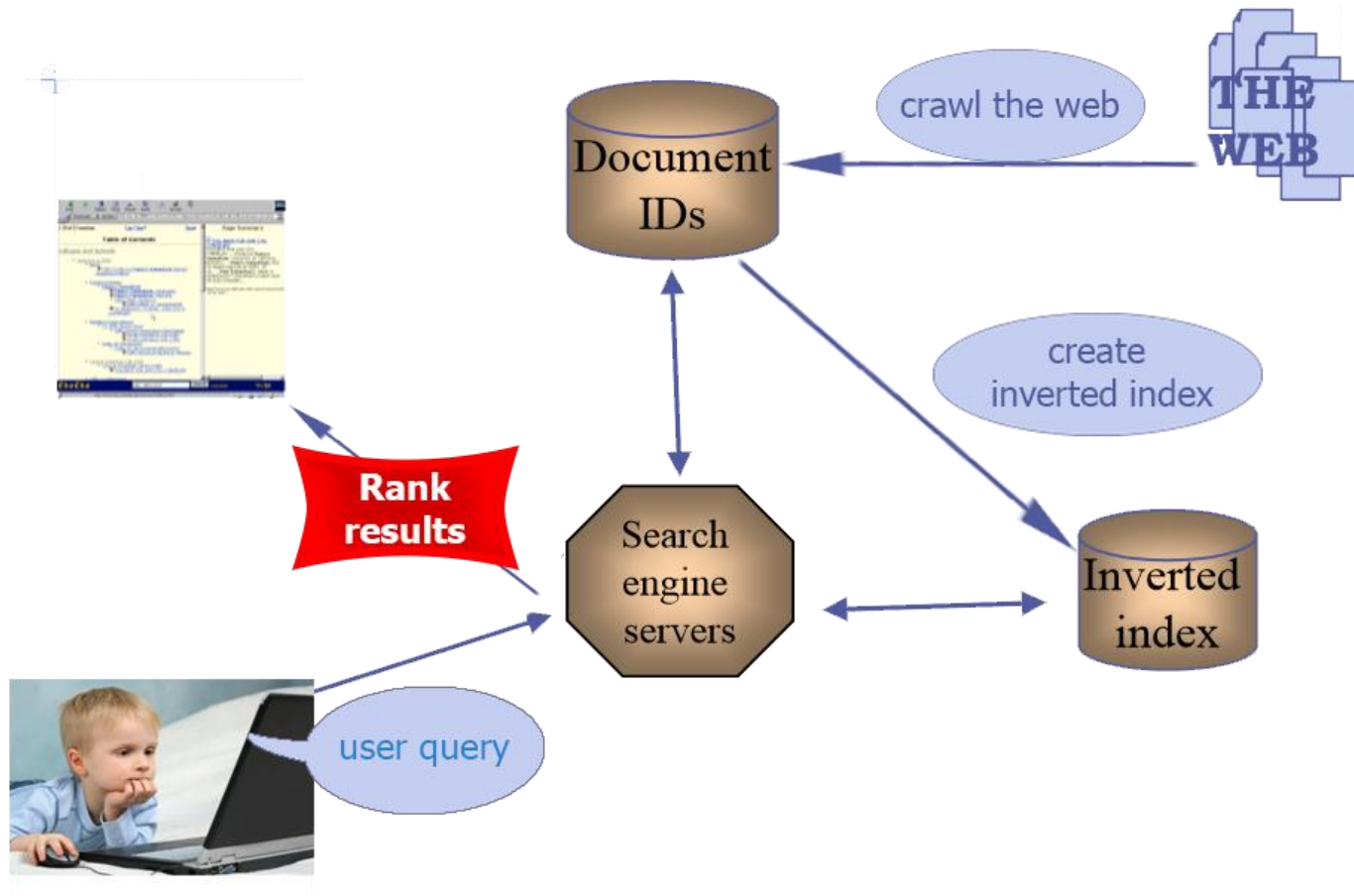
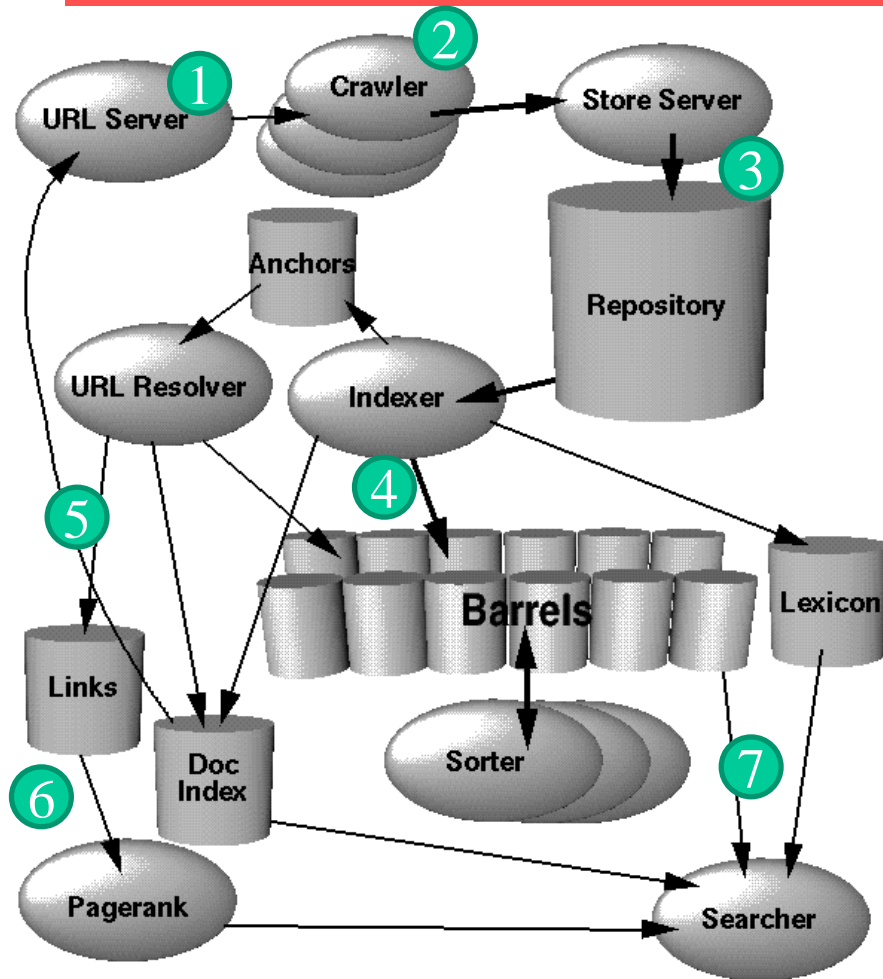

