Przetwarzanie języka naturalnego

Sprawozdanie z projektu

Część Teoretyczna

Indeksowanie stron internetowych przez Google

Czym jest indeksowanie stron internetowych?

Indeksowanie jest to proces dzięki któremu wyszukiwarki uporządkowywują informacje stron internetowych przed wykonaniem wyszukiwania w celu umożliwienia szybkich odpowiedzi na zadane zapytania. Wyszukując dane zapytanie w przeglądarce nie przeszukujemy sieci w czasie rzeczywistym w celu znalezienia interesujących nas informacji, a korzystamy z utworzonego systemu danych w postaci indeksów. Idee indeksowania stron internetowych można zobrazować na przykładzie biurowego systemu kartotek lub “biblioteki”.

A ile tych stron jest?

Na dzień 22 maja stron internetowych było 1 953 256 800, a ich liczba rośnie z każdą sekundą.

Poza stronami internetowymi Google posiada wiele indeksów, które obejmują informacje na temat:

- milionów książek z całego świata
- rozkładów jazdy lokalnych przewoźników
- danych ze źródeł publicznych jak np. Bank Światowy

Etapy działania wyszukiwarki Google

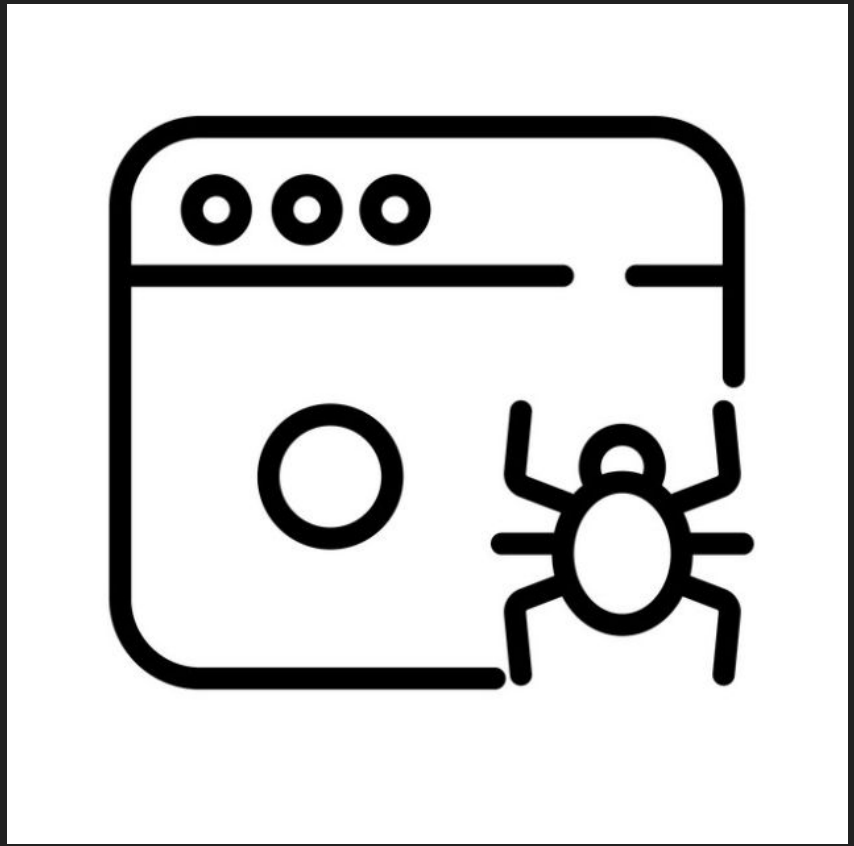
1) Przeszukiwanie (ang. *crawling*) - pobieranie tekstu, zdjęć, filmów ze znalezionych stron internetowych poprzez zautomatyzowane programy (ang. *crawlers*).
2) Indeksowanie (ang. *indexing*) - analizowanie tekstu, zdjęć, filmów ze znalezionych uprzednio stron internetowych, a następnie zachowanie ich w ogromnej bazie danych zwanej “Google index”.
3) Wyszukiwanie rezultatów - kiedy użytkownik korzysta z wyszukiwarki Google ta zwraca informację na podstawie zapytania użytkownika (wykorzystanie m.in. “Google index”).

Etap pierwszy - Skanowanie:

Pierwszym krokiem jest przeszukanie internetu pod kątem tego, jakie strony internetowe istnieją.

