🌐 Topical Focus Analyzer - Przewodnik Użytkownika

💡 Zrozumienie Site Focus Score & Site Radius Score

Zanim zagłębimy się w narzędzie, wyjaśnijmy kluczowe metryki, które oblicza:

- **Site Focus Score (0-100):** Ten wynik mierzy, jak bardzo podobne tematycznie są do siebie adresy URL (reprezentujące strony) w całej witrynie. **Wyższy wynik** wskazuje, że treść witryny jest ściśle zgrupowana wokół określonych tematów, wykazując silną spójność tematyczną i specjalizację. Niższy wynik sugeruje szerszą lub bardziej rozproszoną tematykę.
- **Site Radius Score (0-100):** Ten wynik odzwierciedla, jak ściśle reprezentacje stron grupują się wokół centralnego tematu witryny (obliczonego "centroidu" lub koncepcji "site embedding"). **Wyższy wynik** oznacza, że treść jest zgrupowana blisko siebie (mniejszy "promień"), podczas gdy niższy wynik oznacza, że tematy treści są bardziej rozproszone lub zróżnicowane wzdłuż/wokół centrum witryny.

Dlaczego jest to ważne? Analiza tych wyników pomaga:

- **Poprawić SEO:** Wyszukiwarki preferują witryny z wyraźnym autorytetem tematycznym. Wysokie wyniki **Site Focus Score** i **Site Radius Score** mogą sygnalizować wiedzę specjalistyczną i trafność w określonych obszarach.
- **Udoskonalić strategię treści:** Zrozumieć, czy Twoja treść jest zgodna z zamierzoną niszą, czy też staje się zbyt szeroka. Zidentyfikować luki lub obszary wymagające pogłębienia, aby wzmocnić główne tematy witryny.
- **Poprawić doświadczenie użytkownika:** Dobrze sfokusowana witryna jest często łatwiejsza dla użytkowników w nawigacji i zrozumieniu jej celu.
- **Analizować konkurencję:** Porównać fokus tematyczny i spójność Twojej witryny z konkurencją.

1. 🚀 Wprowadzenie

Witaj w Analizatorze Fokusu Tematycznego! To narzędzie pomaga zrozumieć strukturę witryny internetowej i jej tematykę poprzez analizę map strony oraz, opcjonalnie, treści poszczególnych stron. Dostarcza informacji na temat stopnia skupienia tematycznego witryny (wykorzystując wyjaśnione powyżej **Site Focus Score** i **Site Radius Score**), wizualizuje powiązania między treściami i identyfikuje potencjalne duplikacje treści (kanibalizację).

Ten przewodnik obejmuje konfigurację aplikacji (opartej na wersji `multi_sitemap_app.py`) oraz korzystanie z jej funkcji.

2. 📋 Wymagania wstępne

Zanim zaczniesz, upewnij się, że posiadasz:

- **Python:** Zainstalowaną wersję 3.9 lub nowszą. Możesz pobrać ją ze strony python.org.
- **Klucz API Google AI (Opcjonalnie):** Potrzebny tylko, jeśli chcesz korzystać z funkcji podsumowania generowanego przez AI. Klucz można uzyskać w Google AI Studio.

3. 🛠️ Instalacja i Konfiguracja

Postępuj dokładnie według tych kroków w swoim terminalu lub wierszu poleceń.

1. **Utwórz katalog projektu:**

   Stwórz folder dla aplikacji i przejdź do niego.

   ```
   mkdir topical-focus-analyzer
   cd topical-focus-analyzer
   ```

2. **Utwórz i aktywuj środowisko wirtualne:**

   Izoluje to zależności aplikacji.

   Utwórz:

   ```
   python -m venv venv
   ```

   Aktywuj:

   W systemie Windows (Wiersz poleceń/PowerShell):

   ```
   venv\Scripts\activate
   ```

   W systemach macOS/Linux (Bash/Zsh):

   ```
   source venv/bin/activate
   ```

   Na początku wiersza poleceń powinien pojawić się przedrostek `(venv)`.

3. **Utwórz plik `requirements.txt`:**

   W katalogu `topical-focus-analyzer` utwórz plik o nazwie dokładnie `requirements.txt`. Wklej do niego następującą zawartość:

   ```
   # Główne biblioteki
   requests
   beautifulsoup4
   lxml
   pandas
   numpy>=1.24.4
   scikit-learn>=1.4.2
   plotly
   streamlit
   python-dotenv

   # Podsumowania AI
   google-generativeai

   # Statystyki treści (opcjonalnie)
   trafilatura
   rtree
   ```

   Określone wersje `numpy` i `scikit-learn` zostały podane dla lepszej kompatybilności, zgodnie z ustaleniami podczas testów. Biblioteka `trafilatura` jest zalecana, ale opcjonalna; aplikacja posiada mechanizm zastępczy, jeśli jej instalacja się nie powiedzie.

