Google PageRank Unplugged

Dit werk is gelicenseerd onder een Creative Commons Naamsvermelding-NietCommercieel-GelijkDelen 4.0 Internationaal licentie. Bezoek https://creativecommons.org/licenses/by-nc-sa/4.0/deed.nl om een kopie te zien van de licentie.

Google PageRank Unplugged

Samenvatting

Hoe bepaalt Google welke pagina voorrang krijgt in een zoekactie? Aan de hand van deze unplugged activiteit ontdekken de leerlingen de indexerings- en PageRank-algoritmen van Google.

Doel

De leerlingen leggen met eigen woorden uit hoe zoekmachines zoektermen indexeren en hoe Google te werk gaat om webpagina's voorrang te geven op andere webpagina's.

Duur

100 minuten (2 lesmomenten van 50 minuten).

Leeftijdsgroep

12-14j

Activiteit

Vraag

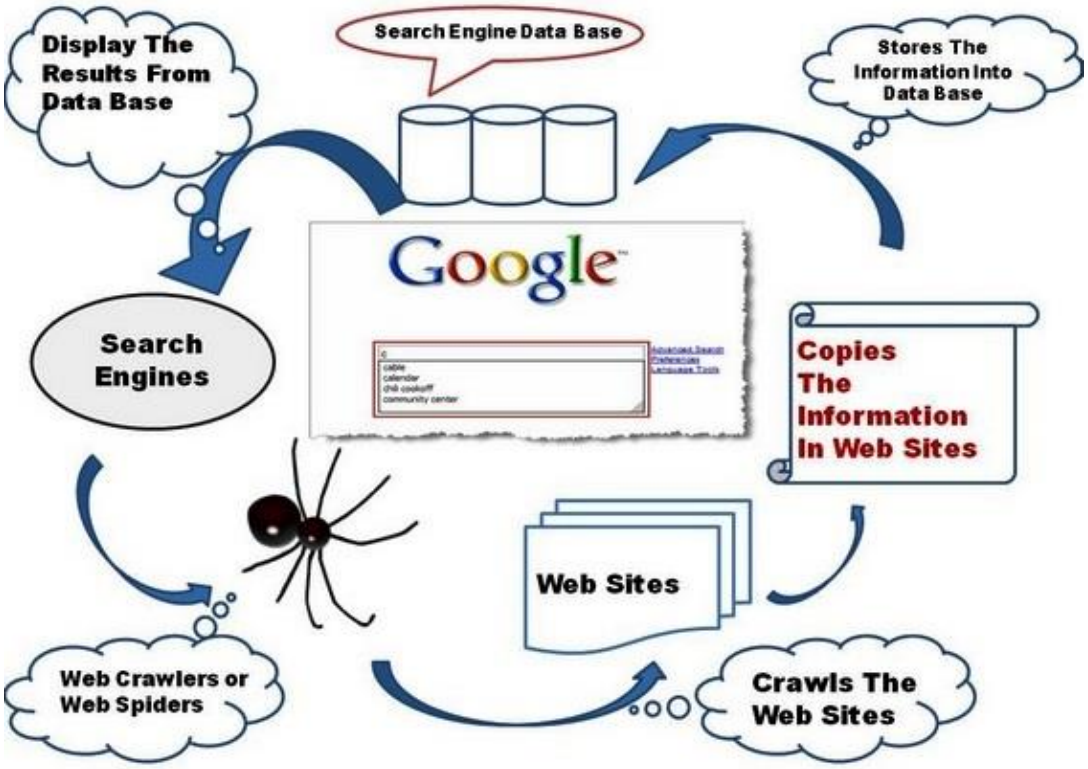
Wanneer jullie thuis of op school informatie moeten zoeken over een bepaald onderwerp; hoe gaan jullie dan te werk? (Duid enkele leerlingen aan en luister naar hun antwoord; mogelijk krijg je steeds het zelfde antwoord: we gaan op het Internet en zoeken via Google, Bing, … naar de term)

Wat krijg je als resultaat? (een opsomming van pagina's die je zoekterm bevatten, met het beste resultaat bovenaan).