Proces ten wykonywany jest przez automatyczne programy zwane crawlerami. Przeszukują one znane strony internetowe pod kątem linków do nieznanych stron.

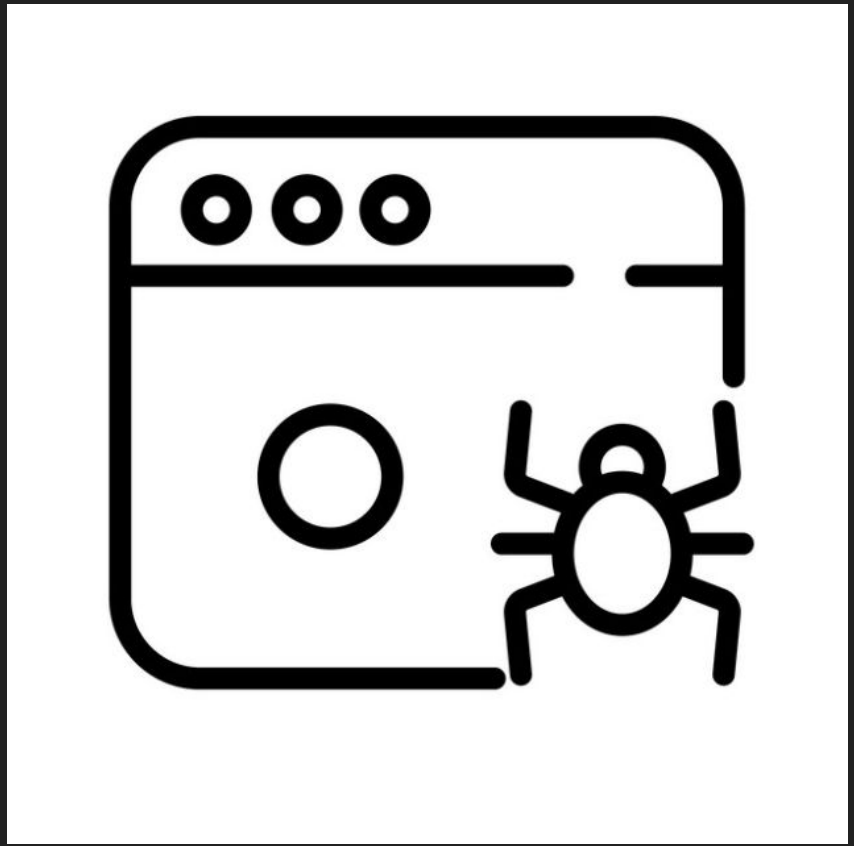
Inną opcją, jest możliwość przesłania do Google mapy naszej witryny, która to zawiera informacje o stronach, filmach i innych plikach w Twojej witrynie oraz o związkach między tymi elementami.



Etap pierwszy - Skanowanie:

Gdy Google znajdzie adres URL strony, może ją odwiedzić (lub ją „zindeksować”), aby dowiedzieć się, co ona zawiera.

Do skanowania miliardów stron internetowych Google używa ogromnej liczby komputerów a aplikacją odpowiedzialną za te czynności jest Googlebot (zwany też robotem, botem lub pająkiem). Googlebot działa według określonych algorytmów, aby wybierać witryny, które należy zindeksować, i ustalać, jak często mają być sprawdzane oraz ile stron z danej witryny pobrać.

