**Wprowadzenie**

W naszej pracy postanowiliśmy sprawdzić, czy restauracja “U Fukiera” szanowanej szefowej kuchni Magdy Gessler jest doceniania przez klientów i czy rzeczywiście spełnia ich oczekiwania. Skorzystaliśmy z opinii zebranych ze strony [tripadvisor.com](http://tripadvisor.com), a następnie zaimportowanych do jednego pliku .CSV. Do analizy tekstu wykorzystaliśmy dwa narzędzia, które naszym zdaniem najlepiej pozwolą nam opisać dane. Wybraliśmy modelowanie tematów, dzięki czemu zidentyfikujemy podgrupy słów, na których skupiają się klienci, oraz asocjacje, które pozwolą nam znaleźć słowa najbardziej skorelowane z danych terminem.

Nasz program naturalnie można wykorzystać przy analizie dowolnej usługi ze względu na swoją uniwersalność.

**Cele systemu**

* Analiza opinii dotyczących danej usługi (w naszym wypadku restauracja “U Fukiera” Magdy Gessler)
  + ocena jakości danej firmy
  + zidentyfikowanie obszarów wymagających poprawy
* Analiza modelowania tematów, sprawdzenie jak kształtują się słowa w podziale na podgrupy
  + pozwoli nam łatwiej przewidzieć na jakiś tematach skupiają się klienci
* Analiza asocjacji związanych z terminami: quality, decorations, taste, romantic, menu, dumpling
  + pozwoli nam sprawdzić jakość poszczególnych komponentów danej usługi

**Wymagania funkcjonalne - co robi system, jakie zadania wykonuje**

1. Zlicza częstotliwość występowania danych słów
2. Tworzy wizualną prezentację danych w postaci chmury słów
3. Modeluje tematy przy użyciu ukrytej alokacji Dirichleta
4. Tworzy asocjacje, czyli znajduje słowa najbardziej skorelowane z danym terminem w macierzy TDM/DTM
5. Wizualizuje asocjacje przy pomocy wykresu

**Wymagania niefunkcjonalne - jak działa system, właściwości**

1. Estetyczna prezentacja danych
2. Szybkość wykonania kodu ze względu na optymalny kod
3. Niezawodność w wykonaniu kodu dla odpowiednich danych
4. Łatwość obsługi dla innych danych
5. Przystępność (prezentacja danych, z której łatwo wyciągnąć wnioski)

**Interfejsy użytkownika i wymagania dotyczące danych**

**Wejście:**

* Plik tekstowy .CSV z opiniami

**Wyjście:**

* Tabela częstości słów
* Chmura słów
* Wykresy ważności słów dla danej ilości tematów w modelu LDA
* Spis korelacji o wybranym poziomie dla wybranych słów
* Wykresy lizakowe z natężeniem korelacji dla wybranych słów

**Wymagania dotyczące danych:**

* dane tekstowe z plików .CSV
* dane tekstowe w języku angielskim

**Słownictwo dokumentacji**

* **Korpus** - zbiór dokumentów używanych do analizy
* **Chmura słów** - specyficzny sposób prezentowania danych tekstowych, im dane słowo częściej pojawia się w tekście, tym jest większe na wizualizacji
* **Macierz TDM** - macierz, gdzie w kolumnach są nazwy dokumentów, w wierszach słowa a wartości w komórkach są częstościami występowania danego słowa w danym dokumencie
* **Macierz DTM** - macierz, gdzie w kolumnach są słowa, w wierszach nazwy dokumentów a wartości w komórkach są częstościami występowania danego słowa w danym dokumencie
* **Utajniona alokacja Dirichleta** (LDA) – nienadzorowany algorytm uczenia maszynowego pomagający zidentyfikować tematy w zbiorze tekstów
* **Temat –** zbiór słów powiązanych ze sobą i często występujących razem
* **Korelacja** – określa wzajemne powiązania pomiędzy wybranymi zmiennymi, w naszym przypadku miara mówi, jak często dane słowa występują w tych samych dokumentach lub w dokumentach o podobnych tematach

**Przypadki użycia (use cases)**

Użytkownik:

* Wczytuje plik .CSV
* Definiuje liczbę tematów
* Wybiera słowo lub słowa do analizy
* Uruchamia analizę
* Wizualizuje wyniki
* Generuje wykresy i raport html

System:

* Wczytuje pliki tekstowe
* Tworzy korpus
* Przetwarza i czyści dane
* Tokenizuje dane
* Zlicza częstość słów i wyświetla te, które występowały najczęściej
* Tworzy chmury słów
* Modeluje tematy i tworzy wykresy słów dla wybranej liczby tematów w modelu LDA
* Znajduje korelacje o wyznaczonym współczynniku dla wybranych słów
* Tworzy wykresy lizakowe z natężeniem dla korelacji dla wybranych słów

**Scenariusze użytkownika (user stories)**

Scenariusz 1. Wczesne wykrywanie ryzyk reputacyjnych na podstawie modelowania tematów i analizy asocjacji

Jako: Analityk danych w dziale Customer Experience

Chcę: Wykryć ewentualne ryzyka reputacyjne analizując opinie klientów zamieszczone na portalu

Aby: Zidentyfikować obszary ryzyk reputacyjnych, by móc im przeciwdziałać już na wczesnym etapie.

Kryteria akceptacji:

* Użytkownik może załadować plik .CSV z opiniami
* Skrypt modeluje tematy i przeprowadza analizę
* Skrypt generuje wykresy słów dla tematów
* Skrypt wykrywa korelacje między słowami
* Skrypt generuje wykres z danymi korelacjami
* Użytkownik samodzielnie ma możliwość ocenić czy słowa wskazane w korelacji do danej usługi są pozytywne czy negatywne oraz czy grożą one potencjalnym kryzysom reputacyjnym

Scenariusz 2. Analiza związków między słowami w opiniach klientów

Jako: Analityk danych w dziale Customer Experience

Chcę: Przeanalizować opinie klientów pod kątem powiązań nimi

Aby: Zidentyfikować jakość całości usługi oraz poszczególnych jej elementów

Kryteria akceptacji:

* Użytkownik może załadować plik .CSV z opiniami
* Skrypt modeluje tematy i przeprowadza analizę
* Skrypt pokazuje jakie słowa są najczęściej poruszane w opiniach
* Skrypt wykrywa korelacje między słowami
* Skrypt generuje wykres z danymi korelacjami
* Użytkownik może zidentyfikować jakie słowa korelują z danym elementem usługi - czy są to słowa pozytywne, czy negatywne