Chapter 10

Google Technology

Google Model



Google Model



ขั้นที่ 1 การรวบรวมข้อมูลเว็บจากการดาวน์โหลดหน้าเว็บ URLserver จะส่งรายชื่อของ URL ไปที่ Crawlers

ขั้นที่ 2 หน้าเว็บที่ถูกดาวน์โหลดแล้วจะถูกส่งไปยัง Store server

ขั้นที่ 3 เซิร์ฟเวอร์จัดเก็บจะทำการบีบอัดแบบ Zlib และนำไปจัดเก็บใน Repository

ขั้นที่ 4 DocID จะถูกกำหนดให้แต่ละหน้าเว็บ ตัวทำดัชนีและตัวเรียงลำดับ จะทำงานร่วมกันในการจัดทำดัชนี

ขั้นที่ 5 ทำการคำนวณค่าของที่ค้นเจอและการเรียงลำดับลิงค์ทั้งหมด โดยมี 2 วิธี คือ Forward Index และ Inverted Index >>

ขั้นที่ 6 URL resolver ใช้ในการจัดอันดับ Rank ของทุกเอกสาร

ขั้นที่ 7 ตัวเรียงลำดับ WordID สำหรับการค้นหาใหม่ๆ

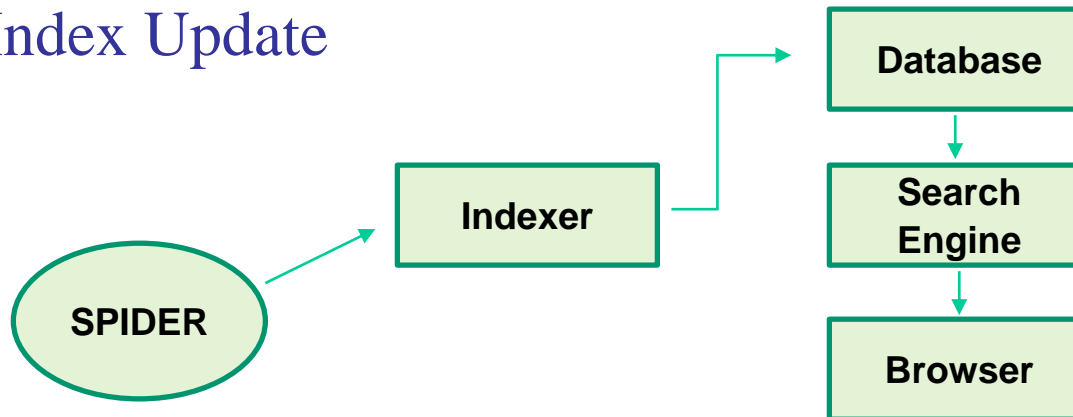
Google Crawling & Indexing

การทำงานของ Google Search

- Query → Google จัดทำดัชนีขึ้น → Index Comparison → Output
- Googlebot (Spider)
 - รวบรวมข้อมูล
 - ค้นหาเว็บไซต์ใหม่ๆ
 - Index Update

