# Indeksowanie stron internetowych przez Google

...

Przetwarzanie Języka Naturalnego

Staniszewski Miłosz, Szołdra Szymon

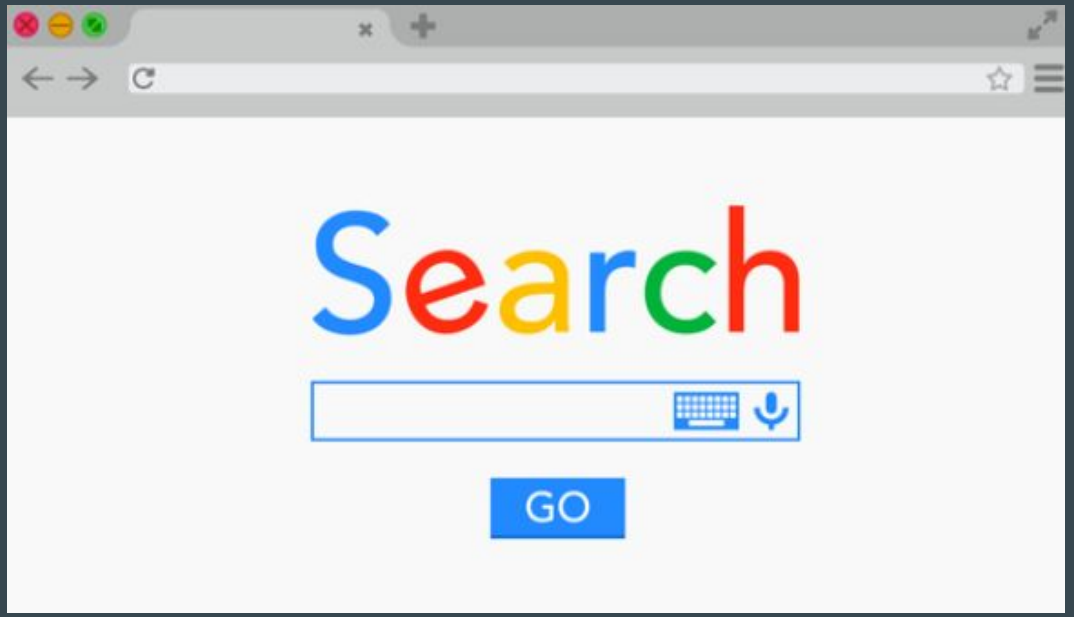
# Wprowadzenie

W 2016 Google podało że ma zaindeksowane ponad 130 bilionów stron, a liczba ta ciągle rośnie z każdym dniem. Dodatkowo, każdego dnia około 15% wyszukiwań to nowe dla wyszukiwarki hasła, dlatego aby dostarczać użytkownikom satysfakcjonujące treści, potrzebny jest sprawny, automatyczny system.

Jak działa proces wyszukiwania informacji w wyszukiwarce Google?

# 3 główne etapy systemu wyszukiwania w Google

1. Crawling
2. Indexing
3. Ranking



# Crawling

# Crawling

Jest to analiza sieci w postaci przeszukiwania jej przez boty. Krok ten jest potrzebny aby wyszukiwarka wiedziała, jakie strony istnieją w internecie. Ponieważ nie istnieje nic takiego jak “rejestr stron internetowych” czy cokolwiek podobnego, systemy wyszukiwarek internetowych muszą nieustannie aktualizować swoją listę znanych stron.

**Cel**: Zdobycie informacji o tym, jakie strony istnieją w internecie.

**Sposób działania**: Boty (tzw. pająki) “pełzają” przez World Wide Web w celu odnalezienia nowych stron i oszacowania, które powinny zostać zindeksowane. Proces ten odbywa się przez wchodzenie w kolejne linki znalezione w drzewie DOM oraz plikach javascript kolejnych stron.

# Crawling - Googlebot

Proces “pełzania” w wyszukiwarce Google jest realizowany przez Googlebot.

