

CC5212-1

PROCESAMIENTO MASIVO DE DATOS

OTOÑO 2019

Lecture 8

Information Retrieval: Ranking

Aidan Hogan

aidhog@gmail.com

Apache Lucene

Lucene

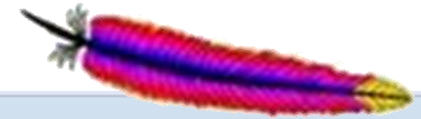


- Inverted Index
 - They built one so you don't have to!
 - Open Source in Java

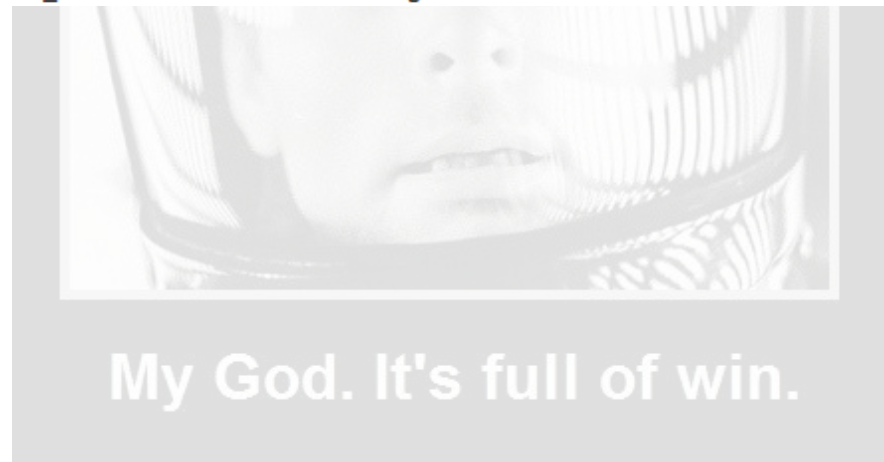


My God. It's full of win.

Apache Lucene

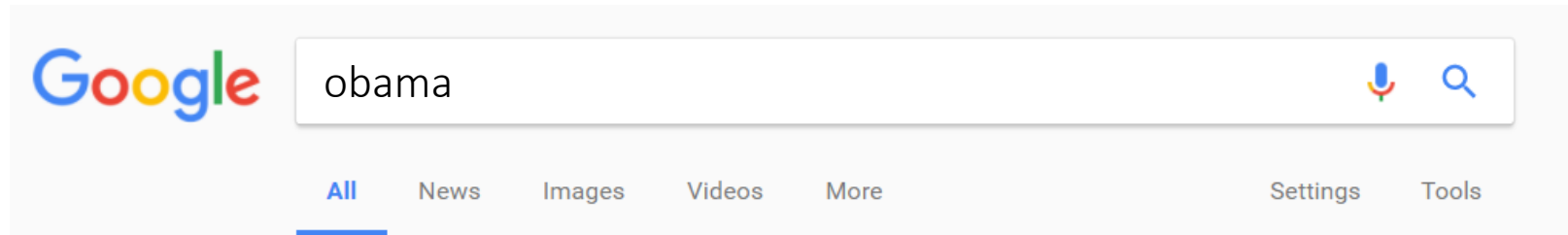
The logo for Apache Lucene, featuring the word "Lucene" in a stylized, green, outlined font with a feather-like tail on the left.

```
Tasks Console
SearchWikiIndex [Java Application] C:\Program Files\Java\jre1.8.0_65\bin\javaw.exe (03-05-2017 12:45:22 a. m.)
Opening directory at lucene
Enter a keyword search phrase:
obama
Running query: obama
Parsed query: TITLE:obam^5.0 ABSTRACT:obam
Matching documents: 255
Showing top 10 results
1 http://es.wikipedia.org/wiki/Obama_Republican Obama Republican
2 http://es.wikipedia.org/wiki/Obama_(Fukui) Obama (Fukui)
3 http://es.wikipedia.org/wiki/Republicanos_por_Obama Republicanos por Obama
4 http://es.wikipedia.org/wiki/Engonga_Obame Engonga Obame
5 http://es.wikipedia.org/wiki/Barack_Obama Barack Obama
6 http://es.wikipedia.org/wiki/Michelle_Obama Michelle Obama
7 http://es.wikipedia.org/wiki/Cartel_%22Hope%22_de_Obama Cartel "Hope" de Obama
8 http://es.wikipedia.org/wiki/Transici3n_presidencial_de_Barack_Obama Transici3n presidencial de Barack Obama
9 http://es.wikipedia.org/wiki/Por_qu3_Obama_ganar3_en_2008_y_en_2012 Por qu3 Obama ganar3 en 2008 y en 2012
10 http://es.wikipedia.org/wiki/Ricardo_Mangue_Obama_Nfubea Ricardo Mangue Obama Nfubea
```




INFORMATION RETRIEVAL: RANKING

How Does Google Get Such Good Results?



About 462,000,000 results (0.71 seconds)


Barack Obama (@BarackObama) · Twitter

<https://twitter.com/BarackObama> 

Well said, Jimmy. That's exactly why we fought so hard for the ACA, and why we need to protect it for kids like Billy. And congratulations! [twitter.com/jimmykimmel...](https://twitter.com/jimmykimmel)


11 hours ago · Twitter

The Office of Barack and Michelle Obama

<https://www.barackobama.com/> 

Welcome to the Office of Barack and Michelle **Obama**. We Love You Back. Play video. The Office of Barack and Michelle **Obama**. © 2017 | Legal & Privacy.

Barack Obama - Wikipedia

https://en.wikipedia.org/wiki/Barack_Obama 

Barack Hussein **Obama** II is an American politician who served as the 44th President of the United States from 2009 to 2017. He is the first African American to ...



How does Google Get Such Good Results?