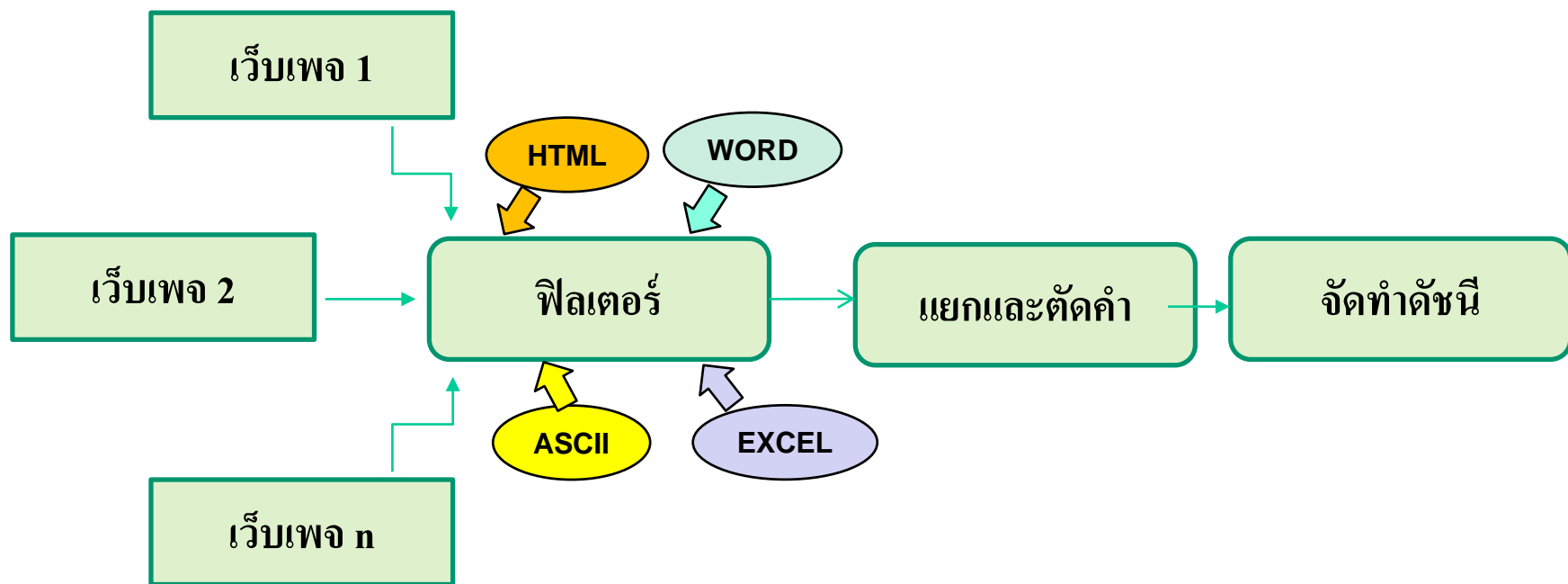
Google Crawling

- Googlebot หรือ Spider
 - รวบรวมข้อมูล (ความถี่ที่พบ และ จำนวนหน้าเว็บ)
 - ค้นหาหน้าเว็บใหม่ๆ
 - Index Update



Google Indexing

- การจัดทำดัชนีจะจัดทำขึ้นโดย อินเด็กเซอร์ โดยจะรับข้อมูลจากสไปเดอร์ มาจัดทำดัชนีอีกที ขั้นตอนการทำงานของอินเด็กเซอร์แบ่งออกได้เป็น 3 ขั้นตอนหลัก คือ
 - กรองคำด้วยฟิลเตอร์
 - แยกและตัดคำ
 - จัดทำดัชนี



Result & PageRank

การแสดงผลการค้นหา

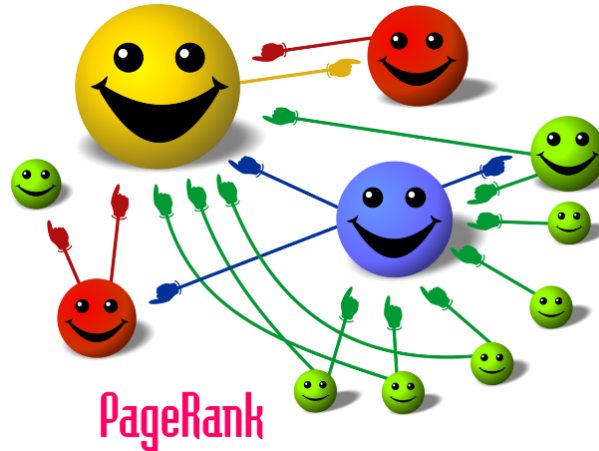
- PageRank ใช้สำหรับวัดความสำคัญของหน้าเว็บ



