

Składowanie danych w systemach Big Data

Konrad Komisarczyk, Mikołaj Malec, Patryk Wrona



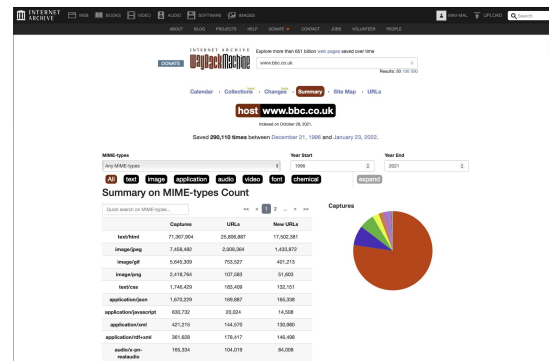


Cel projektu

Celem projektu była analiza zmian stron internetowych w czasie na podstawie danych z Wayback Machine.

Inspiracją do powstania projektu była udostępniana przez Wayback Machine funkcjonalność "Summary".

W celu prezentacji wyników działania systemu przeanalizowano 1500 najpopularniejszych stron internetowych według rankingu Tranco w okresie ostatnich 20 lat (od 2002 do 2022 roku).



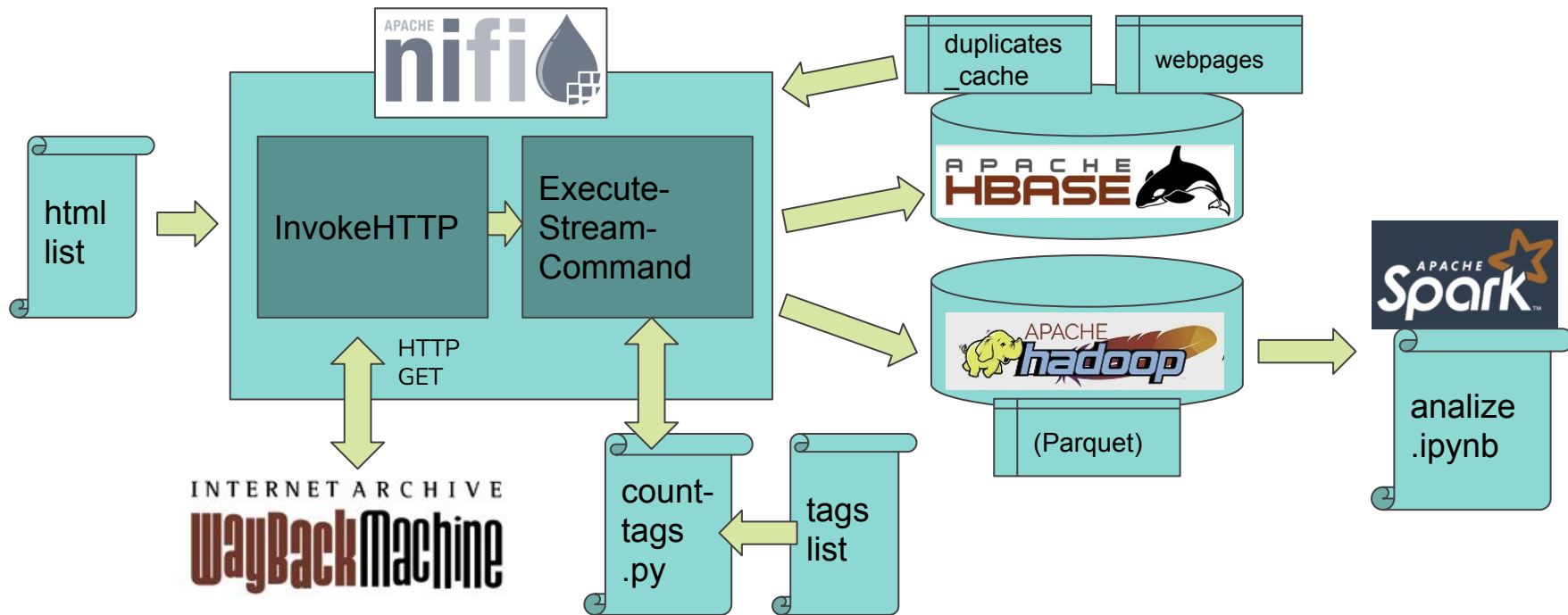
Tranco


A Research-Oriented Top Sites Ranking Hardened Against Manipulation

By [Victor Le Pochat](#), [Tom Van Goethem](#), [Samaneh Tajalizadehkhoob](#), [Maciej Korczyński](#) and [Wouter Joosen](#)