Deel 1

Tijdens deze activiteit leren de leerlingen hoe de indexering van Google verloopt. Dit duurt ongeveer 50 minuten.

Leg kort uit dat Google een index bijhoudt: programmaatjes die web-crawlers heten, gaan bijhouden waar een woord zich bevindt in een pagina.



Figuur 1

Stap 1: Toon 3 tekstjes en laat de leerlingen een index opstellen

Opdracht: schrijf elk woord op en noteer daarachter op welke pagina je het terugvindt

Dit is een heel eenvoudig recept voor spaghetti	Dit recept voor macaroni is heel eenvoudig	Dit recept voor spaghetti is even eenvoudig als dat voor macaroni

Gebruik ***werkblad 1***.

Nu gaat het over 3 pagina's… maar wat als het over een miljard pagina's gaat? Hoe vindt Google ons recept het snelste? En het lekkerste? En wat als je een zin wilt zoeken? Hoe kent de zoekmachine de volgorde van de woorden?

Stap 2: voeg ook de locatie toe op de pagina

Zoekmachines houden naast de pagina ook de plaats van een woord op de pagina bij; heel eenvoudig. Het derde woord krijgt nummer 3, het negenentwintigste krijgt nummer 29.

Opdracht: schrijf voor elk woord de pagina en de plaats binnen de pagina op.

Gebruik ***werkblad 2***.

Stap 3: metawords

Tot nu toe hebben we eenvoudige voorbeeldjes gebruikt om het zoekgedrag van een zoekmachine te verklaren. Misschien weten jullie dit niet, maar de meeste webpagina's hebben redelijk wat structuur; ze bevatten titels, een hoofding, links naar andere pagina's, afbeeldingen, …

Om de oefening eenvoudig te houden, stellen we dat een webpagina een titel en een body heeft, net zoals in de tabel hieronder.

<titelStart>Recept spaghetti<titelStop> <bodyStart>Dit is een heel eenvoudig recept voor spaghetti <bodyStop>	<titelStart>Recept macaroni<titelStop> <bodyStart>Dit recept voor macaroni is heel eenvoudig<bodyStop>	<titelStart>Recepten<titelStop> <bodyStart> Dit recept voor spaghetti is even eenvoudig als dat voor macaroni<bodyStop>

Webpagina's worden geschreven in HTML. HTML gebruikt metawords om afbeeldingen, alinea's, titels, links, … betekenis te geven. In ons voorbeeld duiden we de start van de titel aan met <titelStart> en het einde met <titelStop>.

Webbrowsers tonen de metawords (in werkelijkheid noemen we deze "tags") gelukkig niet aan de gebruiker, maar de metawords zijn er wel! Klik maar eens met je rechtermuisknop op een webpagina en selecteer "paginabron weergeven". Merk op dat de tags <title> en </title> rond de titel van de webpagina staan.

Opdracht: neem de plaats van de metawords mee op in je zoekindex.

*Gebruik **werkblad 3**.*

Eindresultaat: geïndexeerde pagina's

De leerlingen hebben nu 3 pagina's geïndexeerd. Doe alsof je zoekt naar een recept voor macaroni. Wat is het adres van de pagina waar je de informatie zult vinden? Het is mogelijk dat enkele leerlingen toch een ander antwoord geven. Hoe komt het dan dat de ene pagina "hoger staat" dan een andere?

Start een kort gesprek over het aantal webpagina's op het Internet. Haal aan dat er zeer veel wordt geautomatiseerd dmv crawlers. Laat de lln eens berekenen hoe lang het zou duren om 4 miljard pagina's te indexeren aan de hand van de tijd het hun heeft gekost om 1 pagina te indexeren.