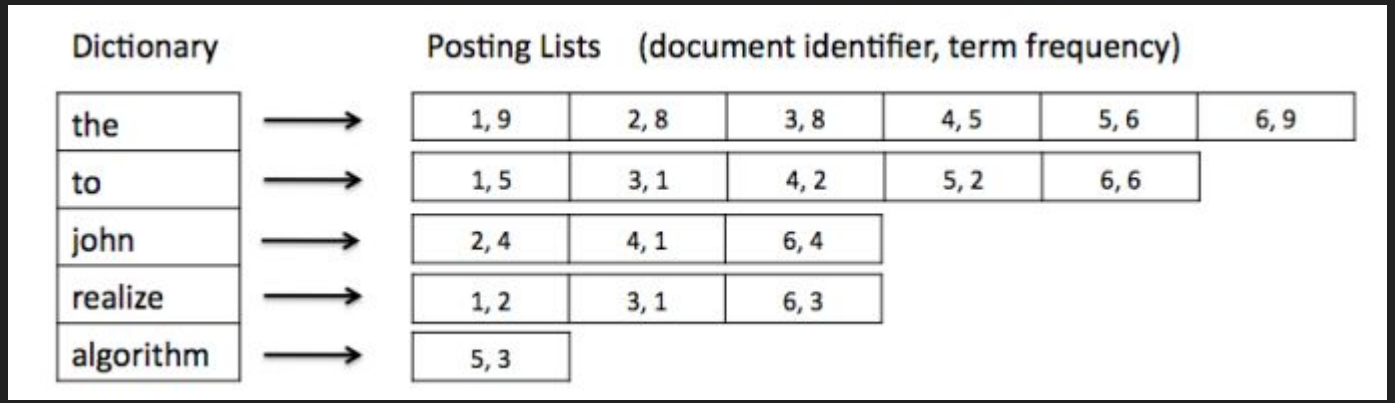


Etap drugi - Indeksowanie:

Po znalezieniu strony Google stara się rozpoznać jej tematykę. Ten etap nazywa się indeksowaniem i obejmuje przetwarzanie oraz analizę treści tekstowych, kluczowych tagów i atrybutów treści, takich jak elementy <title> i atrybuty alt, obrazy, filmy oraz innych treści.

Etap drugi - Indeksowanie: struktura indeksów

Najczęściej pojawiającym się przykładem struktury indeksów jest “inverted index”.



Etap drugi - Indeksowanie: zarządzanie zasobami strony

Podejście pasywne

Udostępnienie strony bez mapy witryn

Zalety:
Należy jedynie przygotować treść strony. Przydatne przy prostych witrynach, gdzie twórcy nie zależy na szybkim uwzględnieniu strony w wynikach wyszukiwania.

Wady:
System może nie znaleźć całej zawartości strony szczególnie jeśli jest ona nowa. Mogą wystąpić problemy z nowymi treściami, które się szybko zobaczyć w wyszukiwarce.

Podejście aktywne

Udostępnienie systemom bezpośredniej listy URL strony (mapę witryn)

Zalety:
Poprawia skuteczność wyświetlania treści w wynikach wyszukiwarki z elementami rozszerzonymi.

Wady:
Wymaga nakładu pracy. Należy udostępnić metadane zasobów, czyli mapę witryn oraz powiązania pomiędzy stronami internetowymi, aplikacjami i stronami AMP.

Etap drugi - Indeksowanie: czynniki wpływające na jakość indeksowania

- Witryna powinna być bogata w przydatne treści, a strony powinny być przejrzyste i dokładne.
- Witryna powinna zawierać słowa, które użytkownicy potencjalnie by użyli, żeby znaleźć daną stronę.
- Elementy <title> oraz atrybuty alt powinny być opisowe, konkretne i dokładne.
- Witryna powinna posiadać przejrzystą hierarchię budowy.
- Witryna powinna stosować zalecenia Google dotyczące obrazów, zdjęć czy innego rodzaju danych.
- W celu ułatwienia dokładnego zrozumienia zawartości strony internetowej witryna powinna pozwolić na zindeksowanie wszystkich jej zasobów, które wpływają na jej renderowanie (np. pliki CSS i JavaScript).
- Wszystkie ważne treści powinny być domyślnie widoczne w witrynie. Ukryte elementy HTML również są indeksowane, ale jako, że użytkownik ich nie widzi są traktowane z mniejszym prioretytem.