PageRank

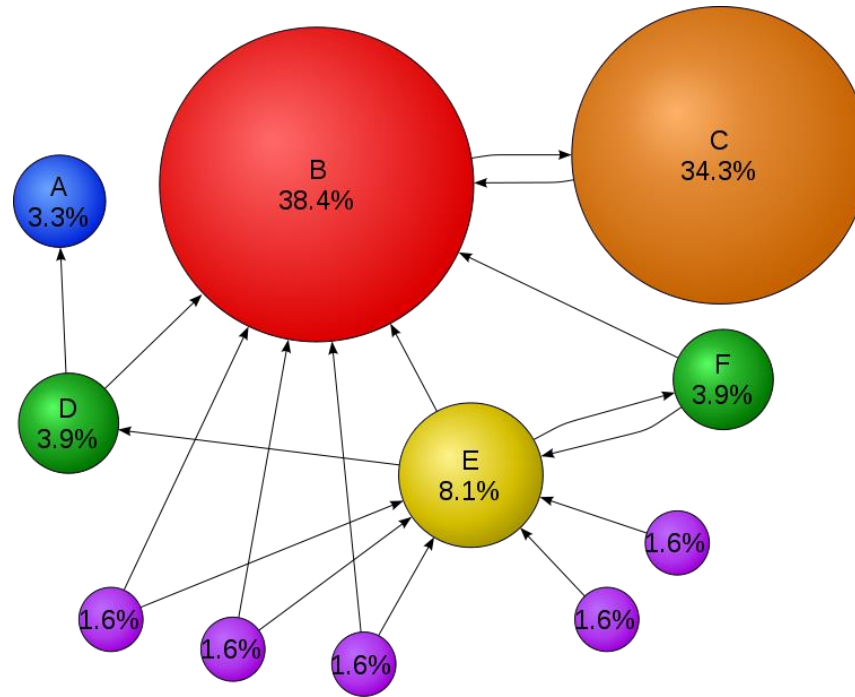
- เพจเรงก์ (PageRank) คือ ค่าตัวเลขบ่งบอกถึงความสำคัญของเว็บเพจนั้น ๆ และการอ้างอิงจากหน้าอื่น
- ค่าระหว่าง 0 ถึง 10
- เว็บเพจที่มี PageRank สูงแสดงก่อน
- ค่าของเพจเรงก์ถูกคำนวณจากจำนวนการอ้างอิงจากหน้าอื่น

PageRank



- PageRank คล้ายกับการโหวตคะแนนเสียง ความสำคัญจะขึ้นกับ**จำนวนลิงก์ที่เข้ามา**
เกี่ยวข้องมายังเพจ เป็นคะแนนโหวต ถ้าไม่มีลิงก์ ก็ถือว่าไม่มีคะแนนโหวต
- ลิงก์ต่างๆ ไม่ได้มีความสำคัญเหมือนกันทั้งหมด Google ได้พยายามที่จะแบ่งแยกลิงก์ต่างๆออก ว่าเป็นสแปม(ขยะ)หรือไม่ มีความปลอดภัยหรือไม่
- การจะทำให้เว็บไซต์ที่เราสร้างนั้นมีอันดับสูงสุดในการค้นหา เราจะต้องพัฒนาเว็บไซต์ให้ Google สามารถรวบรวมข้อมูลและทำดัชนีจากเว็บของเราให้มากที่สุด

PageRank



- หน้า B มีค่าเพจเรงก์สูงสุด เพราะมีจำนวนหน้าที่ลิงก์เข้าหามากสุด
- หน้า C มีเพจเรงก์สูงกว่าหน้า E แม้ว่าจะมีหน้าที่ลิงก์มาหาน้อยกว่า

Google Query

1.คำที่ใช้ในการค้นหาจะถูกนำมาใช้ทั้งหมด

Google จะแสดงผลเฉพาะหน้าที่มีคำตรงกับคำที่ใช้ในการค้นหาทั้งหมดเท่านั้น

ตัวอย่างเช่น คำที่ใช้ในการค้นหา คือ "compact fold-up bicycle" Google จะทำการค้นหาหน้าที่มีคำ "compact" and "fold-up" and "bicycle" โดยที่ผู้ค้นหาไม่จำเป็นต้องใส่ and ในระหว่างคำเอง การทำงานแบบนี้เรียกว่า "implicit AND"



compact fold-up bicycle



Web

Shopping

Images

Videos

News

More ▾

Search tools

About 1,880,000 results (0.28 seconds)

Folding Bikes Online Shop - lazada.co.th



Ad [folding-bikes.lazada.co.th/](https://www.lazada.co.th/folding-bikes) ▾

Quality Outdoor Gear & Equipment. Shop Professional Sports Brands!

โปรของใช้ในบ้าน ลดถึง 80% - ช้อปง่ายๆ ช้อปที่สาขาตัว - เครื่องสำอางค์แบรนด์ดัง

Folding bikes & compact bicycles by Brompton, Dahon, Tern ...

www.nycewheels.com/folding-bike.html ▾

Small enough to fit in your closet, in the trunk of your car, or the luggage rack of a train, a **fold up bike** can be taken anywhere. Our selection of good quality ...

Fold-up bikes are not to be overlooked - NYCeWheels

www.nycewheels.com › ... › [Folding bike articles](#) ▾

Some **fold-up bicycles** are meant to fold strictly for storage, while other **fold-up bicycles** are meant to fold **compact** enough to be taken anywhere. Some **fold-up** ...