1. Przy odwiedzeniu strony, w pierwszej kolejności analizowana jest możliwość indeksacji przez wyszukiwarkę. Jeżeli nie ma przeciwskazań do zaindeksowania strony przez Google (jak np. blokujący tag *noindex* lub dyrektywy w pliku *robots.txt* uniemożliwiające wyszukiwarce odwiedzenie witryny), Googlebot przechodzi do faktycznego procesu skanowania strony .

# Crawling - Googlebot

2. Googlebot zaczyna od jej wyrenderowania na silniku Chrome. W dzisiejszym internecie jest to ważny krok, ponieważ strony znaczną część swojej treści pobierają za pomocą skryptów JavaScriptowych.
3. Robot podczas skanowania strony działa wobec określonych algorytmów. Ustalają one, które strony indeksować, jak często je odwiedzać aby zaktualizować swoją listę oraz ile stron z danej witryny pobrać. Roboty są tak zaprojektowane, aby nie przeciążać odwiedzanej witryny - przykładowo, zbyt duża częstotliwość błędów 500 HTTP dla bota oznacza, że trzeba spowolnić jego działanie.

# Client i Server side rendering, a Googlebot

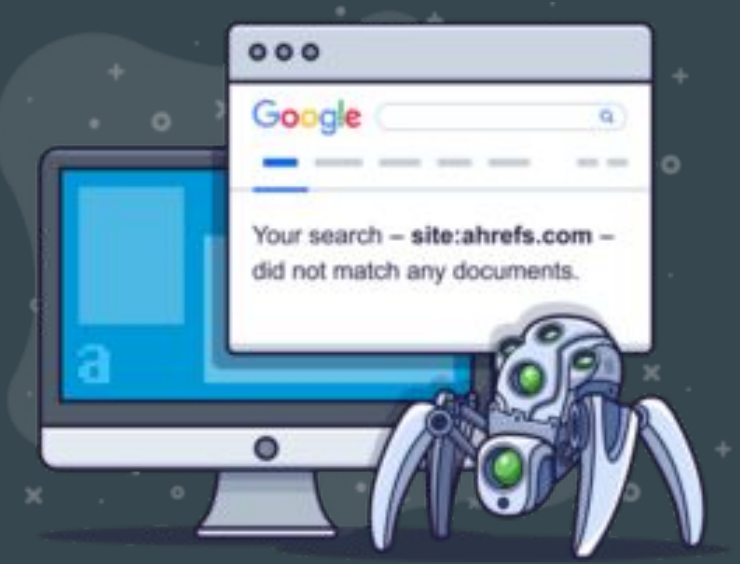
Jeszcze na początku poprzedniej dekady zdecydowana większość stron internetowych była renderowana na serwerze. W połowie dekady popularność zdobyły rozwiązania Single Page Application takie jak chociażby React. Jednak z powodu problemów z Googlebotem programiści powracają do renderowania stron internetowych na serwerze - (rozwiązania Next.js, Nuxt.js). Przekłada się to na zdecydowaną poprawę pozycjonowania stron napisanych w ten sposób w wynikach wyszukiwania Google.

# Googlebot a semantyka w HTML

Tworząc stronę internetową warto pamiętać o zachowaniu zasad semantyki HTML. Oczywiście możemy zbudować stronę przy pomocy samych znaczników , jednak będzie to miało negatywny skutek w możliwości poprawnego odczytania struktury strony przez Googlebota.

Google bardzo lubi poprawne używanie znaczników HTML, zachowywanie struktury nagłówków itd.

# Indexing


Google
Your search – site:ahrefs.com – did not match any documents.

# Indeksowanie

Po procesie crawlingu, przeskanowana strona zostaje dodana do indeksu googla - kolekcji stron internetowych.

Każdy taki wpis do bazy danych dotyczący pojedynczej strony zawiera dodatkowe informacje zebrane przez crawlera.