[Download the latest Tranco list](#)

Architektura rozwiązania



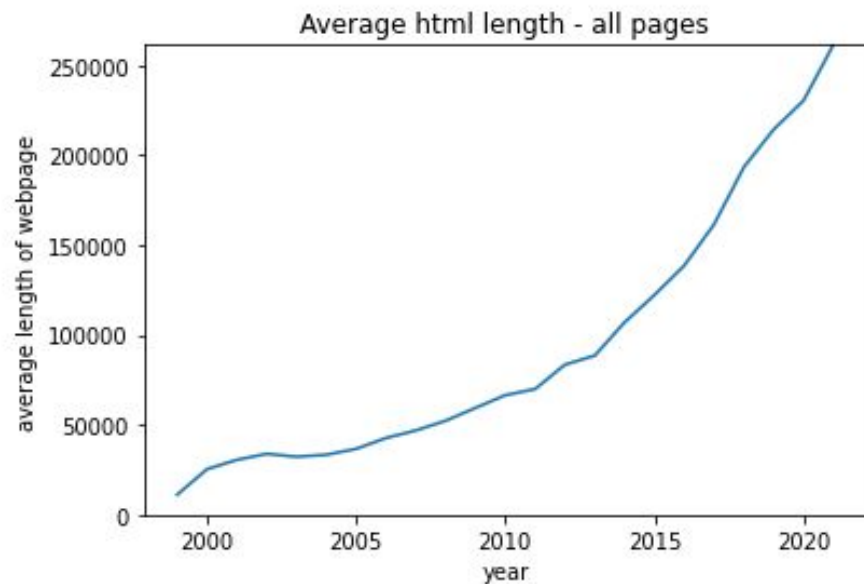


Analiza tagów 1500 najbardziej popularnych stron internetowych na przestrzeni 20 lat