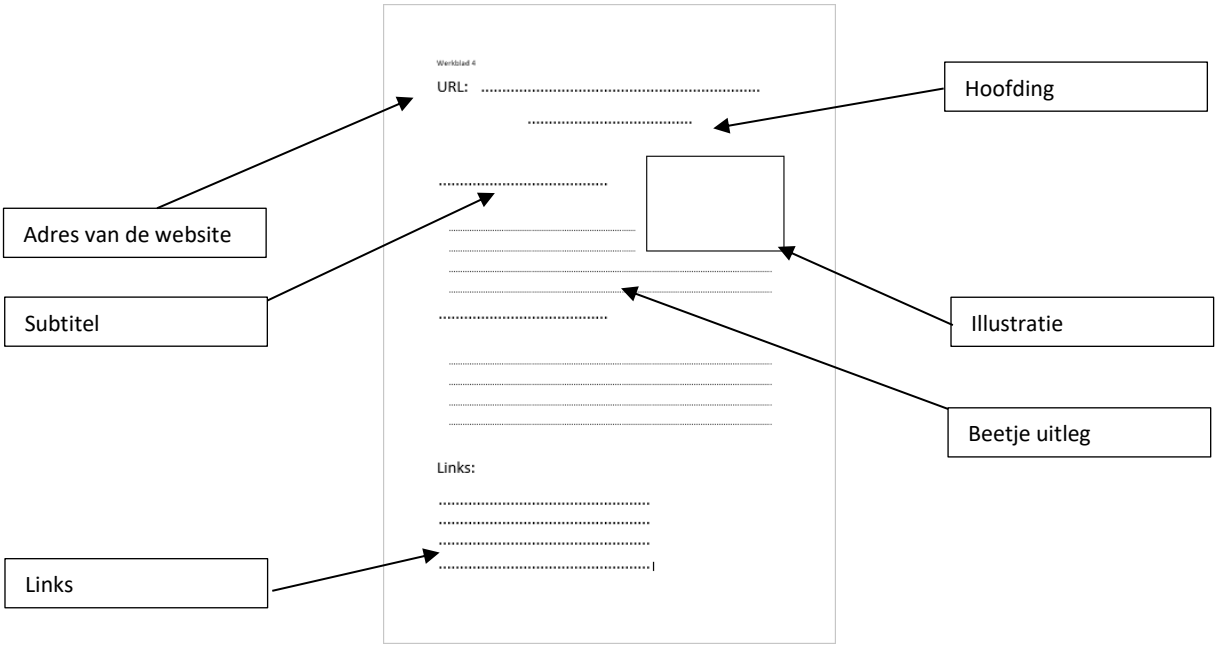
Deel 2 - Welke pagina's zijn "beter" dan andere?

Stap 1: Webpagina's maken

Leerlingen maken in kleine groepjes (max! 3) een webpagina op papier. Kies een onderwerp uit het thema waarrond je leerlingen op dit moment werken. De webpagina bevat zeker de volgende elementen:

- URL
- Hoofding (<heading>)
- Afbeelding (<img>)
- Subtitel(s)
- Enkele alinea's
- Sterretjes (vind ik leuk)
- Links naar andere pagina's.

Maak de webpagina's op **werkblad 4**.



Praat met de lln. Heeft er een groepje een koppeling (link) gelegd met een andere pagina? Wat gebeurt er als je op een link klikt? Vraag je lln om te kiezen met welke pagina ze willen linken.

Stap 2: Ranken

Doe alsof je een zoekactie uitvoert naar 1 van de pagina's van de lln.

