



# Trefwoorden, Woordwolken, Verwante woorden

## Open dag op locatie

Gosse Bouma

g.bouma@rug.nl  
Informatiekunde  
Rijksuniversiteit Groningen

5 maart 2010

# Overzicht

## Zoekwoorden en woorden tellen

**Zoekwoorden** (op het web) en **woorden tellen** (op een webpagina, in een tekst) levert een schat aan informatie op

## Informatie

[www.let.rug.nl/gosse/opendag](http://www.let.rug.nl/gosse/opendag)

# Verwante Woorden

## Toepassingen

- **Advertenties**

- Google Sponsored links
- Welke producten passen bij welke zoektermen?

- **Trefwoorden / Wordwolken**

- Wat zijn de belangrijkste woorden in een tekst?
- Geven een indruk van het onderwerp

- **Vaste Verbindingen**

- Welke woorden komen vaak samen voor?
- harde+schijf, buitenaardse+wezens, Sven+Kramer

# Trends in Google Zoekwoorden



# Google Ads

wielrennen - Google zoeken - Mozilla Firefox

File Edit View History Bookmarks Tools Help deljicio.us

Aanmelden

Het Internet Afbeeldingen Nieuws Maps Nieuw! Discussie

Google™ wielrennen

Zoek:  het Internet  pagina's in het Nederlands  pagina's

Het Internet Resultaten 1 - 10 van circa 4.860.000 voor **wielrennen** (0,11 seconden)

[wielrennen.startpagina.nl](http://wielrennen.startpagina.nl)  
wielrennen.startpagina.nl, de beste link-pagina op het gebied van het **wielrennen**.  
[wielrennen.startpagina.nl/](http://wielrennen.startpagina.nl/) - 124k - [In cache](#) - [Gelijkwaardige pagina's](#)

[Wielrensite.nl - al het wielernieuws op één site](http://Wielrensite.nl)  
5 13:49 - Ludovic Auger stopt met **wielrennen**. 3 13:45 - Schmidt wordt ploegleider bij Team C.S.C. 21 13:40 - Di Luca drie maanden geschorst ...  
[www.wielrensite.nl/](http://www.wielrensite.nl/) - 20k - [In cache](#) - [Gelijkwaardige pagina's](#)

[Rabobank - Wielrennen](http://Rabobank - Wielrennen)  
Rabobank investeert fors in **wielrennen** in de meest brede zin van het woord. Zo is Rabobank hoofdsponsor van de Koninklijke Nederlandse Wielren Unie (KNWU), ...  
[www.rabobank.nl/particulieren/sponsoring/wielrennen/](http://www.rabobank.nl/particulieren/sponsoring/wielrennen/) - 17k - [In cache](#) - [Gelijkwaardige pagina's](#)

[Wielrennen.blog.nl](http://Wielrennen.blog.nl)  
Bedankt voor je bezoek aan **wielrennen.blog.nl** de digitale ontmoetingsplaats voor alle aanhangers van de wielersport. ...  
[wielrennen.blog.nl/](http://wielrennen.blog.nl/) - 100k - [In cache](#) - [Gelijkwaardige pagina's](#)

Gesponsorde Koppelingen

[Het Laatste Sportnieuws](#)  
Nu direct het laatste sportnieuws op je iGoogle startpagina!  
[www.google.nl/ig](http://www.google.nl/ig)

[Duursportwinkel](#)  
Speciaalzaak in triathlon, duathlon zwemmen, hardlopen en fietsen  
[www.duursportwinkel.nl](http://www.duursportwinkel.nl)

[Uw bericht hier?](#)

# Trefwoorden voor een product

## Google Ads

<b>Zoekterm</b>	<b>Advertentie</b>
centrino	dell.nl
racefiets	gazelle.nl
wielrennen	google.nl (sportnieuws)
wielrennen	vermageren.com
lance armstrong	quadrand.com (polsbandjes)

# Google Advertenties

## Google Adwords

- Adverteerders (bedrijven/instellingen) bieden geld voor een zoekterm (adword)
- Hoe meer geld je biedt, hoe groter de kans dat je advertentie getoond wordt.
- Adverteerder betaalt aan Google iedere keer wanneer er op een advertentie geclickt wordt.