Google that one movie where the guy breaks his leg and spies on his neighbor

Web Videos News Images Shopping More Search tools

About 64,700,000 results (0.91 seconds)

[Rear Window \(1954\) - IMDb](#)
www.imdb.com/title/tt0047396/ - Internet Movie Database
★★★★★ Rating: 8.6/10 - 274,497 votes

Google da da da dum symphony

Web Videos News Shopping Images More Search tools

About 107,000 results (0.36 seconds)



Beethoven - Symphony No. 5 in C Minor (1) - YouTube
www.youtube.com/watch?v=W2qW6fOtAMY



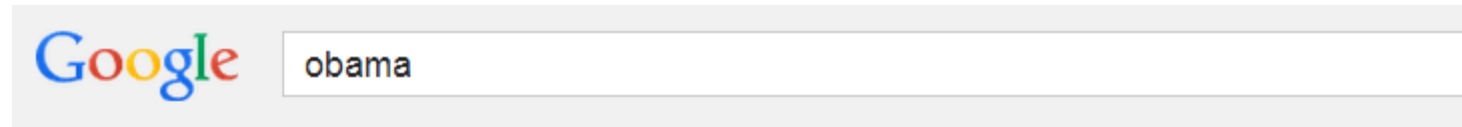
sometimes when i'm

- sometimes when i'm **alone i use comic sans**
- sometimes when i'm **alone i google myself**
- sometimes when i'm **alone i cry**
- sometimes when i'm **all alone**
- sometimes when i'm **dreaming**
- sometimes when i'm **sad i like to cut myself**
- sometimes when i'm **dreaming lyrics**
- sometimes when i'm **alone**
- sometimes when i'm **driving on the road at night**
- sometimes when i'm **alone i wonder**

Google Search I'm Feeling Lucky

TWO ASPECTS OF RANKING:
RELEVANCE VS. IMPORTANCE

Two Sides to Ranking: Relevance



Web Images News Videos More ▾ Search tools

About 16,700,000 results (0.23 seconds)

Broccoli - Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/Broccoli ▾

Broccoli is an edible green plant in the cabbage family, whose large flowering head is used as a vegetable. The word **broccoli** comes from the Italian plural of ...

[Cauliflower](#) - [Romanesco broccoli](#) - [Broccoli \(disambiguation\)](#) - [Broccolini](#)

Broccoli - The World's Healthiest Foods

www.whfoods.com/genpage.php?tname=foodspice&dbid=9 ▾

Broccoli can provide you with some special cholesterol-lowering benefits if you will cook it by steaming. The fiber-related components in **broccoli** do a better job ...

News for broccoli

Mistakes We All Make With Spaghetti, Steak And ...

Huffington Post - 2 days ago

But in her new book *Brassicas: Cooking the World's Healthiest Vegetables*, she says plunking **broccoli**, cauliflower or Brussels sprouts into ...



Two Sides to Ranking: Importance



Google

obama

Web Images News Videos More Search tools

About 48,100,000 results (0.26 seconds)

Mount Obama - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Mount_Obama

Mount Obama (known as **Boggy Peak** until August 4, 2009) is the highest point in the nation of Antigua and Barbuda and on the island of Antigua. It lies in the far ...

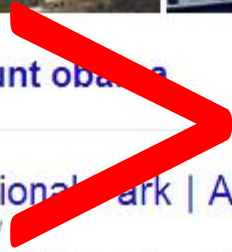
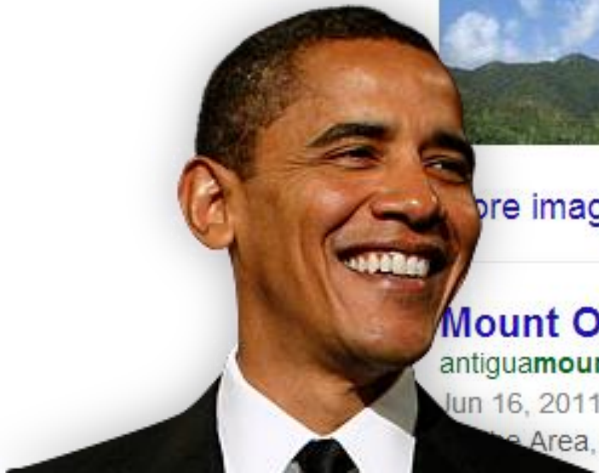
Images for mount obama Report images



