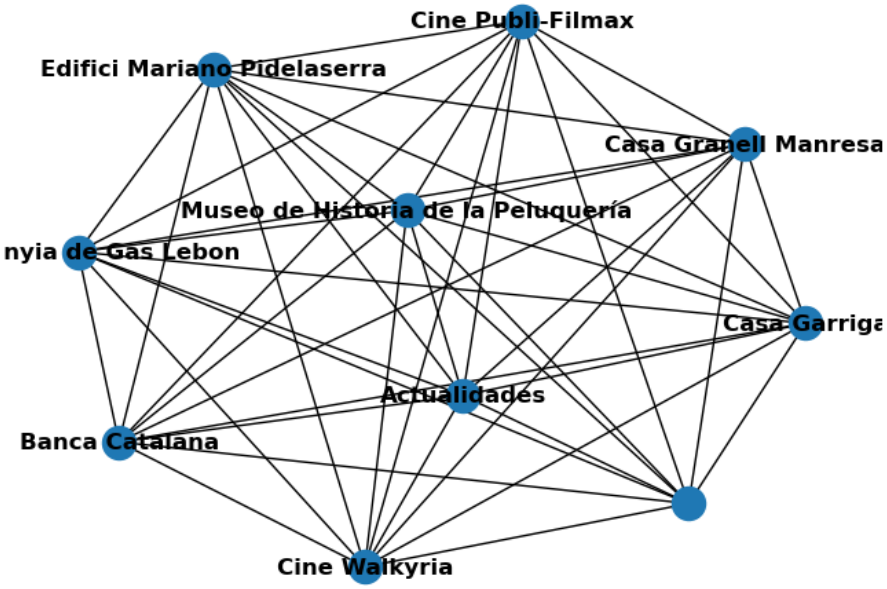


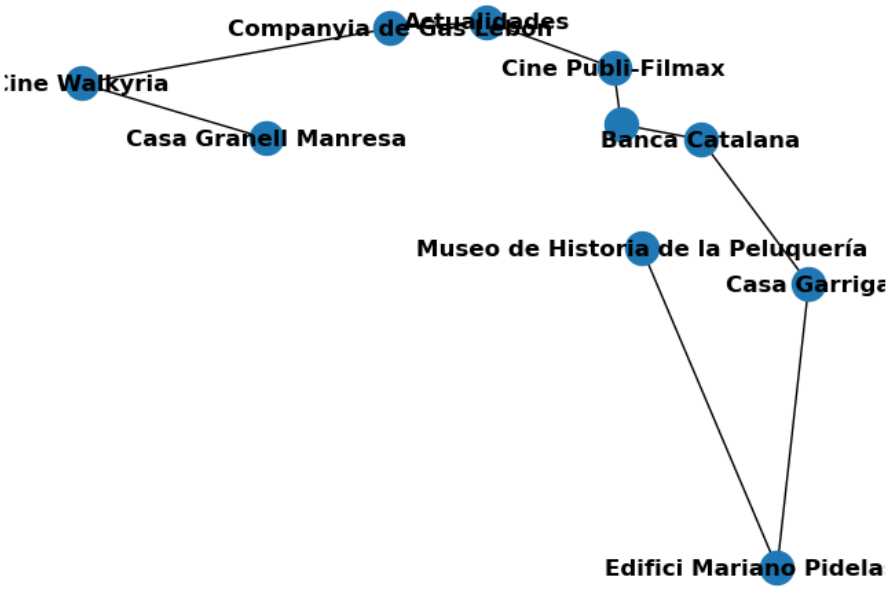



Hôtel du 17 quai de Bourbon
2-14 rue François-Miron, Paris
24 rue Chanoinesse, Paris
Tour Sud
Cabaret du Père Lunette
37 rue de la Huchette, Paris
Le Champo
Nouvel Odéon
2-6, place Daupl



5.2 Barcelona




Companyia de Gas Lebon
Actualidades
Cine Publi-Filmax
ine Walkyria
Casa Granell Manresa
Banca Catalana
Museo de Historia de la Peluquería
Casa Garriga
Edifici Mariano Pidela

6 Interfejs użytkownika

Interfejs użytkownika składa się z dwóch głównych sekcji: pola wprowadzania danych użytkownika oraz panelu sterowania. W pierwszej sekcji użytkownik może wprowadzić swoje imię, nazwisko, miasto i wybrać pliki, a następnie kliknąć przycisk "Submit"w celu przetworzenia danych. Druga sekcja zawiera pole "Distance"i przycisk "Zoom", które odblokowują się po wczytaniu mapy, umożliwiając dalszą interakcję.

Rysunek 1: Interfejs użytkownika