# Szczegółowe dane o stronach

Te szczegółowe informacje, to między innymi:

- “Mapa” wszystkich pozostałych miejsc w internecie, do których strona prowadzi za pomocą linków.
- Informacje dotyczące tych linków - przykładowo, to gdzie są umiejscowione na stronie (dlatego istotne jest wspomniane wcześniej poprawne używanie znaczników HTML), czy link jest reklamą lub nie itp.
- Dokładne dane dotyczące charakteru treści strony oraz trafności względem jej związku z poruszanym tematem.

# Co dalej?

Dane z tej kolekcji to dokładnie te dane, które zobaczymy po wyszukaniu jakiegokolwiek hasła w wyszukiwarce. Jednak aby wyświetlić nam najlepsze wyniki, Google dodatkowo filtruje i sortuje strony na podstawie swoich kryteriów oceniania. To właśnie w tym celu Googlebot pobiera takie dane jak trafność strony względem jej związku z poruszanym tematem - dzięki temu strony o podobnym charakterze da się ze sobą porównać.

# Ranking

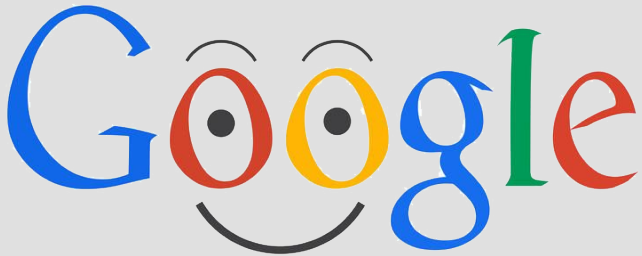
# Od czego zależy kolejność wyników w Google?

1. Historia wyszukiwania i osobiste preferencje. W zależności od tego czy Google uzna że interesujemy się motoryzacją czy zwierzętami wynik wyszukiwania frazy “jaguar” będzie zdecydowanie inny.
2. Geolokalizacja. Jeśli wpiszemy frazę “fryzjer” lista wyników będzie dostosowana do miejsca, w którym obecnie się znajdujemy.
3. Obecna popularność frazy, liczba wyszukiwań przez innych użytkowników
4. Liczba odnośników do danej witryny z innych witryn.

Tak naprawdę większości tych rzeczy możemy się jedynie domyślać. Nie ma żadnej osoby na świecie, która zna dokładny algorytm pozycjonowania stron. Eksperci SEO metodą prób i błędów dochodzą do najbardziej skutecznych metod wybijania się w rankingu.

Wymieniliśmy kilka z nich, lecz z pewnością wpływ na pozycjonowanie mają setki zmiennych.

# Dziękujemy za uwagę



# Użyte źródła:


https://blogs.perficient.com/2016/12/21/how-googles-search-results-work-crawling-indexing-and-ranking/#:~:text=In%20a%20nutshell%2C%20this%20process,should%20rank%20for%20relevant%20queries

https://developers.google.com/search/docs/fundamentals/how-search-works#:~:text=Googlebot%20uses%20an%20algorithmic%20process,fast%20to%20avoid%20overloading%20it.

https://www.eactive.pl/blog-o-seo/crawling-indeksowanie-strony-ranking-stron-podstawy-pozycjonowania-w-wyszukiwarce/

https://widoczni.com/blog/indeksowanie-crawling-ranking/#:~:text=Crawling%20%2D%20to%20analiza%20stron%20internetowych,(lub%20zg%C5%82oszonego)%20adresu%20URL.
https://morningscore.io/how-does-google-rank-websites/#:~:text=To%20rank%20websites%2C%20Google%20uses,on%20a%20search%20result%20page.

https://developers.google.com/search/docs/crawling-indexing/robots/intro
https://www.artefakt.pl/blog/seo/dlaczego-widze-inne-strony-w-google-niz-znajomi-7-czynnikow-od-ktorych-zalezy-kolejnosc

https://pl.linkedin.com/pulse/od-czego-zale%C5%BCy-kolejno%C5%9B%C4%87-w-wynikach-wyszukiwania-google-cherubin