**5. Interfejsy użytkownika:**

* **Wejście:**
  + Plik z danymi w formacie .csv.
* **Wyjście:**
  + Chmury słów i wykresy:
    - Chmura słów dla recenzji pozytywnych.
    - Chmura słów dla recenzji negatywnych.
    - Wykresy Top Terms w tematach LDA dla recenzji pozytywnych i negatywnych.
  + Lista dziesięciu najczęściej występujących słów w recenzjach pozytywnych i negatywnych.

**6. Wymagania dotyczące danych**

* System przyjmuje dane w formacie .csv, kodowane w UTF-8**.**
* Skrypt zakłada, że dane tekstowe są w języku angielskim.
* System wymaga kolumn:
  + Hotel\_Name.
  + Postive\_Review.
  + Negative\_Review.
* System nie obsługuje plików większych niż 100 MB.
* System usuwa puste recenzje oraz recenzje zawierające wyłącznie stopwords lub znaki specjalne.

**Słownictwo dokumentacji:**

* **Recenzja**: Opinia użytkowników na temat hotelu, zawarta w pliku CSV w kolumnach Negative\_Review i Positive\_Review.
* **Czyszczenie tekstu:** Proces usuwania z tekstu znaków specjalnych, linków, cyfr, stopwords, wielokrotnych spacji oraz konwersji tekstu do małych liter.
* **LDA (Latent Dirichlet Allocation):** Algorytm statystyczny służący do modelowania tematów w zbiorze dokumentów, który przypisuje tematy na podstawie współwystępowania słów w tekstach.

**Przypadki użycia (use cases)**

* **Użytkownik:**
  + Wczytuje plik .csv z recenzjami hotelowymi.
  + Uruchamia analizę tekstu i modelowanie tematów.
  + Wyświetla wykresy i chmury słów.
  + Przegląda listę najczęściej używanych słów w recenzjach pozytywnych i negatywnych.
* **Skrypt/system:**
  + Wczytuje dane z pliku .csv.
  + Czyści dane tekstowe z niepotrzebnych znaków, linków, stopwords itd.
  + Dzieli dane na recenzje pozytywne i negatywne.
  + Tworzy korpusy i przekształca je do formatu akceptowanego przez algorytm LDA.
  + Wykonuje topic modeling osobno dla pozytywnych i negatywnych recenzji.
  + Wyświetla w konsoli dziesięć najczęściej używanych słów w każdej grupie recenzji.
  + Generuje:
    - Wykresy Top Terms dla każdego z tematów.
    - Chmurę słów dla recenzji pozytywnych.
    - Chmurę słów dla recenzji negatywnych.

**Testowe przypadki użycia:**

* Test z plikiem .csv zawierającym tylko pozytywne recenzje.
* Test z plikiem .csv zawierającym tylko negatywne recenzje.
* Test z plikiem .csv zawierającym zarówno pozytywne, jak i negatywne recenzje.
* Test z plikiem .csv zawierającym puste lub niepełne recenzje.

**Scenariusze użytkownika (user stories)**

**Scenariusz 1: Przeprowadzenie analizy różnicy opinii między hotelami**

* **Jako**: Analityk sieci hoteli w agencji turystycznej
* **Chcę**: Przeprowadzić analizę, jakie tematy i słowa pojawiają się w recenzjach dla poszczególnych hoteli.
* **Aby**: Przedstawić raport, które hotele powinny być rozważane przy tworzeniu ofert wyjazdów.

**Kryteria akceptacji:**

* System umożliwia filtrowanie danych po kolumnie Hotel\_Name.
* Użytkownik może wyświetlić wyniki dla wskazanego hotelu.
* Wygenerowane chmury słów i wykresy przedstawiają wybrany zakres danych.
* Użytkownik ma dostęp do Top Terms i chmur słów z podziałem na hotele.

**Scenariusz 2: Wyszukiwanie najlepszego hotelu na wakacje**

* **Jako**: Turysta planujący wakacje bez udziału agencji turystycznych
* **Chcę**: Przeanalizować dostępne na portalach hoteli ze względu na ich recenzje.
* **Aby**: Wybrać hotel, który najlepiej odpowiada moim oczekiwaniom i unikać miejsc z powtarzającymi się problemami

**Kryteria akceptacji:**

* Użytkownik może wczytać plik .csv z recenzjami hoteli.
* System przeprowadza analizę tekstów oraz analizę tematów dla pozytywnych i negatywnych recenzji.
* System generuje chmury słów i wykresy tematów, dzięki którym użytkownik może szybko zorientować się, co jest najczęściej chwalone i krytykowane w danym hotelu.
* Użytkownik może porównać kilka hoteli pod pozytywnych i negatywnych opinii.