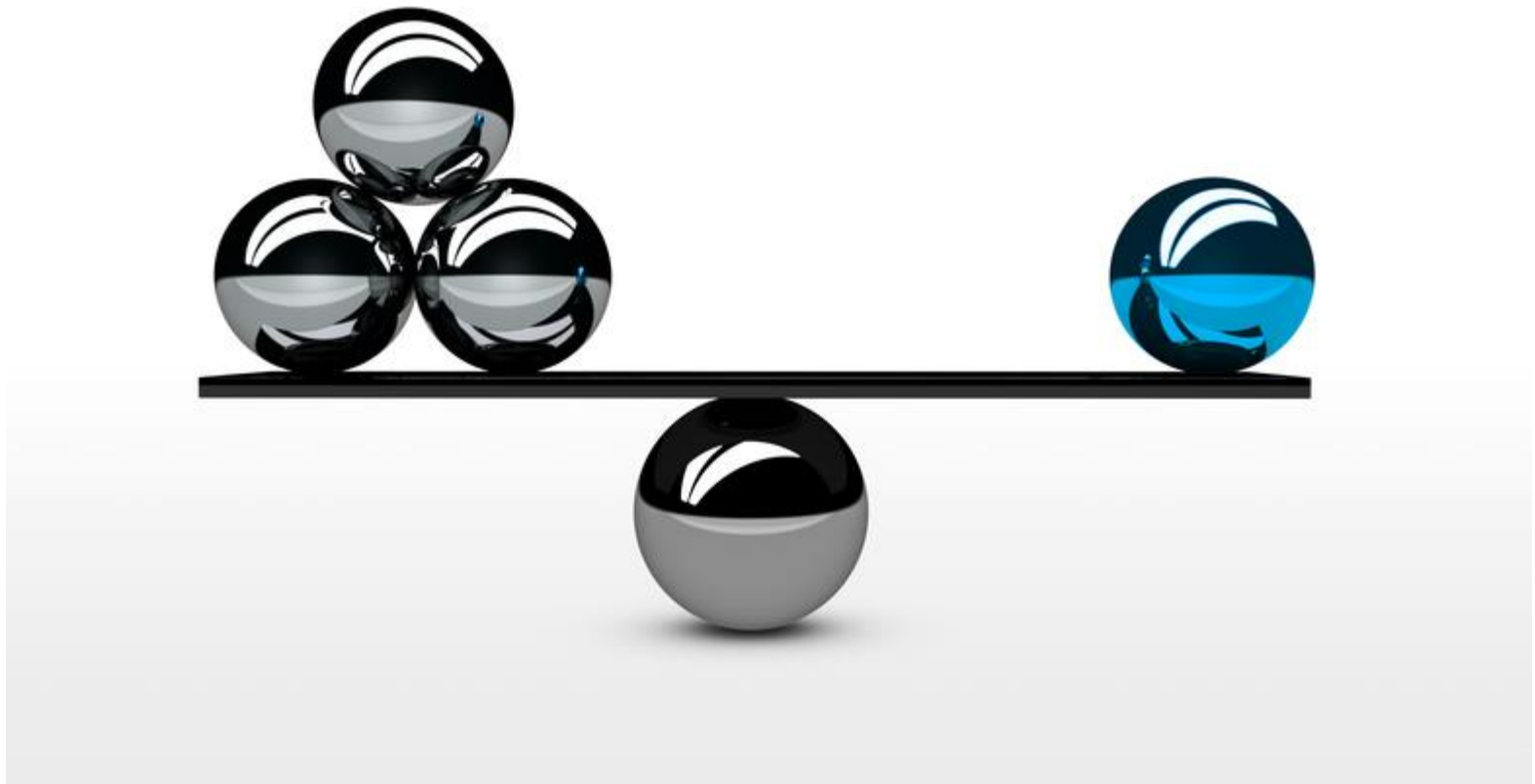
More images for mount obama

Mount Obama National Park | Antigua and Barbuda
antiguamountobama.com/

Jun 16, 2011 - As the **Mount Obama** Committee continues its work in the Mount Obama National Park Area, the committee organized a site visit to the O...



Relevance vs. Importance: A Balancing Act



RANKING:

RELEVANCE

Example Query

Which of these three keyword terms is most “important”?




Google

movie freedom wallace

Web Images News Videos More Search tools

About 4,290,000 results (0.29 seconds)

[Braveheart In Defiance Of The English Tyranny! BRAVO ...](#)

 www.youtube.com/watch?v=WLrrBs8JBQo
Feb 25, 2008 - Uploaded by popthetime
... actor starring as the "William **Wallace**" character in the **movie** - B...
... Braveheart **Freedom** Speech (HD) by ...

[Braveheart - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/Braveheart
Braveheart is a 1995 epic historical drama war **film** directed by and starring Mel Gibson. Gibson portrays ... **Wallace** instead shouts the word "**Freedom!**" and the ...

[Braveheart \(1995\) - Quotes - IMDb](#)
www.imdb.com/title/tt0112573/quotes
... (1995) Quotes on IMDb: Memorable quotes and exchanges from **movies**, TV series and more... ... William **Wallace**: It's all for nothing if you don't have **freedom**.

Matches in a Document

The image shows a browser window displaying the Wikipedia page for 'Braveheart'. The browser's address bar shows the URL 'https://en.wikipedia.org/wiki/Braveheart'. The page title is 'Braveheart' and the subtitle is 'From Wikipedia, the free encyclopedia'. The main content area contains a paragraph about the film, mentioning Mel Gibson and William Wallace. A search bar in the top right corner contains the text 'freedom' and shows '1 de 7' results. A red box highlights the search bar and the 'freedom' text. Another red box highlights the search results in the bottom left corner, showing 'freedom' and '7 occurrences'. The page also features a sidebar with navigation links and a movie poster for 'Braveheart' on the right.

W Braveheart - Wikipedia x

Es seguro | https://en.wikipedia.org/wiki/Braveheart

Not log freedom 1 de 7

Article Talk Read Edit View history Search Wikipedia

Braveheart

From Wikipedia, the free encyclopedia

For other uses, see [Braveheart \(disambiguation\)](#).

Braveheart is a 1995 American [epic war film](#) directed by and starring [Mel Gibson](#). Gibson portrays [William Wallace](#), a 13th-century Scottish warrior who led the Scots in the First War of Scottish Independence against King Edward I of England. The story is inspired by [Blind Harry's epic poem *The Actes and Deidis of the Illustre and Vallyeant Campioun Schir William Wallace*](#) and was adapted for the screen by [Randall Wallace](#).

The film was nominated for ten [Academy Awards](#) at the 68th Academy Awards and won five: [Best Picture](#), [Best Director](#), [Best Cinematography](#), [Best Makeup](#), and [Best Sound Editing](#).

Contents [hide]

1 Plot

2 Cast

3 Production

freedom

- 7 occurrences

Upload file

Braveheart

MEL · GIBSON

Every man dies,
not every man
really lives.

BRAVEHEART

Matches in a Document

The image shows a screenshot of a web browser displaying the Wikipedia page for 'Braveheart'. The browser's address bar shows the URL 'https://en.wikipedia.org/wiki/Braveheart'. The page title is 'Braveheart' and the subtitle is 'From Wikipedia, the free encyclopedia'. The main content area contains a paragraph about the film, mentioning it is a 1995 American epic war film directed by and starring Mel Gibson. A search bar in the top right corner shows the search term 'movie' and the result count '3 de 16'. Two red boxes highlight search results for the term 'freedom', showing '7 occurrences'. Two orange boxes highlight search results for the term 'movie', showing '16 occurrences'. The movie poster for 'Braveheart' is visible on the right side of the page.

W Braveheart - Wikipedia x

Es seguro | https://en.wikipedia.org/wiki/Braveheart

Not log movie 3 de 16

Article Talk Read Edit View history Search Wikipedia

Braveheart

From Wikipedia, the free encyclopedia

For other uses, see [Braveheart \(disambiguation\)](#).

Braveheart is a 1995 American [epic war film](#) directed by and starring [Mel Gibson](#). Gibson portrays [William Wallace](#), a 13th-century Scottish warrior who led the Scots in the First War of Scottish Independence against King Edward I of England. The story is inspired by [Blind Harry's epic poem *The Actes and Deidis of the Illustre and Vallyeant Campioun Schir William Wallace*](#) and was adapted for the screen by [Randall Wallace](#).

The film was nominated for ten [Academy Awards](#) at the 68th Academy Awards and won five: [Best Picture](#), [Best Director](#), [Best Cinematography](#), [Best Makeup](#), and [Best Sound Editing](#).

Contents [hide]

1 Plot

2 Cast

3 Production

Upload file

freedom

- 7 occurrences

movie

- 16 occurrences

Braveheart

MEL · GIBSON

Every man dies,
not every man
really lives.