2.ผลที่ได้จากการค้นหาจะตรงกับคำที่ใช้เรียกค้น

Google จะ แสดงผลเฉพาะหน้าที่มีคำตรงกับคำที่ใช้ในการค้นหาอย่างถูกต้องเท่านั้น

ตัวอย่างเช่น

If you search for ...	Google won't find ...
cheap	inexpensive
tv	television
effects	influences
children	kids
car	automobile
Calif OR CA	California

Google จะขยายความตัวย่อด้วย เช่น

If you search for ...	Google finds ...
NYC	New York City
SF	San Francisco
GNP	Gross National Product

3. ค้นหาใกล้เคียงกับคำที่ใช้ในการเรียกค้น

Google จะทำการค้นหาหน้าที่มีคำที่ได้จากการแปลงคำค้นหาที่ใช้ด้วย ตัวอย่างเช่น

ใช้คำค้นหาว่า **child bicycle helmet** ระบบจะทำการค้นหาคำที่ใส่รวมถึงคำที่มีรากศัพท์เดียวกันกับคำที่ใส่ด้วย โดยคำที่ระบบใช้จะเป็นดังนี้ “child,” “children,” or “children's,” “bicycle,” “bicycles,” “bicycle's,” “bicycling,” or “bicyclists,” and “helmet” or “helmets.”

4. Stopwords

สำหรับคำบางคำที่ถือเป็น Stopwords(เช่น the, on, where, how, de, la, ตัวเลขตัวเดียว, ตัวอักษรตัวเดียว เป็นต้น) Google จะถือว่าคำประเภทนี้ไม่มีความหมายและไม่เอามาใช้ในการค้นหา

ตัวอย่างเช่น ค้นหาคำว่า What to read for a London tourist คำที่ Google นำมาใช้ในการค้นหาคือ "read" and "London" and "tourist"



What to read for a London tourist



Web

Shopping

News

Images

Videos

More ▾

Search tools

About 102,000,000 results (0.43 seconds)

US tourist locked inside London bookshop - The Guardian

www.theguardian.com › [World](#) › [UK News](#) › [London](#) ▾ The Guardian ▾

Oct 17, 2014 - US **tourist** locked inside **London** bookshop. Police called to store

Read 'em and sleep: how one tweet led to a literary lock-in. After trapping a ...

London Tourist Information Centres - visitlondon.com

www.visitlondon.com/tag/tourist-information-centre ▾ London ▾

Call into one of **London's** Tourist Information Centres for local information, maps and ... There are Tourist Information Centres all over **London**. ... **Read More** ...

U.S. tourist gets trapped in London bookstore - CNN.com

www.cnn.com/2014/10/17/travel/tourist-trapped-in-bookstore/ CNN ▾

Oct 17, 2014 - U.S. **tourist** gets trapped in **London** bookstore. By Barry Neild, CNN ...

The theater boxes have been turned into **reading** spaces. Paris Left Bank ...

5. จำกัดจำนวนคำ

Google นั้นจำกัดจำนวนคำที่ใช้ในการค้นหาได้มากที่สุดอยู่ที่ **32** คำ(ไม่รวม **Stopwords**) หากมีการใช้คำจำนวนมากกว่า **32** คำในการค้นหา Google จะใช้เฉพาะ **32** คำแรกเท่านั้นในการค้นหา



: about above abrade abridge abroad abrupt abscond absent absinthe amy amen



Web

Images

Maps

Shopping

More ▼

Search tools

About 5,850 results (1.08 seconds)

"amen" (and any subsequent words) was ignored because we limit queries to 32 words.

Did you mean: aardvark aback abacus abalone abandon
abashed abbey abbreviate abdicate abdomen abduct aberration
abhor abide ability object able abnormal aboard abode abolish
abolitionist abort about above **abraded** abridge abroad abrupt
abscond absent absinthe amy amen

wordlist - MIT

web.mit.edu/kilroi/Public/.../wordli... ▼ Massachusetts Institute of Technology ▼



AACHEN AAL AARDVARK AARDVARKS AARON ABACI **ABACK ABACUS**
ABACUSES ABAPT **ABALONE** ABALONES **ABANDON** ABANDONED ... ABILITIES
ABILITY **OBJECT** **ABJECTED** **ABJECTING** **ABJECTLY** **ABJECTS** **ABJURATION** ...
ABNORMALLY **ABOARD** **ABODE** **ABODED** **ABODES** **ABODING** **ABOLISH** ...

6.ระยะห่างระหว่างคำ

Google จะพิจารณาระยะห่างระหว่างคำที่ใช้ในการค้นหาด้วยตัวอย่างเช่นคำว่า **snake grass** กับคำว่า **snake in the grass**

- **snake grass** หมายถึงพืชชนิดหนึ่ง การค้นหาคำ "snake" and "grass" แต่จะให้ความสำคัญกับหน้าเว็บที่มีทั้งสองคำอยู่ติดกันมากกว่า

6.ระยะห่างระหว่างคำ









[Web](#) [Images](#) [Shopping](#) [Videos](#) [News](#) [More ▾](#) [Search tools](#)

About 1,780,000 results (0.29 seconds)

Images for snake grass

[Report images](#)



[See results for snake grass](#)

Grass snake
The grass snake is a water snake

[More images for snake grass](#)

Equisetum - Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/Equisetum ▾ Wikipedia ▾