De groepjes mogen hun pagina's een score (rang) geven adhv de volgende criteria:

- 1 punt per keer dat het sleutelwoord in een alinea staat
- 1 punt per keer dat de term in een hoofding staat
- 1 punt per keer dat de term in de paginatitel staat
- 1 punt per keer dat de term in de url staat
- 1 punt per keer dat er naar de pagina wordt gelinkt

Gebruik **werkblad 5**

Uitbreiding: Stap 3: SEO

Laat de leerlingen eens nadenken over hoe ze hun ranking zouden kunnen verbeteren (eerst wat meer nadenken over inhoud, daarna proberen te triggeren dat leerlingen meer links moeten laten leggen naar hun pagina).

Doe eens opnieuw

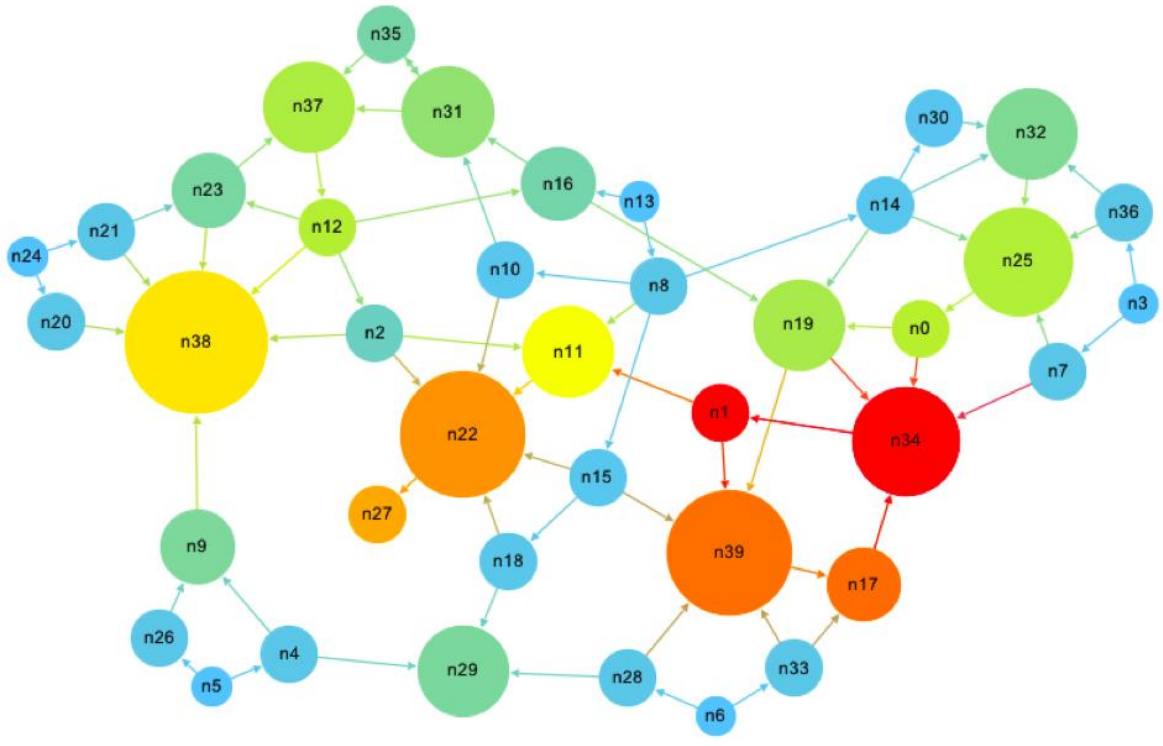
Concepten

Zoekmachines zijn gebouwd opdat gebruikers informatie kunnen vinden op het wereldwijde web (www). Deze machines maken gebruik van web crawlers om het hele www te indexeren. Deze indexen worden gebruikt om te bepalen welke pagina een bepaalde term of zoekterm bevat.

Om te weten of een pagina goed genoeg is om te verschijnen als antwoord op je vraag, analyseert het zoekprogramma de inhoud van elke pagina. Deze analyse is complex en wijzigt voortdurend. Typisch zoekt het programma hoe vaak een term voorkomt, hoe vaak synoniemen van de term voorkomen op een bepaalde pagina, of de term(en) voorkomen in de titel van de webpagina, … Daarnaast wordt gekeken of er links zijn met andere pagina's.