Braveheart

Matches in a Document

The image shows a screenshot of a web browser displaying the Wikipedia page for 'Braveheart'. The browser's address bar shows the URL 'https://en.wikipedia.org/wiki/Braveheart'. A search bar in the top right corner contains the text 'wallace' and shows '44 de 88' results. The main content of the page is the article for 'Braveheart', which includes a description of the 1995 film and a movie poster. Three search results are highlighted with colored boxes: a red box for 'freedom' (7 occurrences), an orange box for 'movie' (16 occurrences), and a green box for 'wallace' (88 occurrences). The 'wallace' box is also highlighted in the search bar at the top right.

W Braveheart - Wikipedia x

Es seguro | https://en.wikipedia.org/wiki/Braveheart

Not log wallace 44 de 88

Article Talk Read Edit View history Search Wikipedia

Braveheart

From Wikipedia, the free encyclopedia

For other uses, see [Braveheart \(disambiguation\)](#).

Braveheart is a 1995 American epic war film directed by and starring Mel Gibson. Gibson portrays William Wallace, a 13th-century Scottish warrior who led the Scots in the First War of Scottish Independence against King Edward I of England. The story is inspired by Blind Harry's epic poem *The Actes and Deidis of the Illustre and Vallyeant Campioun Schir William Wallace* and was adapted for the screen by Randall Wallace.

The film was nominated for ten Academy Awards at the 68th Academy Awards and won five: Best Picture, Best Director, Best Cinematography, Best Makeup, and Best Sound Editing.

Contents [hide]

1 Plot

2 Cast

3 Production

Upload file

freedom

- 7 occurrences

movie

- 16 occurrences

wallace

- 88 occurrences

Braveheart

MEL GIBSON

Every man dies, not every man really lives.

Usefulness of Words

Google

Google

Web Images Videos News More ▾ Search tools

About 835,000,000 results (0.34 seconds)

movie

- occurs very frequently

Google

Web Images Videos Books More ▾ Search tools

About 198,000,000 results (0.32 seconds)

freedom

- occurs frequently

Google

Web Images Books News More ▾ Search tools

About 49,200,000 results (0.31 seconds)

wallace

- occurs occasionally

Estimating Relevance

- Rare words more important than common words
 - **wallace** (49M) more important than **freedom** (198M)
more important than **movie** (835M)
- Words occurring more frequently in a document indicate higher relevance
 - **wallace** (88) more matches than **movie** (16) more matches than **freedom** (7)

Relevance Measure: TF-IDF

- TF: Term Frequency

- Measures occurrences of a term in a document

- $tf(t, d)$... various options

- Raw count of occurrences

$$tf(t, d) = \text{count}(t, d)$$

- Logarithmically scaled

$$tf(t, d) = \log(\text{count}(t, d) + 1)$$

- Normalised by document length

$$tf(t, d) = \frac{\text{count}(t, d)}{\sum_{t' \in d} \text{count}(t', d)}$$

$$tf(t, d) = \frac{\text{count}(t, d)}{\max_{t' \in d} \text{count}(t', d)}$$

- A combination / something else 😊

Relevance Measure: TF-IDF

- **IDF: Inverse Document Frequency**
 - How common a term is across **all** documents
 - $\text{idf}(t, D)$...
 - Logarithmically scaled document occurrences

$$\text{idf}(t, D) = \log\left(\frac{|D|+1}{|\{d \in D : t \in d\}|+1}\right)$$

- Note: The more rare, the larger the value

Relevance Measure: TF-IDF

- **TF-IDF**: Combine Term Frequency and Inverse Document Frequency:

$$\text{tf-idf}(t, d) = \text{tf}(t, d) \times \text{idf}(t, D)$$

- Score for a query
 - Let query $q = (t_1, \dots, t_n)$
 - Score for a query: $\text{score}(q, d) = \sum_{t \in q} \text{tf-idf}(t, d)$(There are other possibilities)

Relevance Measure: TF-IDF



Term Frequency

$$\text{tf}(t, d) = \text{count}(t, d)$$

Inverse Document Frequency

$$\text{idf}(t, D) = \log_2 \left(\frac{|D|}{|\{d \in D : t \in d\}| + 1} \right)$$

$$\text{tf-idf}(t, d) = \text{tf}(t, d) \times \text{idf}(t, D)$$

t	$\text{tf}(t, d)$
movie	16
freedom	7
wallace	43

Relevance Measure: TF-IDF



Term Frequency

$$\text{tf}(t, d) = \text{count}(t, d)$$

Inverse Document Frequency

$$\text{idf}(t, D) = \log_2 \left(\frac{|D|}{|\{d \in D : t \in d\}| + 1} \right)$$

$$\text{tf-idf}(t, d) = \text{tf}(t, d) \times \text{idf}(t, D)$$

t	$\text{tf}(t, d)$	$ \{d \in D : t \in d\} $
movie	16	835,000,000
freedom	7	198,000,000
wallace	43	49,200,000

Relevance Measure: TF-IDF



Term Frequency

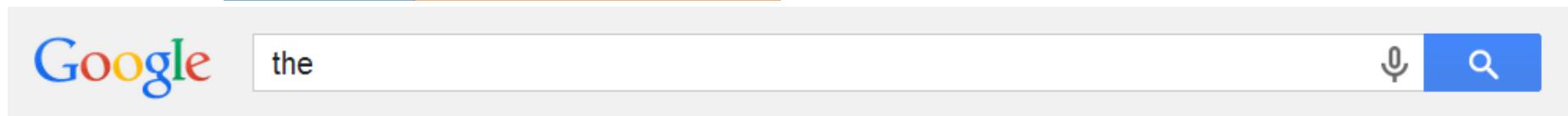
$$tf(t, d) = \text{count}(t, d)$$

Inverse Document Frequency

$$idf(t, D) = \log_2 \left(\frac{|D|}{|\{d \in D : t \in d\}| + 1} \right)$$

$$tf-idf(t, d) = tf(t, d) \times idf(t, D)$$

t	$tf(t, d)$	$ \{d \in D : t \in d\} $	$\frac{ D }{ \{d \in D : t \in d\} + 1}$
movie	16	835,000,000	
freedom	7	198,000,000	
wallace	43	49,200,000	