4. **Zainstaluj zależności:**

   Uruchom to polecenie, gdy Twoje środowisko wirtualne jest aktywne:

   ```
   pip install -r requirements.txt
   ```

5. **Utwórz strukturę plików projektu:**

   Utwórz niezbędny podkatalog i puste pliki Python. Rzeczywisty kod dla tych plików powinien zostać uzyskany na podstawie wcześniejszych rozmów dotyczącej rozwoju (ten przewodnik skupia się na konfiguracji).

   Utwórz podkatalog `modules`:

   ```
   mkdir modules
   ```

   (Użyj `md modules` w Wierszu poleceń Windows, jeśli `mkdir` zawiedzie)

   Utwórz następujące puste pliki:

   W systemach macOS/Linux:

   ```
   touch modules/__init__.py
   touch modules/sitemap_finder.py
   touch modules/sitemap_parser.py
   touch modules/content_extractor.py
   touch modules/simple_vectorizer.py
   touch modules/dimensionality_reducer.py
   touch modules/analyzer.py
   touch modules/llm_summarizer.py
   touch multi_sitemap_app.py
   touch .env
   ```

   W systemie Windows (PowerShell):

   ```
   New-Item -ItemType File -Path "modules\__init__.py" -Force
   New-Item -ItemType File -Path "modules\sitemap_finder.py" -Force
   New-Item -ItemType File -Path "modules\sitemap_parser.py" -Force
   New-Item -ItemType File -Path "modules\content_extractor.py" -Force
   New-Item -ItemType File -Path "modules\simple_vectorizer.py" -Force
   New-Item -ItemType File -Path "modules\dimensionality_reducer.py" -Force
   New-Item -ItemType File -Path "modules\analyzer.py" -Force
   New-Item -ItemType File -Path "modules\llm_summarizer.py" -Force
   New-Item -ItemType File -Path "multi_sitemap_app.py" -Force
   New-Item -ItemType File -Path ".env" -Force
   ```

   Musisz wypełnić te pliki (zwłaszcza `multi_sitemap_app.py` i pliki wewnątrz `modules/`) kodem Python opracowanym wcześniej.

4. ⚙️ Konfiguracja

Skonfiguruj opcjonalną funkcję Podsumowania AI:

1. Otwórz plik `.env` znajdujący się w katalogu `topical-focus-analyzer`.
2. Dodaj swój klucz API Google AI w następującym sposób (zastąp `your_google_api_key_here` swoim rzeczywistym kluczem):

   ```
   GOOGLE_API_KEY="your_google_api_key_here"
   ```

3. Zapisz i zamknij plik.

Jeśli nie dodasz klucza lub klucz będzie nieprawidłowy, opcja "Generate AI Summary" w aplikacji będzie wyłączona.

5. ▶️ Uruchamianie Aplikacji

1. **Upewnij się, że środowisko wirtualne jest aktywne:** Jeśli zamknąłeś terminal, wróć do katalogu projektu i aktywuj je ponownie (patrz Krok 3.2).
2. **Uruchom aplikację Streamlit:** Wykonaj to polecenie:

   ```
   streamlit run multi_sitemap_app.py
   ```

3. **Uzyskaj dostęp do aplikacji:** Twoja domyślna przeglądarka internetowa powinna automatycznie otworzyć nową kartę z aplikacją (zazwyczaj pod adresem `http://localhost:8501`). Jeśli nie, ręcznie przejdź pod ten adres w przeglądarce.

6. 🖱️ Korzystanie z Aplikacji

Interfejs aplikacji jest podzielony na pasek boczny (sidebar) do konfiguracji i główny obszar do wyświetlania wyników.