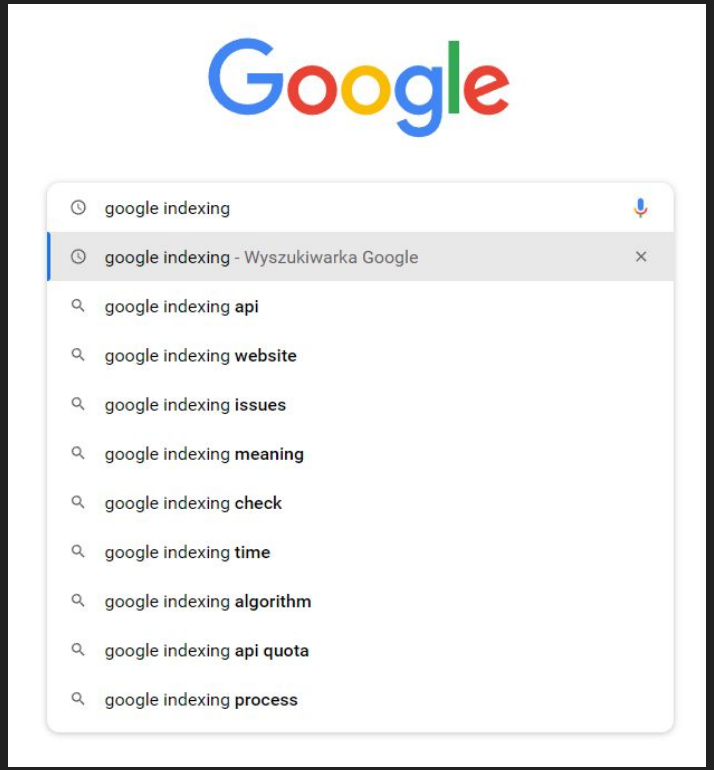
Wiele innych czynników wpływa na jakość indeksowania, w tym celu należy dokładnie przebadać dokumentację Google, która instruuje jak tworzyć witryny, aby zostały dobrze zindeksowane.

Etap trzeci - Wyświetlanie wyników wyszukiwania

Po wpisaniu przez użytkownika zapytania systemy Google’a wyszukują w indeksie pasujące strony, a następnie zwracają te najlepsze i najtrafniejsze w ich ocenie.

Na wyniki wyszukiwania ma wpływ 5 głównych czynników:

- Znaczenie
- Trafność
- Jakość
- Użyteczność
- Kontekst



Znaczenie - zrozumienie co kryje się da danych pytaniem. Pomagają w tym modele językowe, które poprawiają literówki, rozumieją słowa i ich intencje np. “zmiana jasności monitora”, a “dostosowanie jasności monitora”, rozumieją jakich rodzajów informacji szuka użytkownik np. “gotowanie zdjęcie” pokaże zdjęcia/grafiki przepisów.

Trafność - analiza treści strony w celu określenia, czy jest ona interesująca dla użytkownika np. obecność na stronie internetowej tych samych haseł co w wyszukiwanym haśle.

Jakość - wybieranie fachowych, rzetelnych i wiarygodnych treści np. obecność linków czy odniesień do treści na innych popularnych stronach.

Użyteczność - sprawdzenie jakości/użyteczności strony np. czy jest dostosowana do komórek oraz jak szybko strona się wczytuje.

Kontekst - wyszukiwanie informacji pod kątem ustawień wyszukiwarki np. język, lokalizacja oraz ostatnich wyszukiwań. Będąc w Krakowie dostaniemy inne wyniki wpisując “naprawa laptopa” niż miałoby to miejsce w Berlinie.

Algorytm PageRank

Najbardziej znanym a także pierwszym algorytmem określających ważność strony internetowej w wyszukiwarce Google był PageRank, którego nazwa nawiązuje do nazwiska założyciela tej firmy - Larry'ego Page'a.

PageRank jest rozwinięciem znanej od dawna heurystyki, wedle której jakość tekstu jest proporcjonalna do liczby tekstów na niego się powołujących. Ulepszenie zaproponowane przez autorów Google polegało na ważeniu jakości odnośników wskazujących na rozpatrywany tekst ich własną wartością PageRank

Szczegóły właściwego algorytmu nigdy nie zostały upublicznione i są jednymi ze ściśle strzeżonych tajemnic Google. Ponadto algorytm ten był nieustannie poprawiany od samego powstania w 1998 roku.

Od 2016 roku niemożliwym jest pobranie wartości PageRank dla danej strony, mimo tego, wewnątrz wyszukiwarki Google algorytm ten pozostaje w użyciu.

Dziękujemy za uwagę

Radosław Bujak
Michał Borowski
Michał Gątkowski

Bibliografia

1. https://computersciencewiki.org/index.php/Web-indexing
2. https://developers.google.com/search/docs/advanced/guidelines/how-search-works
3. https://support.google.com/webmasters/answer/7645831
4. https://developers.google.com/search/docs/advanced/structured-data/intro-structured-data
5. https://developers.google.com/search/docs/advanced/guidelines/webmaster-guidelines
6. https://developers.google.com/search/docs/beginner/intro-indexing
7. https://www.google.com/search/howsearchworks/how-search-works/ranking-results
8. https://www.deepcrawl.com/knowledge/technical-seo-library/search-engine-indexing/
9. https://www.internetlivestats.com/
10. https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/37043.pdf