About 11,410,000,000 results (0.27 seconds)

$$|D| = 11,410,000,000$$

Relevance Measure: TF-IDF



Term Frequency

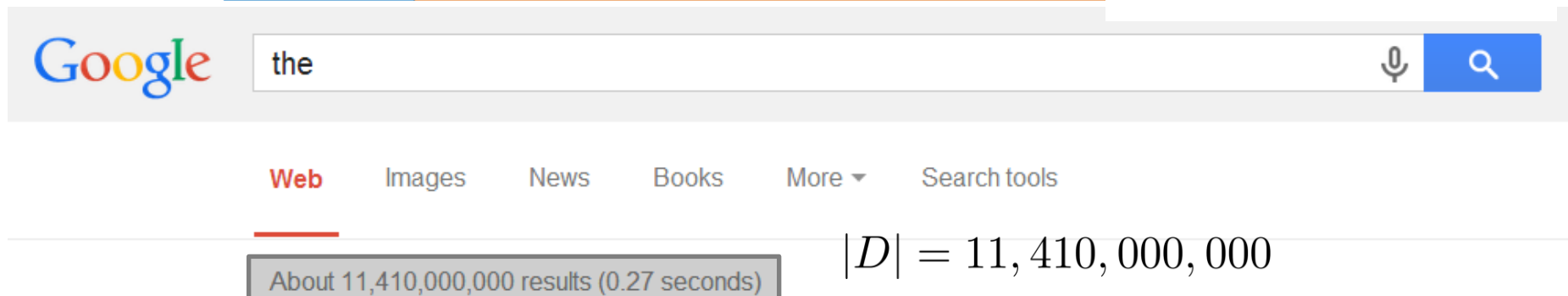
$$\text{tf}(t, d) = \text{count}(t, d)$$

Inverse Document Frequency

$$\text{idf}(t, D) = \log_2 \left(\frac{|D|}{|\{d \in D : t \in d\}| + 1} \right)$$

$$\text{tf-idf}(t, d) = \text{tf}(t, d) \times \text{idf}(t, D)$$

t	$\text{tf}(t, d)$	$ \{d \in D : t \in d\} $	$\frac{ D }{ \{d \in D : t \in d\} + 1}$
movie	16	835,000,000	13.66
freedom	7	198,000,000	57.62
wallace	43	49,200,000	231.91



Relevance Measure: TF-IDF



Term Frequency

$$\text{tf}(t, d) = \text{count}(t, d)$$

Inverse Document Frequency

$$\text{idf}(t, D) = \log_2 \left(\frac{|D|}{|\{d \in D : t \in d\}| + 1} \right)$$

$$\text{tf-idf}(t, d) = \text{tf}(t, d) \times \text{idf}(t, D)$$

t	$\text{tf}(t, d)$	$ \{d \in D : t \in d\} $	$\frac{ D }{ \{d \in D : t \in d\} + 1}$	$\text{idf}(t, D)$
movie	16	835,000,000	13.66	3.77
freedom	7	198,000,000	57.62	5.84
wallace	43	49,200,000	231.91	7.85

Relevance Measure: TF-IDF



Term Frequency

$$\text{tf}(t, d) = \text{count}(t, d)$$

Inverse Document Frequency

$$\text{idf}(t, D) = \log_2 \left(\frac{|D|}{|\{d \in D : t \in d\}| + 1} \right)$$

$$\text{tf-idf}(t, d) = \text{tf}(t, d) \times \text{idf}(t, D)$$

t	$\text{tf}(t, d)$	$ \{d \in D : t \in d\} $	$\frac{ D }{ \{d \in D : t \in d\} + 1}$	$\text{idf}(t, D)$	$\text{tf-idf}(t, d)$
movie	16	835,000,000	13.66	3.77	60.36
freedom	7	198,000,000	57.62	5.84	40.94
wallace	43	49,200,000	231.91	7.85	337.87

Relevance Measure: TF-IDF



Term Frequency

$$\text{tf}(t, d) = \text{count}(t, d)$$

Inverse Document Frequency

$$\text{idf}(t, D) = \log_2 \left(\frac{|D|}{|\{d \in D : t \in d\}| + 1} \right)$$

$$\text{tf-idf}(t, d) = \text{tf}(t, d) \times \text{idf}(t, D)$$

t	$\text{tf}(t, d)$	$ \{d \in D : t \in d\} $	$\frac{ D }{ \{d \in D : t \in d\} + 1}$	$\text{idf}(t, D)$	$\text{tf-idf}(t, d)$
movie	16	835,000,000	13.66	3.77	60.36
freedom	7	198,000,000	57.62	5.84	40.94
wallace	43	49,200,000	231.91	7.85	337.87

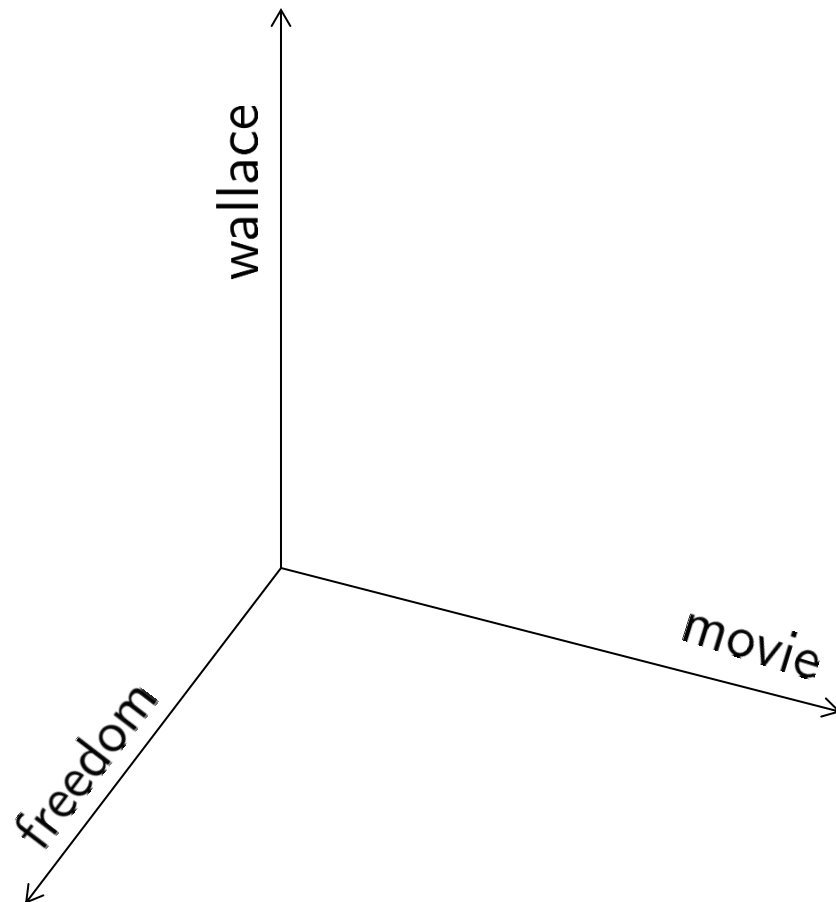
$$\text{score}(q, d) = \sum_{t \in q} \text{tf-idf}(t, d)$$

$$\text{score}((\text{movie, freedom, wallace}), \text{http://en.wikipedia.org/Braveheart}) \approx 439.17$$