Google noemt dit algoritme PageRank (genoemd naar de stichter van Google, Larry Page). Als een pagina veel genoemd wordt (gelinkt wordt) op andere pagina's (die ook van goede kwaliteit zijn), dan zal Google sneller geloven dat die eerste pagina erg belangrijk is.

De pagina's in het wereldwijde web worden beschouwd als knopen in een gerichte graaf. Elke hyperlink van een pagina X naar pagina A is dan een boog in deze graaf van knoop X naar knoop A en kan worden opgevat als een "stem" van pagina X voor pagina A. Als er veel pagina's naar A linken, wordt A "populairder", d.w.z. stijgt de PageRank van A.



Figuur 2: Voorbeeld van een gerichte graaf

Maar de PageRank van A hangt niet alleen af van het aantal "stemmen" (hyperlinks naar A), maar ook van de populariteit van de pagina's die hun stem uitbrengen, dus van de PageRank van die pagina's die naar A linken. De PageRank van pagina A wordt dan volgens het originele concept van Page en Brin berekend met de volgende formule:

$$PR(A) = (1-d) + d\left(\frac{PR(T_1)}{C(T_1)} + \frac{PR(T_2)}{C(T_2)} + \cdots + \frac{PR(T_n)}{C(T_n)}\right)$$

- PR(x) is de PageRank van pagina x
- C(x) is het aantal uitgaande links van pagina x
- T1, T2, T3, … zijn de pagina's die naar A verwijzen
- d stelt de waarschijnlijkheid voor dat een surfer een willekeurige pagina opvraagt

Naverwerking

Op werkblad 5 gebruiken de leerlingen een eenvoudige manier om webpagina's punten te geven. Tijdens de nabespreking kun je dieper ingaan op het eigenlijke algoritme en de formule uittesten met behulp van een vast aantal pagina's.

Spoor de leerlingen aan om te onderzoeken wat "sponsored links" zijn. Wanneer verschijnen ze? Hoe werken ze? Krijgt iedereen steeds de zelfde resultaten?

Werkblad 1

Noteer in de tabel op welke pagina's elk woord voorkomt.

Dit is een heel eenvoudig recept voor spaghetti	Dit recept voor macaroni is heel eenvoudig	Dit recept voor spaghetti is even eenvoudig als dat voor macaroni

	Pagina
Dit	1, 2, 3
Recept	
Voor	
Spaghetti	
Is	
Heel	
Eenvoudig	
Een	
Macaroni	
Even	
Als	
dat	

Zoekmachines geven een nummer aan elke pagina op het Internet (en dat zijn er veel, ongeveer 4,5 miljard, maar geen paniek, computers werken graag met getallen!). Stel dat het Internet enkel bestond uit de bovenste 3 pagina's, dan zou de eerste pagina nummer 1 krijgen, de tweede nummer 2 en de derde nummer 3.

Zo houden zoekmachines een woordenlijst bij van woorden en de pagina's waarop ze terug te vinden zijn. Stel dat je het woord macaroni zou zoeken, dan krijg je pagina 2 en 3. Maar wat als je de woorden spaghetti en macaroni zou zoeken? In dat geval geeft de zoekmachine gemeenschappelijke paginanummers terug: spaghetti (1,3) en macaroni (2,3), resultaat is 3.

Werkblad 2

Dit is een heel eenvoudig recept voor spaghetti	Dit recept voor macaroni is heel eenvoudig	Dit recept voor spaghetti is even eenvoudig als dat voor macaroni

	Pagina en volgnummer
Dit	1-1, 2-1, 3-1
Recept	
Voor	
Spaghetti	
Is	
Heel	
Eenvoudig	
Een	
Macaroni	
Even	
Als	
dat	

Welke resultaten gaan we nu krijgen? We zoeken eenvoudig spaghetti; voor eenvoudig krijgen we 1-5 en 3-7 en voor spaghetti krijgen we 1-8 en 3-4. We zijn zeker dat beide woorden op de pagina's voorkomen. Maar welke pagina geeft het beste resultaat? Op pagina 1 heeft eenvoudig plaats 5 en spaghetti plaats 8. We weten met zekerheid dat eenvoudig voor spaghetti komt, omdat 5 voor 8 komt.