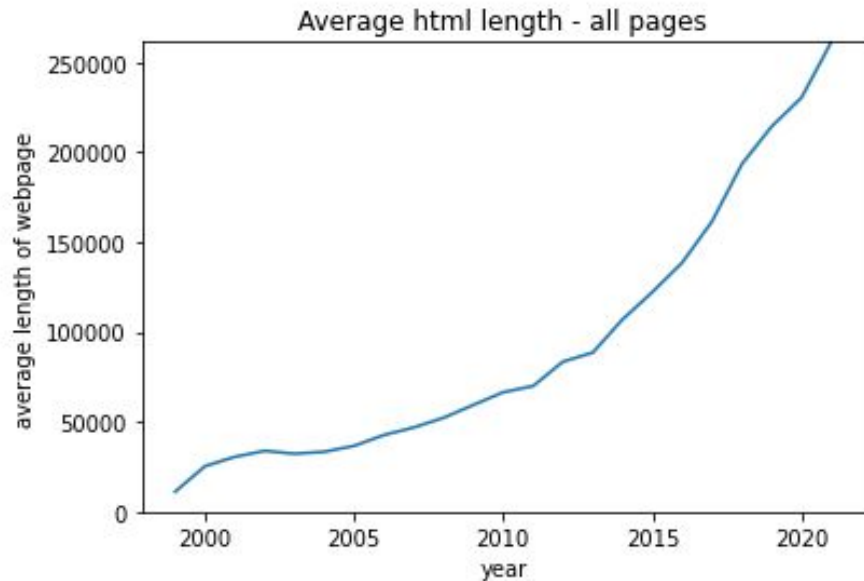




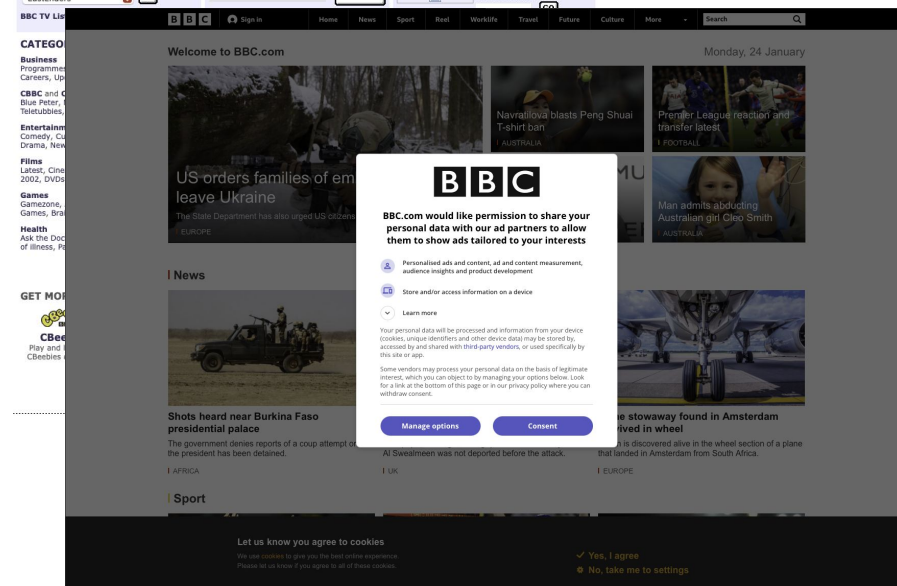
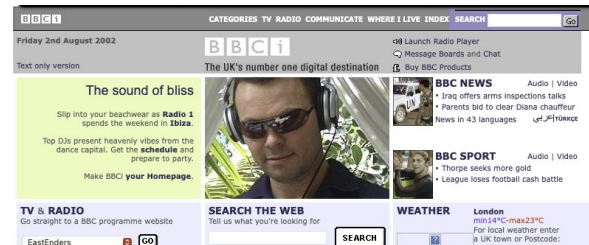
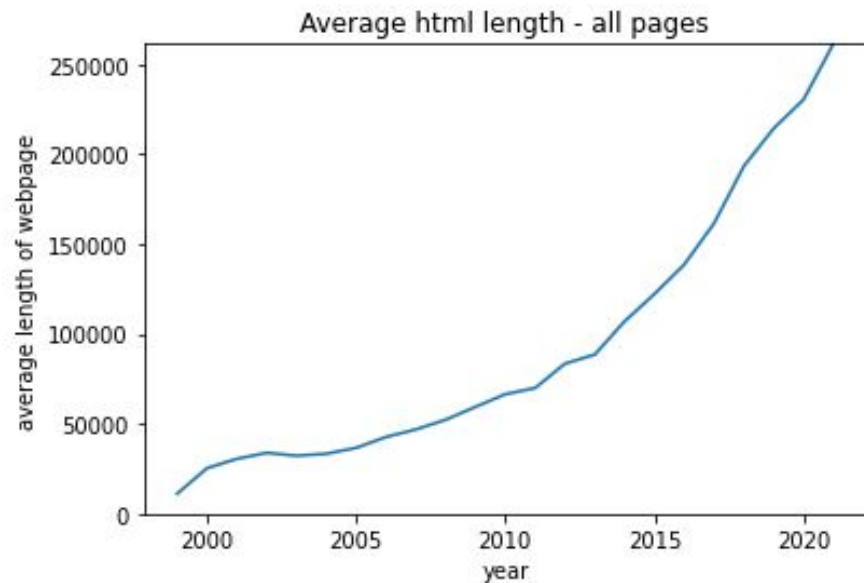
Analiza: coraz dłuższe strony