1. **Wprowadź domenę (Enter Domain):** W pasku bocznym wpisz docelową domenę (np. `streamlit.io`).
2. **Znajdź mapy strony (Find Sitemaps):** Kliknij przycisk "Find Sitemaps".
3. **Wybierz mapy strony (Select Sitemaps):** Zaznacz pola wyboru obok mapy(-) strony, które chcesz przeanalizować. Możesz użyć pola wyboru "Select All Sitemaps".
4. **Skonfiguruj filtry (Configure Filters - Opcjonalnie):** Rozwiń sekcję "URL Include/Exclude Filters", aby dodać słowa kluczowe, które muszą (lub nie mogą) być obecne w analizowanych adresach URL. Wybierz logikę (AND/OR) dla filtrów uwzględniających.
5. **Ustaw opcje analizy (Set Analysis Options):**
   - Zdecyduj, czy włączyć **"Analyze Page Content"**. Włączenie tej opcji jest wolniejsze, ale znacznie dokładniejsze.
   - Jeśli analizujesz treść, wybierz **"Vectorization Mode"** (Content Only, URL Path Only lub Combined). Dostosuj **"Content Weight"**, jeśli używasz trybu Combined.
   - Ustaw **"Max URLs to Process"** (uwaga: ten suwak może znajdować się poza sekcją "Advanced" w niektórych wersjach).
   - Dostosuj **"Advanced Analysis Options"** (TF-IDF, t-SNE, Metryki, Kanibalizacja) w razie potrzeby. Domyślne ustawienia są zazwyczaj odpowiednie, ale zapoznaj się z poniższym wyjaśnieniem dotyczącym parametrów Perplexity, k1 i k2.

   #### Zrozumienie Opcji Zaawansowanych Analizy (Advanced Analysis Options)

   Te ustawienia, znajdujące się w rozwijanej sekcji "Advanced Analysis Options", pozwalają na precyzyjne dostrojenie procesu analizy. Ich modyfikacja jest opcjonalna i zazwyczaj potrzebna tylko wtedy, gdy domyślne wyniki wydają się nieprawidłowe lub gdy chcesz poeksperymentować.

   - **t-SNE Perplexity**

     Co robi: Ten parametr wpływa na algorytm t-SNE używany do tworzenia wizualizacji "Visual Map". Koncepcyjnie odnosi się do liczby najbliższych sąsiadów branych pod uwagę dla każdego punktu podczas tworzenia mapy w niskiej wymiarowości.
     Wpływ: Niższe wartości podkreślają strukturę lokalną (potencjalnie pokazując więcej małych, zwartych klastrów). Wyższe wartości bardziej skupiają się na strukturze globalnej (potencjalnie łącząc mniejsze klastry lub pokazując szersze relacje).
     Wskazówki: Domyślna wartość (**15**) jest rozsądnym punktem wyjścia, zwłaszcza dla małych i średnich zbiorów danych. Dla bardzo małych zbiorów danych (< 50 URL), spróbuj wartości bliższych 5. Dla dużych zbiorów danych (> 1000 URL), możesz eksperymentować z wartościami do **50**. Jeśli mapa wygląda jak pojedyncza, gęsta kula, wartość perplexity może być zbyt wysoka. Jeśli wygląda zbyt fragmentarycznie bez wyraźnych grup, może być zbyt niska.

   - **Site Focus Score Scaling (k1)**

     Co robi: Ten suwak (domyślnie **5.0**) dostosowuje czułość obliczania **Site Focus Score**. Skaluje on, jak średnie podobieństwo między wektorami URL przekłada się na końcowy wynik 0-100.
     Wpływ: Wyższa wartość (`↑`) powoduje szybszy wzrost wyniku wraz ze wzrostem średniego podobieństwa. Oznacza to, że wynik staje się bardziej czuły i witryny potrzebują wyższego wewnętrznego podobieństwa, aby osiągnąć najwyższe wyniki. Niższa wartość (`↓`) powoduje wolniejszy wzrost wyniku.
     Wskazówki: Dostosuj tę wartość, jeśli wyniki dla różnych witryn wydają Ci się konsekwentnie zbyt wysokie (obniż `k1`) lub zbyt niskie (zwiększ `k1`) w stosunku do Twojej interpretacji. Domyślna wartość 5.0 zapewnia umiarkowaną czułość. Eksperymentuj w razie potrzeby, aby skalibrować zakres wyników.

   - **Site Radius Score Scaling (k2)**

     Co robi: Ten suwak (domyślnie **5.0**) dostosowuje czułość obliczania **Site Radius Score**, a konkretnie jak maksymalna odległość dowolnego wektora URL od centralnego tematu witryny (centroidu) wpływa na wynik poprzez formułę logarytmiczną.
     Wpływ: Wyższa wartość (`↑`) powoduje szybszy spadek wyniku wraz ze wzrostem maksymalnej odległości. To sprawia, że wynik jest bardziej czuły na elementy odstające lub rozproszenie treści (bardziej zwarta witryna tematyczna uzyskuje wyższy wynik). Niższa wartość (`↓`) powoduje wolniejszy spadek wyniku, co oznacza, że witryna musi być znacznie bardziej rozproszona, aby wynik znacząco spadł.
     Wskazówki: Dostosuj tę wartość, jeśli wyniki Radius Score wydają się szczególnie z intuicją. Jeśli pozorne stosunkowo zwarte witryny uzyskują niskie wyniki (rozważ lekkie zwiększenie `k2`, aby uczynić go bardziej czułym na małe odchylenia) lub jeśli zróżnicowane witryny uzyskują bardzo wysokie wyniki (rozważ lekkie zmniejszenie `k2`, aby uczynić go mniej czułym). Domyślna wartość 5.0 oferuje zrównoważoną czułość.