Op pagina 2 zien we dat eenvoudig na spaghetti komt. Voor deze zoekactie zal pagina 1 dus voorlopig het beste resultaat geven.

Werkblad 3

<titelStart>Recept spaghetti<titelStop> <bodyStart>Dit recept voor spaghetti is heel eenvoudig<bodySto p>	<titelStart>Recept macaroni<titelStop > <bodyStart>Dit is een heel eenvoudig recept voor macaroni<bodySto p>	<titelStart>Recepten<titelSto p> <bodyStart> Dit recept voor spaghetti is even eenvoudig als dat voor macaroni <bodyStop>

	Pagina en volgnummer
Dit	1-6, 2-6, 3-6
Recept	1-2, 1-7, 2-2, 2-11, 3-6
Voor	
Spaghetti	
Is	
Heel	
Eenvoudig	
Een	
Macaroni	
Even	
Als	
dat	
<titelStart>	
<titelStop>	
<bodyStart>	
<bodyStop>	

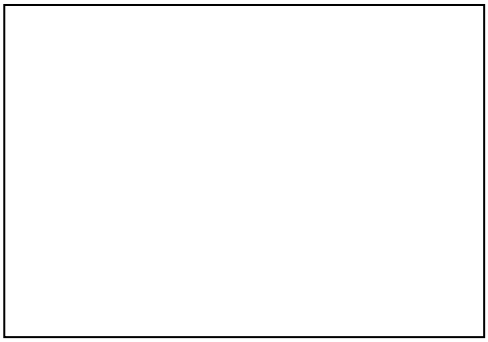
De leerlingen hebben nu 3 pagina's geïndexeerd. Doe alsof je zoekt naar een recept voor macaroni. Wat is het adres van de pagina waar je de informatie zult vinden? Het is mogelijk dat enkele leerlingen toch een ander antwoord geven. Hoe komt het dan dat de ene pagina "hoger staat" dan een andere?

Werkblad 4

URL: ……

……

……



……

……

……

……

……

……

……

……

Links:

……

……

……

……

Werkblad 5

Welke zoekterm kies je?.....

- 1 punt per keer dat het sleutelwoord in een alinea staat
- 1 punt per keer dat de term in een hoofding staat
- 1 punt per keer dat de term in de paginatitel staat
- 1 punt per keer dat de term in de url staat
- 1 punt per keer dat er naar de pagina wordt gelinkt

Gebruik de onderstaande tabel om je pagina een score te geven, trek streepjes (dit heet “turven”) om de voorkomens te tellen. Daarna bereken je het eindtotaal:

Zoekterm staat in een alinea	
Zoekterm staat in een hoofding	
Zoekterm staat in het adres	
Er wordt naar je pagina gelinkt	
Zoekterm staat in de titel	

Bronnen

PageRank. (2017) In Wikipedia. Geraadpleegd op 11 mei 2017, van https://nl.wikipedia.org/wiki/PageRank

MacCormick J. (2012). *Nine algorithms that changed the future*. Princeton, NJ: Princeton University Press

[Voorstelling van een webcrawler]. Verkregen op 11 mei 2017, van http://seo-advisors.com/wp-content/uploads/2013/07/Web-Crawler-Route-Map-Chart.jpg

[Willekeurig gegenereerde graaf]. Verkregen op 11 mei 2017, van http://computationalculture.net/wp-content/uploads/2012/08/ILLUSTRATION3.PNG.png