Vector Space Model (a mention)

t	$\text{tf}(t, d)$
movie	16
freedom	7
wallace	43

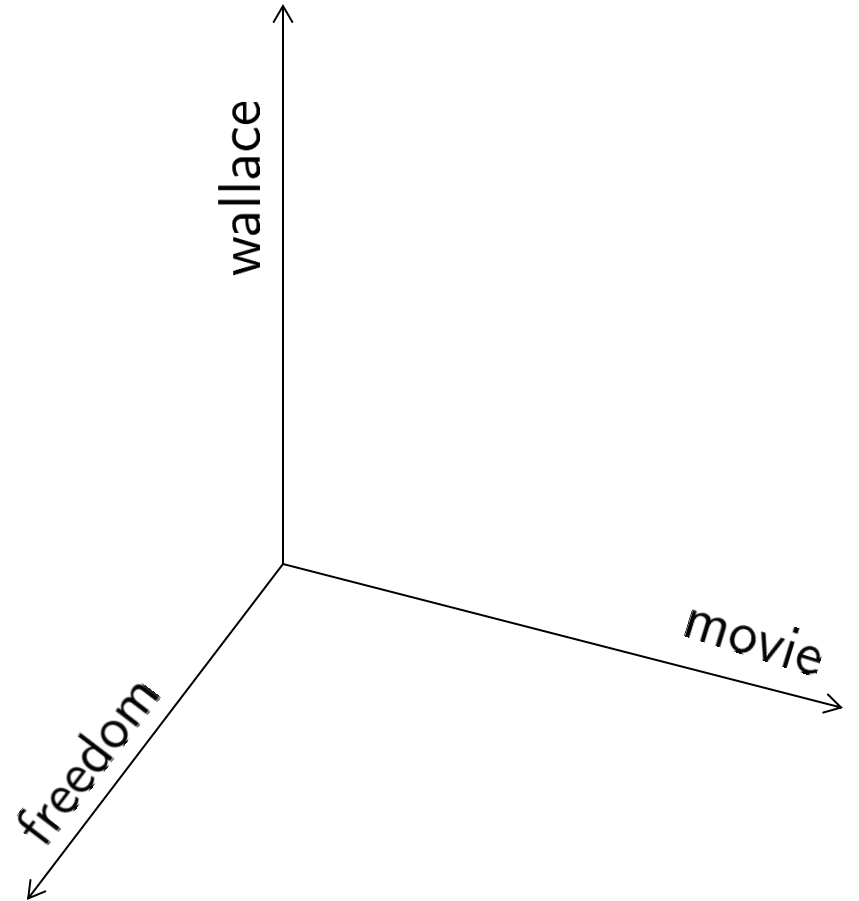
$$l = \sqrt{\sum_{t \in q} \text{tf}(t, d)^2}$$



Vector Space Model (a mention)

t	$\text{tf}(t, d)$	$\text{tf}(t, d)^2$
movie	16	256
freedom	7	49
wallace	43	1,894

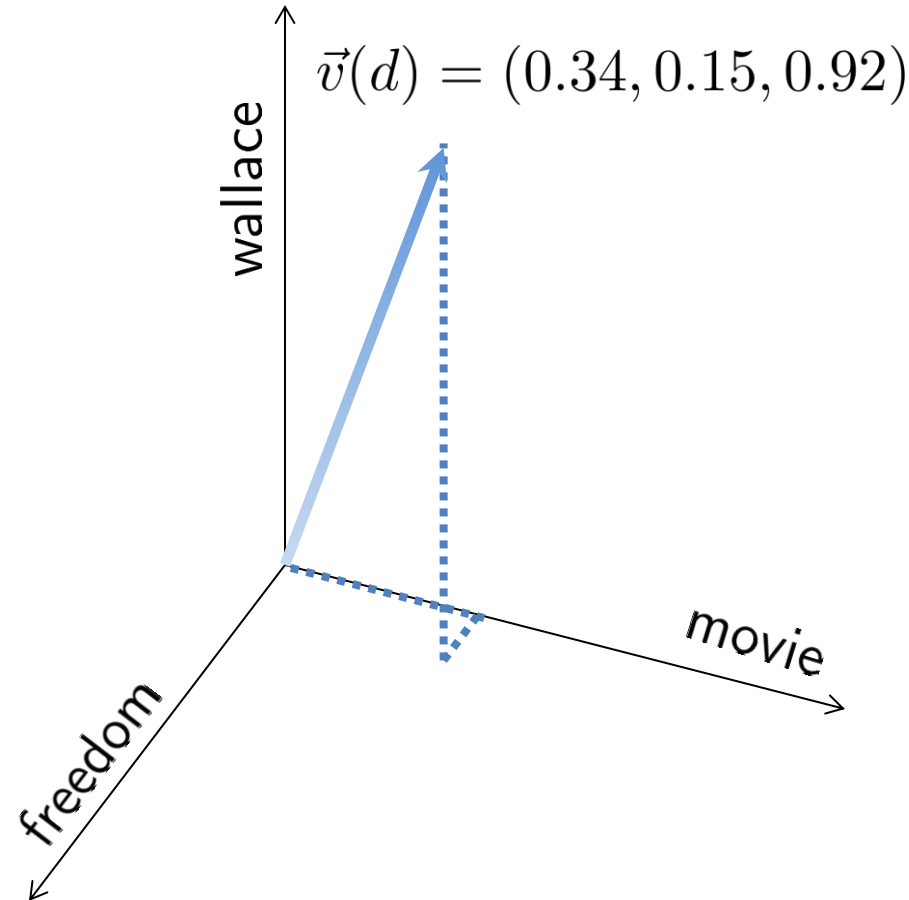
$$l = \sqrt{\sum_{t \in q} \text{tf}(t, d)^2}$$