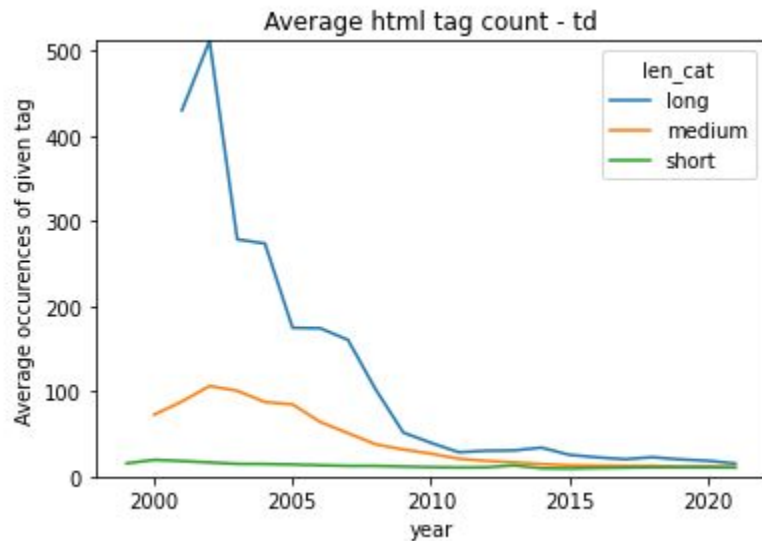
Analiza: coraz dłuższe strony



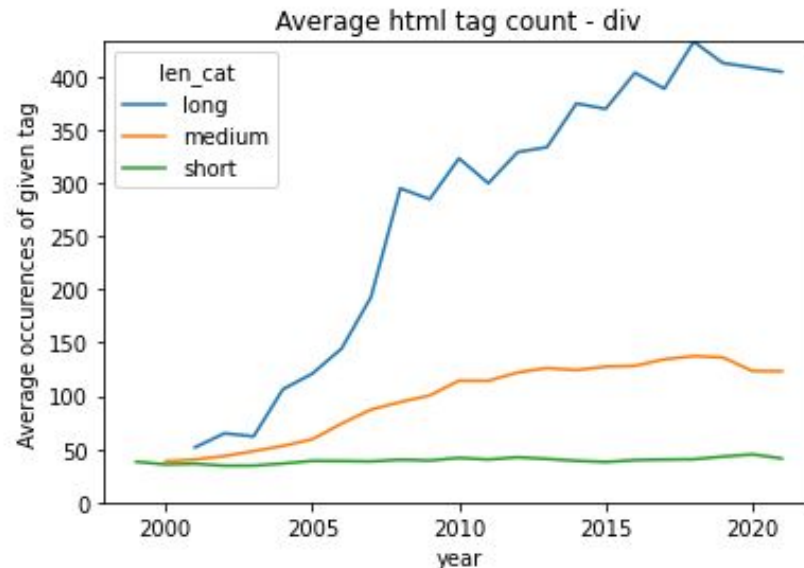
Analiza: coraz dłuższe strony



Analiza: `<td>` vs `<div>`



`<td>` - definiuje standardową komórkę w tabeli HTML



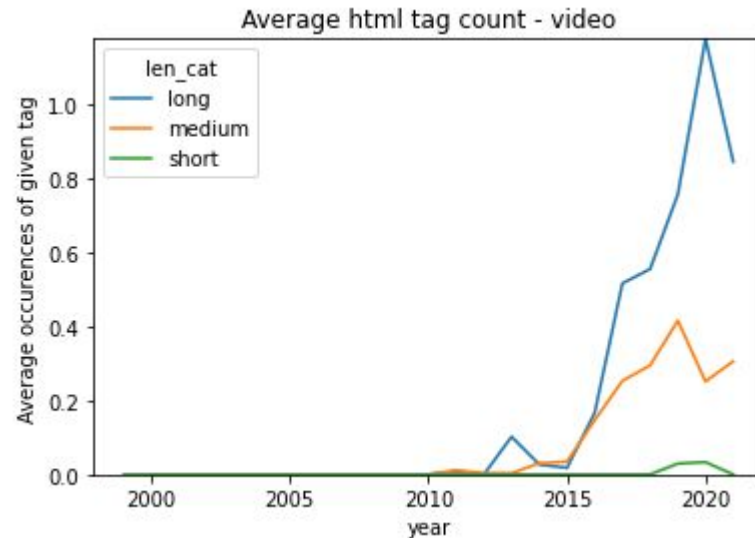
`<div>` - nie ma żadnego specjalnego znaczenia

Tables vs. DIV/CSS

The screenshot shows the BBC website from Friday 2nd August 2002. The layout is a dense grid of content. At the top, there's a navigation bar with 'BBC1' and 'CATEGORIES TV RADIO COMMUNICATE WHERE I LIVE INDEX SEARCH'. Below this, the main content area is divided into several sections: 'The sound of bliss' with a photo of a DJ, 'BBC NEWS' with a list of headlines, 'BBC SPORT' with a headline about the Thorne jocks, 'TV & RADIO' with a search bar, 'SEARCH THE WEB', 'WEATHER' for London, 'DON'T MISS' with a contest, 'WHERE I LIVE' with local information, 'What's On near you', and 'BBC World Service'. At the bottom, there's a 'GET MORE FROM THE BBC' section with links to 'Cbeebies', 'Annual Report', 'Learning', and 'WebGuide'. The footer contains 'About the BBC | Jobs at the BBC | myBBC | Help | Feedback'.