6. **Skonfiguruj Podsumowanie AI (Configure AI Summary - Opcjonalnie):** Włącz przełącznik ("toggle"), jeśli Twój klucz API jest skonfigurowany w pliku `.env`.
7. **Przetwórz wybrane mapy strony (Process Selected Sitemaps):** Kliknij przycisk "Process Selected Sitemaps", aby rozpocząć analizę. Aplikacja wyświetli aktualizacje statusu.
8. **Przeglądaj wyniki (Explore Results):** Po zakończeniu przetwarzania, przejrzyj wyniki, korzystając z zakładek (tabs):
   - 📊 **Overview:** Kluczowe metryki (**Site Focus Score** i **Site Radius Score**), podsumowanie AI, rozkład typów stron źródeł, kluczowe adresy URL.
   - 🔗 **URL Details:** Przeszukiwalna tabela wszystkich przetworzonych adresów URL i ich danych.
   - 🗺️ **Visual Map:** Interaktywny wykres t-SNE pokazujący relacje między adresami URL.
   - ⚠️ **Cannibalization:** Tabela potencjalnie zduplikowanych par treści na podstawie podobieństwa.
   - 🔍 **Content Inspector:** Wyświetla wyodrębniony tekst użyty do analizy (jeśli włączono analizę treści).
   - 📋 **Processing Log:** Szczegółowe logi z przebiegu analizy.

7. 🧩 Zrozumienie Komponentów (Główne Pliki)

Aplikacja opiera się na współpracy kilku plików:

- `multi_sitemap_app.py`
  Główny plik aplikacji, który uruchamiasz. Tworzy interfejs webowy (przy użyciu Streamlit), obsługuje interakcje użytkownika, wywołuje różne moduły w odpowiedniej kolejności i wyświetla końcowe wyniki.
- `.env`
  Prosty plik tekstowy (nie Python) do bezpiecznego przechowywania Twojego klucza API Google poza głównym kodem.
- `requirements.txt`
  Lista wszystkich zewnętrznych bibliotek Python, których aplikacja potrzebuje do działania.
- `modules/` (Katalog)
  Zawiera główną logikę, podzieloną dla lepszej organizacji:
- `sitemap_finder.py`
  Znajduje potencjalne adresy URL map strony dla danej witryny (sprawdzając `robots.txt` i popularne lokalizacje).
- `sitemap_parser.py`
  Pobiera i przetwarza zawartość plików map strony (XML, TXT, pliki indeksowe, skompresowane pliki), aby wyodrębnić listę adresów URL stron, stosując filtry uwzględniania/wykluczania.
- `content_extractor.py`
  (Wersja uproszczona) Jeśli analiza treści jest włączona, ten moduł pobiera kod HTML każdego adresu URL i próbuje wyodrębnić tylko główny tekst artykułu/treści, czyszcząc go na potrzeby analizy, używając: `trafilatura` (jeśli dostępna) lub podstawowych heurystyk `beautifulsoup`.
- `simple_vectorizer.py`
  (Wersja uproszczona) Przetwarza ścieżki URL i/lub oczyszczoną treść strony. Konwertuje dane tekstowe na wektory numeryczne TF-IDF (koncepcyjnie "page embeddings") przy użyciu `scikit-learn`. Obsługuje tryby: tylko URL, tylko treść oraz połączony.
- `dimensionality_reducer.py`
  Redukuje wysokowymiarowe wektory do 2 wymiarów przy użyciu t-SNE w celu wizualizacji. Oblicza również geometryczny centroid (koncepcyjnie "site embedding") wektorów i odległości od niego.
- `analyzer.py`
  Oblicza metryki wysokiego poziomu, takie jak **Site Focus Score** i **Site Radius Score** (używając zaktualizowanej skali logarytmicznej). Identyfikuje potencjalną kanibalizację treści, znajdując pary adresów URL o wysokim podobieństwie wektorowym.
- `llm_summarizer.py`
  Formatuje dane analityczne w prompt i używa API Google Gemini (poprzez bibliotekę `google-generativeai`) do wygenerowania podsumowania wyników w języku naturalnym, jeśli opcja jest włączona i skonfigurowana.