Vector Space Model (a mention)

t	$\text{tf}(t, d)$	$\text{tf}(t, d)^2$	$\frac{\text{tf}(t, d)}{l}$
movie	16	256	0.34
freedom	7	49	0.15
wallace	43	1,894	0.92

$$l = \sqrt{\sum_{t \in q} \text{tf}(t, d)^2}$$



Dividing by l normalises the length of vector to 1

Vector Space Model (a mention)

- Cosine Similarity

$$\text{sim}(d, d') = \vec{v}(d) \cdot \vec{v}(d')$$

t	$\vec{v}(d)$	$\vec{v}(d')$	\times
movie	0.34	0.49	0.17
freedom	0.15	0.82	0.12
wallace	0.93	0.30	0.28

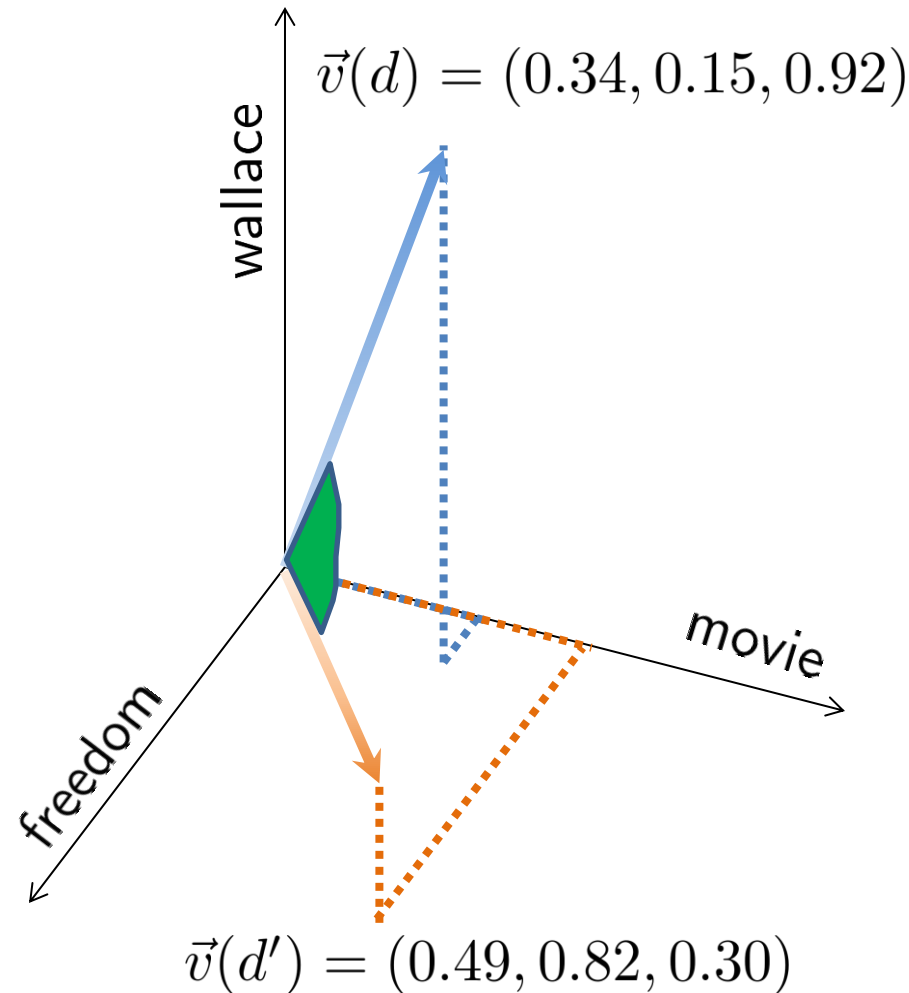
$$\text{sim}(d, d') \approx 0.57$$

Σ

- Note:

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}| |\mathbf{b}| \cos(\angle(\mathbf{a}, \mathbf{b}))$$

$$|\vec{v}(d)| = |\vec{v}(d')| = 1$$



Hence the similarity is the cosine of the **angle** between the vectors

Relevance Measure: TF-IDF

- **TF-IDF**: Combine Term Frequency and Inverse Document Frequency:

$$\text{tf-idf}(t, d) = \text{tf}(t, d) \times \text{idf}(t, D)$$

- Score for a query

- Let query $q = (t_1, \dots, t_n)$

- Score for a query: $\text{score}(q, d) = \sum_{t \in q} \text{tf-idf}(t, d)$

(There are other possibilities)

... we could also use cosine similarity between query and document using **TF-IDF** weights

Two Sides to Ranking: Relevance



The image shows a Google search interface with the search term 'obama' entered. The search results are filtered to the 'Web' tab. The first result is a Wikipedia entry for 'Broccoli', followed by an article from 'The World's Healthiest Foods' about the health benefits of broccoli. Below these, there are two news snippets: one titled 'News for broccoli' and another titled 'Mistakes We All Make With Spaghetti, Steak And E...'. A large red 'X' is drawn over the second news snippet. In the bottom left corner, there is a cutout image of Barack Obama smiling, and in the bottom right corner, there is a cutout image of a head of broccoli.

Google

Web Images News Videos More ▾ Search tools

About 16,700,000 results (0.23 seconds)

Broccoli - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Broccoli** ▾
Broccoli is an edible green plant in the cabbage family, whose large flowering head is used as a vegetable. The word **broccoli** comes from the Italian plural of ...
Cauliflower - Romanesco broccoli - Broccoli (disambiguation) - Brocolini

Broccoli - The World's Healthiest Foods
www.whfoods.com/genpage.php?tname=foodspice&dbid=9 ▾
Broccoli can provide you with some special cholesterol-lowering benefits if you will cook it by steaming. The fiber-related components in **broccoli** do a better job ...

News for broccoli

Mistakes We All Make With Spaghetti, Steak And E...
Huffington Post - 2 days ago
But in her new book Brassicas: Cooking the World's Healthiest Vegetables, she says plunking **broccoli**, cauliflower or Brussels sprouts into ...