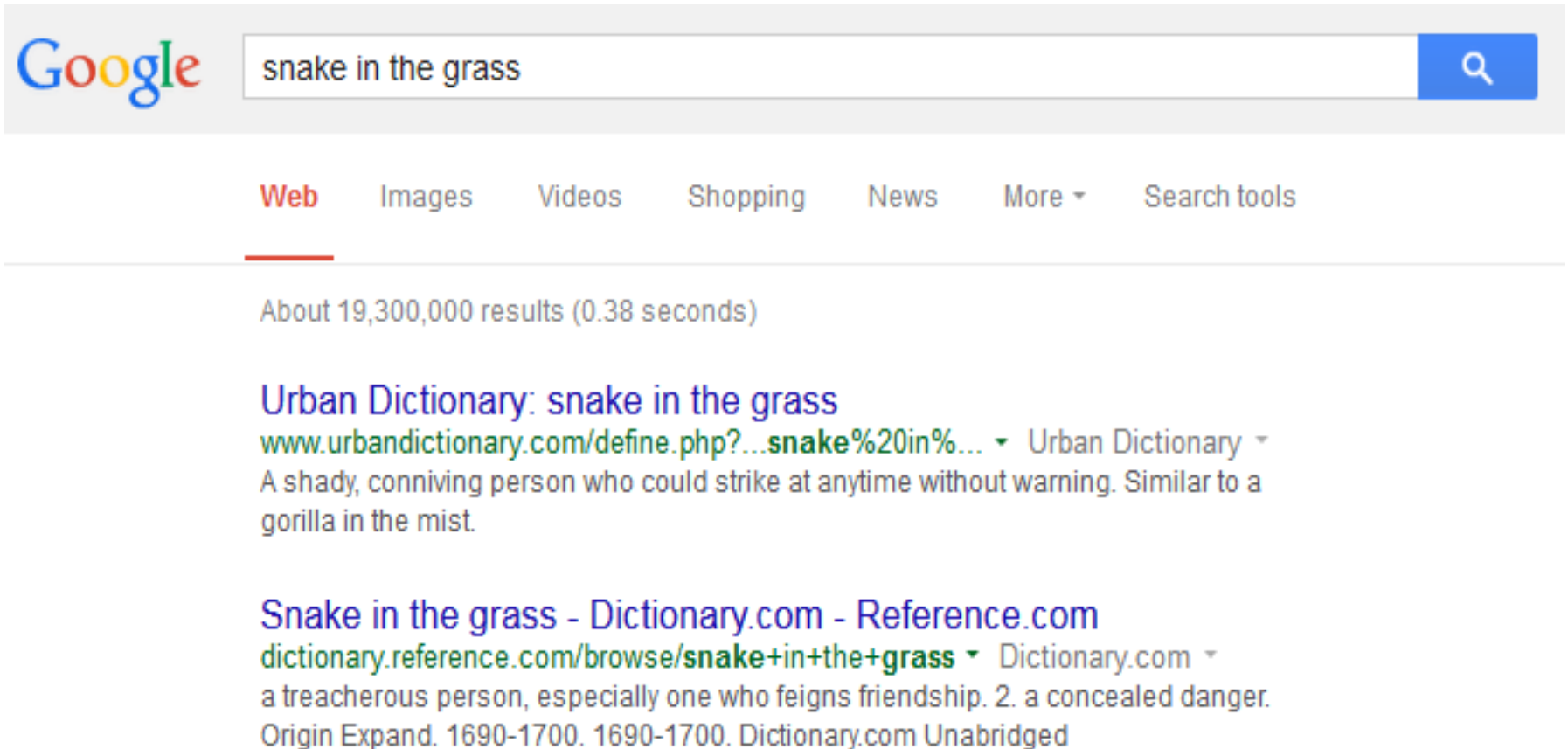
Equisetum (/ˈɛkwɪˈsiːtəm/; horsetail, **snake grass**, puzzleglass) is the only living genus in Equisetaceae, a family of vascular plants that reproduce by spores ...

[Hippuris - Equisetum arvense - Equisetum telmateia - Category:Equisetum](#)

6.ระยะห่างระหว่างคำ

- **snake in the grass** เป็นสำนวน

Google จะไม่สนใจ in, the ที่เป็น stopwords และจะทำการค้นหาคำว่า "snake" and "grass" แต่จะให้ความสำคัญกับหน้าที่มีทั้ง 2 คำปรากฏโดยมีระยะห่างระหว่างกันเป็นระยะ 2 คำมากกว่า



The screenshot shows a Google search interface with the query "snake in the grass" entered in the search bar. Below the search bar, there are tabs for "Web", "Images", "Videos", "Shopping", "News", "More", and "Search tools". The "Web" tab is selected. The search results show "About 19,300,000 results (0.38 seconds)". The first result is from Urban Dictionary, titled "Urban Dictionary: snake in the grass", with a URL that includes "snake%20in%...". The definition provided is: "A shady, conniving person who could strike at anytime without warning. Similar to a gorilla in the mist." The second result is from Dictionary.com, titled "Snake in the grass - Dictionary.com - Reference.com", with a URL that includes "snake+in+the+grass". The definition provided is: "a treacherous person, especially one who feigns friendship. 2. a concealed danger. Origin Expand. 1690-1700. 1690-1700. Dictionary.com Unabridged".