# Opdracht 1

[www.let.rug.nl/gosse/opendag](http://www.let.rug.nl/gosse/opendag)

## Welke universiteiten en HBO's gebruiken Google Ads?

- Welke onderwijsinstelling (HBO, Universiteit) in Groningen maakt het meeste reclame via Google Ads?
  - Zoek bv naar *studeren/studievoorlichting/voorlichting Groningen*
- Welke Nederlandse universiteit maakt het meeste reclame via Google Ads?
  - Zoek bv naar *studeren Groningen/Leiden/Utrecht/Amsterdam*



# Trefwoorden en Woordwolken

## Trefwoorden vinden

- **Trefwoorden** geven een idee van de inhoud van de tekst
- **Woordwolken** geven een idee van de inhoud van een website
  - **del.icio.us, flickr, ...**
- Woordwolken zijn **de meest gebruikte** woorden
- Veel gebruikte woorden zijn vaak goede trefwoorden

# del.icio.us

## ■ My Delicious Tags

semanticweb rdf imported fietsen wikipedia ontology xml nlp  
corpustaalkunde biografie CL readability search tutorial ir sparql qa wikifier searchengine  
metadata google annotation ie perl heei xpath web terminology visualization dutch reference javascr  
semantic monetdb R database openaccess linux dictionary w3c catch java ontologies prolog wordnet tekstm  
statistics php journal weer travel owl api imix boulder internet information-retrieval coref clustering rss rst ajax skos cultur  
hypemyns markup papers cyclo nederland semanticwiki pagerank clarin uima dependencies opensource firefox tren  
relationdiscovery Machinelearning biomed webklas googlemaps similarity freebase webcorpus weather zoeken citepeer cief subside  
culturalheritage

# TagCrowd



# Wordenwolken van webpagina's

## Opdracht 2

- Kies een webpagina met veel tekst (kennislink, nu.nl, ...)
- Maak de woordwolk op [tagcrowd.com](https://tagcrowd.com)

# Woordenwolken

## Van woordwolk naar trefwoorden

- Algemene woorden (**en, een, de, het, van, ...**) komen wel veel voor maar zijn geen goede trefwoorden
- Sommige woordcombinaties (**rode bloedlichaampjes, stoffelijk overschot**) wil je als 1 trefwoord behandelen

# Vaste Verbindingen

## Vaste verbindingen vinden

- Vaste verbindingen zijn combinaties van 2 of meer woorden met een speciale betekenis
- *Max Havelaar, buitenaardse wezens, kunstmatige intelligentie, patates frites, rector magnificus, toffe peer*

# Vaste Verbindingen

## Vaste verbindingen vinden in een tekst

- Vaste verbindingen zijn woorden die relatief vaak naast elkaar staan
- Als ik *toffe* lees, is de kans vrij groot dat *peer* het volgende woord is.
- Je wilt dus weten voor welke woordparen **Woord1+Woord2** geldt dat ze ongeveer even vaak voorkomen als **Woord1** of **Woord2** alleen....

# Vaste Verbindingen

## Opdracht 3

Vaste verbindingen gezocht in 1 miljoen woorden Wikipedia.

- Welke verbindingen vinden we?



# Tot Slot

## (zoek-)woorden tellen

- Trends onderzoeken
- Advertenties plaatsen
- Belangrijke woorden in een tekst vinden
- Vaste verbindingen ontdekken

## Informatiekunde in Groningen

- Werken met web-data en grote tekstbestanden
- Informatie zoeken
- Voor web-toepassingen of taalkundig onderzoek