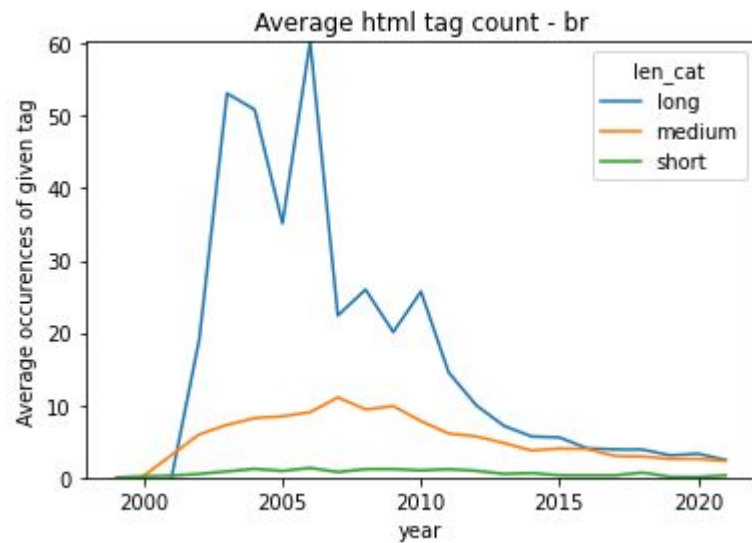
The screenshot shows the BBC website from Monday 24 January. The layout is more modern and clean than the 2002 version. A large cookie consent pop-up is centered on the screen, asking for permission to share personal data with ad partners. The background shows a grid of news stories, including 'US orders families of emigrants to leave Ukraine', 'Naval forces blasts Peng Shuai T-shirt ban', and 'Premier League reaction and transfer latest'. The footer includes 'Let us know you agree to cookies' and 'Yes, I agree' / 'No, take me to settings' buttons.

Analiza: <video>



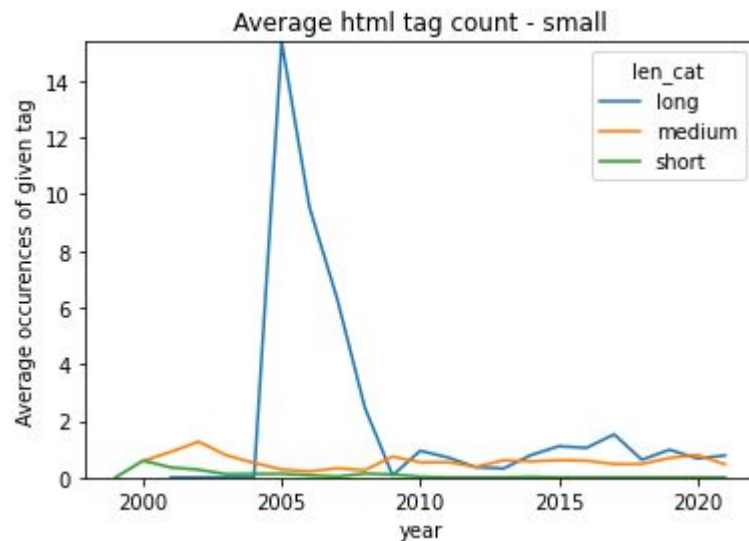
<video> - służy do osadzania treści wideo w dokumencie, na przykład klipu filmowego lub innych strumieni wideo

Analiza: `
`



`
` - tworzy w tekście
przełamanie linii

Analiza: <small>



Kiedyś:

<small> - tekst normalnego rozmiaru jest wyświetlany jako drobny druczek (ang. small print)

Laurence Tureaud <small>(Mr.T)</small>

Laurence Tureaud (Mr.T)

Teraz:

<small> - zmniejsza rozmiar czcionki tekstu

**Dziękujemy
za uwagę**





Bibliografia zdjęć

<https://archive.org/web/>

<https://nifi.apache.org>

<https://hbase.apache.org>

<https://hadoop.apache.org>

<https://spark.apache.org/docs/latest/api/python/index.html#>

<https://tranco-list.eu/#download>

<https://www.bbc.com>