7. ลำดับคำ

Google จะให้ความสำคัญกับหน้าที่มีคำเรียงตามลำดับตรงกับที่ใช้ในการค้นหามากกว่า

ตัวอย่างเช่น New York Library กับ New Library of York ที่จะได้ผลการค้นหาที่ต่างกัน ถึงแม้จะมีคำเหมือนกันก็ตาม

8.อักษรใหญ่-เล็ก

Google จะไม่สนใจเรื่องอักษรใหญ่เล็ก ตัวอย่างเช่น
คำค้นหา

- Red Cross
- red cross
- RED CROSS

จะได้ผลการค้นหาเหมือนกัน

9.อักขระพิเศษ

สำหรับ Google แล้ว จะมีอักขระพิเศษบางตัว ที่จะไม่ถูกนำมาใช้ในการค้นหา เช่น ! ? , . ; [] @ / # < > . เนื่องจาก Google จะให้ความสำคัญกับคำที่อยู่ติดกับอักขระเหล่านั้นมากกว่า แต่จะมีในบางกรณีที่จะนำอักขระเหล่านี้มาใช้ เช่น C++, \$99 เป็นต้น ตัวอย่างเช่น

- Dr. Ruth

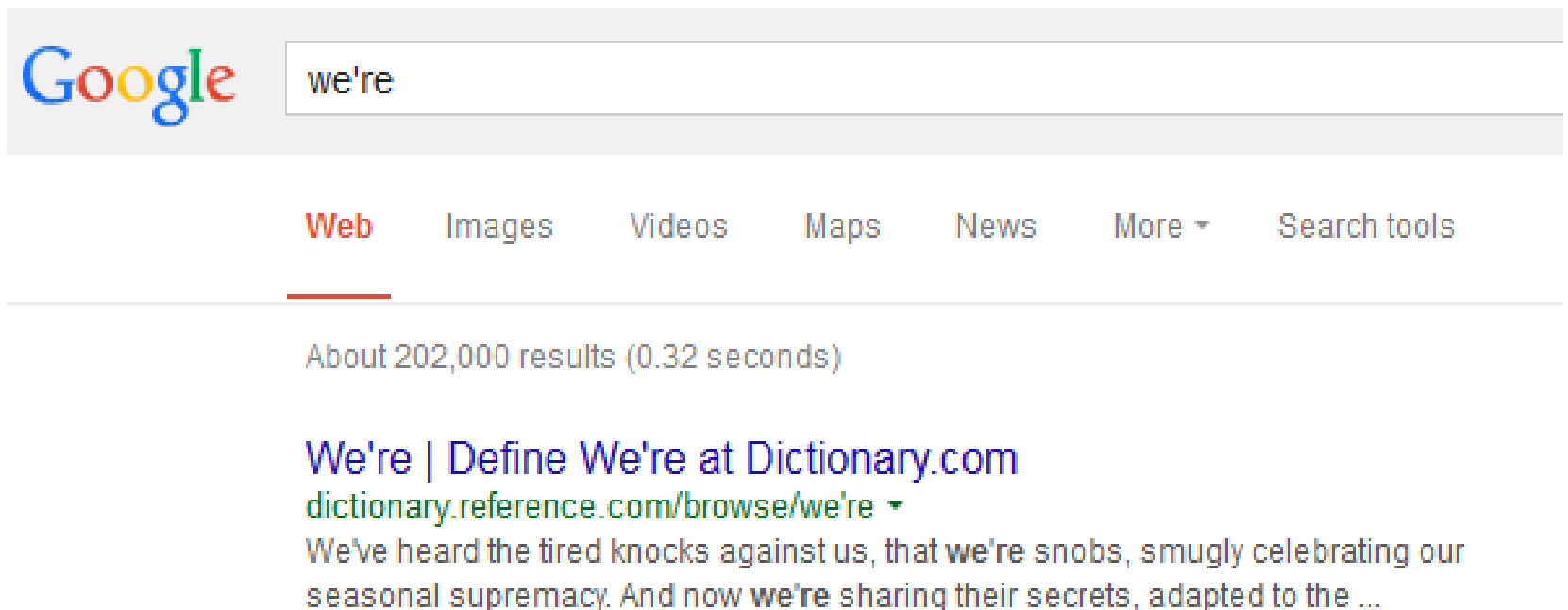
- Dr Ruth

จะให้ผลการค้นหาเหมือนกัน

10.Apostrophes

คำที่มีเครื่องหมาย ' จะถือว่าแตกต่างกับคำที่ไม่มี ตัวอย่างเช่น

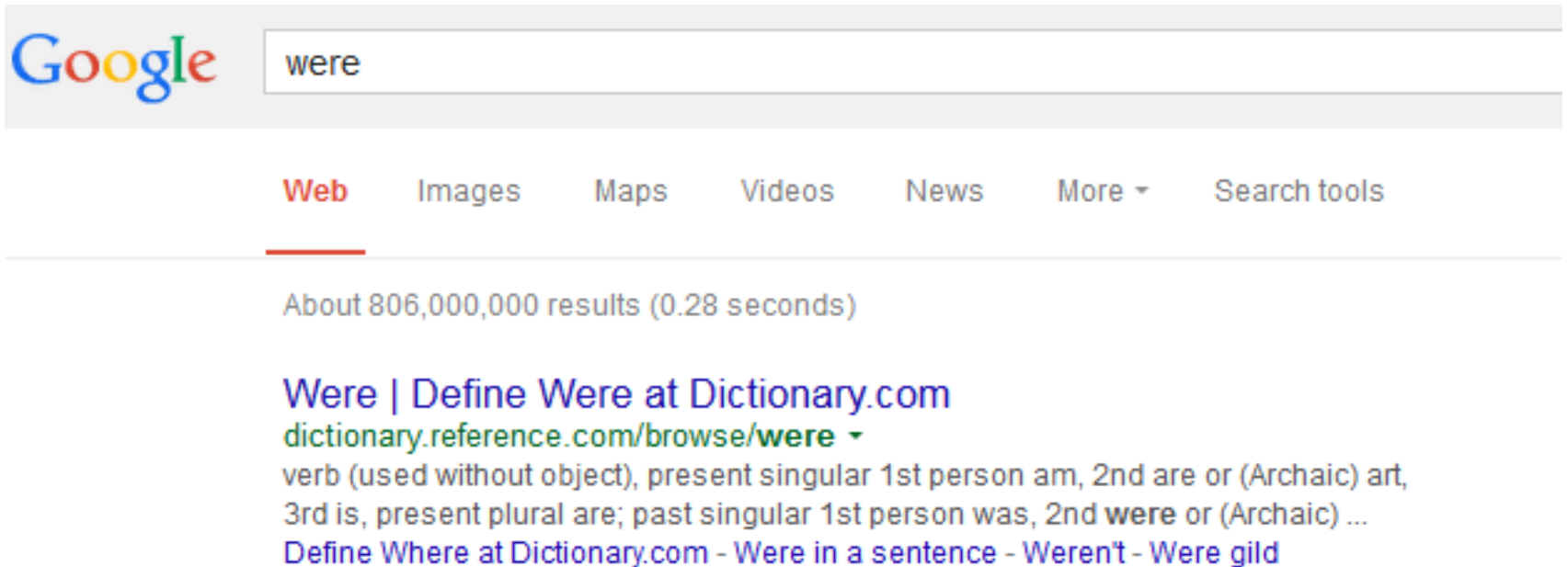
we're จะถูกเปลี่ยนเป็น "we're" แต่ไม่ใช่ "were"



The screenshot shows a Google search interface. The search bar contains the text "we're". Below the search bar, there are tabs for "Web", "Images", "Videos", "Maps", "News", "More", and "Search tools". The "Web" tab is selected. Below the tabs, it says "About 202,000 results (0.32 seconds)". The first search result is titled "We're | Define We're at Dictionary.com" with a URL "dictionary.reference.com/browse/we're". The snippet of the result reads: "We've heard the tired knocks against us, that we're snobs, smugly celebrating our seasonal supremacy. And now we're sharing their secrets, adapted to the ...".

10.Apostrophes

และ were จะถูกเปลี่ยนเป็น "were" แต่ไม่ใช่ "we're"



The screenshot shows a Google search interface. The search bar contains the word "were". Below the search bar, the "Web" tab is selected. The search results show "About 806,000,000 results (0.28 seconds)". The top result is titled "Were | Define Were at Dictionary.com" with a URL "dictionary.reference.com/browse/were". The snippet below the title reads: "verb (used without object), present singular 1st person am, 2nd are or (Archaic) art, 3rd is, present plural are; past singular 1st person was, 2nd were or (Archaic) ... Define Where at Dictionary.com - Were in a sentence - Weren't - Were gild".

Google were

Web Images Maps Videos News More ▾ Search tools

About 806,000,000 results (0.28 seconds)

Were | Define Were at Dictionary.com
dictionary.reference.com/browse/were ▾
verb (used without object), present singular 1st person am, 2nd are or (Archaic) art, 3rd is, present plural are; past singular 1st person was, 2nd were or (Archaic) ...
[Define Where at Dictionary.com](#) - [Were in a sentence](#) - [Weren't](#) - [Were gild](#)

11.เครื่องหมายขีด

คำบางคำอาจจะเขียนไม่เหมือนกันเช่น E-mail , E mail, Email แต่จะให้ความหมายที่เหมือนกัน ดังนั้นหาก Google พบเครื่องหมายขีดในคำค้นหา Google จะทำการแปลงคำค้นหาให้ครอบคลุมในทุกกรณี ดังตัวอย่างต่อไปนี้

-
- คำค้นหา E-mail
 - สิ่งที Google ค้นหาคือ "E-mail", "E mail" และ "Email"
 - คำค้นหา E mail
 - สิ่งที Google ค้นหาคือ "E-mail" และ "E mail"
 - คำค้นหา Email
 - สิ่งที Google ค้นหาคือ "Email"