Field-Based Boosting

- Not all text is equal: titles, headers, etc.

```
<!DOCTYPE html>
<html lang="en" dir="ltr" class="client-nojs">
<head>
<meta charset="UTF-8" />
<title>Barack Obama - Wikipedia, the free encyclopedia</title>
```



The screenshot displays the Wikipedia article for Barack Obama. At the top, the HTML source code is visible, with the title tag `<title>Barack Obama - Wikipedia, the free encyclopedia</title>` highlighted in blue. Below the code, the article's main content is shown. The title "Barack Obama" is highlighted in orange. The article text begins with a redirect notice: *"Obama" redirects here. For other uses, see Obama (disambiguation).* This is followed by a note: *This article is about the 44th president of the United States. For his father, see Barack Obama, Sr.* The main body of text starts with: **Barack Hussein Obama II** (/bəˈrɑːk huːˈseɪn ouˈbɑːmə/[ⓘ]; born August 4, 1961) is the 44th and current President of the United States, and the first African American to hold the office. Born in Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he served as president of the *Harvard Law Review*. He was a community organizer in Chicago before earning his law degree. He worked as a civil rights attorney and taught constitutional law at the University of Chicago Law School.

On the right side of the article, there is a portrait of Barack Obama with the caption "Barack Obama".

On the left side, the Wikipedia logo and navigation menu are visible, including links for Main page, Contents, Featured content, Current events, Random article, Donate to Wikipedia, and Wikimedia Shop.

Anchor Text

- See how the Web views/tags a page

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html>
<head>
  <title>What I watched last night ...</title>
</head>
<body>
<p>Last night I was pretty bored so I made some popcorn and watched
<a href="http://en.wikipedia.org/Braveheart">a movie about William Wallace called Braveheart</a>.
Set in Scotland it has lots of blood and gore.
</p>
</body>
</html>
```

Anchor Text

- See how the Web views/tags a page

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
"http://www.w3.org/TR/
<html>
<head>
  <title>What I watched
</head>
<body>
<p>Last night I was pret
<a href="http://en.wiki
Set in Scotland it has
</p>
</body>
</html>
```

Google da da da dum symphony

Web Videos News Shopping Images More Search tools

About 107,000 results (0.36 seconds)



Beethoven - Symphony No. 5 in C Minor (1) - YouTube
www.youtube.com/watch?v=W2qW6fOtAMY

Lucene uses relevance scoring

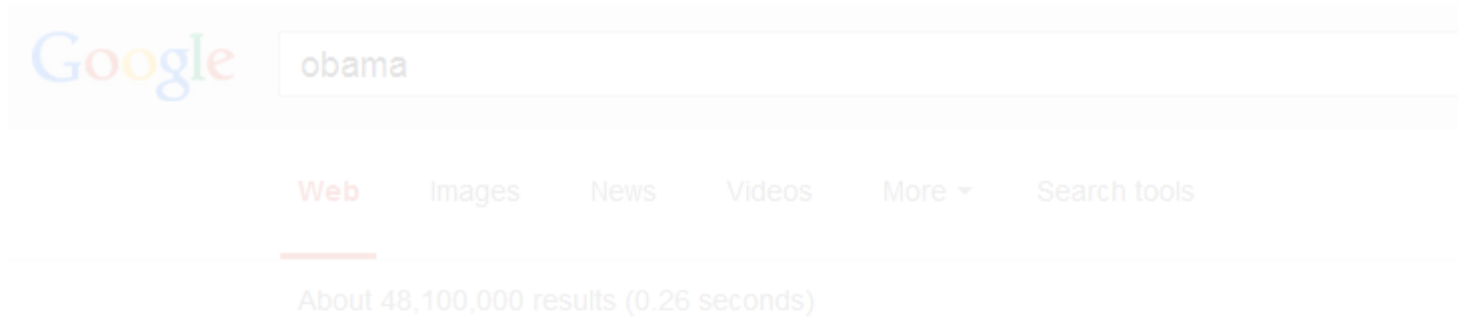


```
Tasks Console
SearchWikiIndex [Java Application] C:\Program Files\Java\jre1.8.0_65\bin\javaw.exe (03-05-2017 12:45:22 a. m.)
Opening directory at lucene
Enter a keyword search phrase:
obama
Running query: obama
Parsed query: TITLE:obam^5.0 ABSTRACT:obam
Matching documents: 255
Showing top 10 results
1 http://es.wikipedia.org/wiki/Obama_Republican Obama Republican
2 http://es.wikipedia.org/wiki/Obama_(Fukui) Obama (Fukui)
3 http://es.wikipedia.org/wiki/Republicanos_por_Obama Republicanos por Obama
4 http://es.wikipedia.org/wiki/Engonga_Obame Engonga Obame
5 http://es.wikipedia.org/wiki/Barack_Obama Barack Obama
6 http://es.wikipedia.org/wiki/Michelle_Obama Michelle Obama
7 http://es.wikipedia.org/wiki/Cartel_%22Hope%22_de_Obama Cartel "Hope" de Obama
8 http://es.wikipedia.org/wiki/Transici3n_presidencial_de_Barack_Obama Transici3n presidencial de Barack Obama
9 http://es.wikipedia.org/wiki/Por_qu3_Obama_ganar3_en_2008_y_en_2012 Por qu3 Obama ganar3 en 2008 y en 2012
10 http://es.wikipedia.org/wiki/Ricardo_Mangue_Obama_Nfubea Ricardo Mangue Obama Nfubea
```

RANKING:

IMPORTANCE

Two Sides to Ranking: Importance

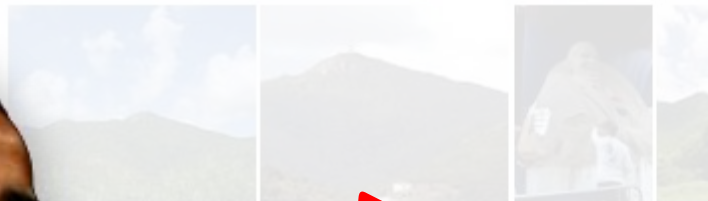


How could we determine that Barack Obama is more important than Mount Obama as a search result for "obama" on the Web?



Images for mount obama

Report images



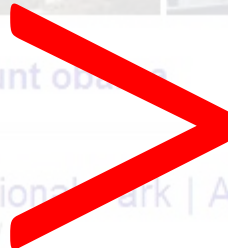
More images for mount obama

Mount Obama National Park | Antigua a

antiguamountobama.com/

Jun 16, 2011 - As the Mount Obama Committee continu

Area, the committee organized a site visit to the C



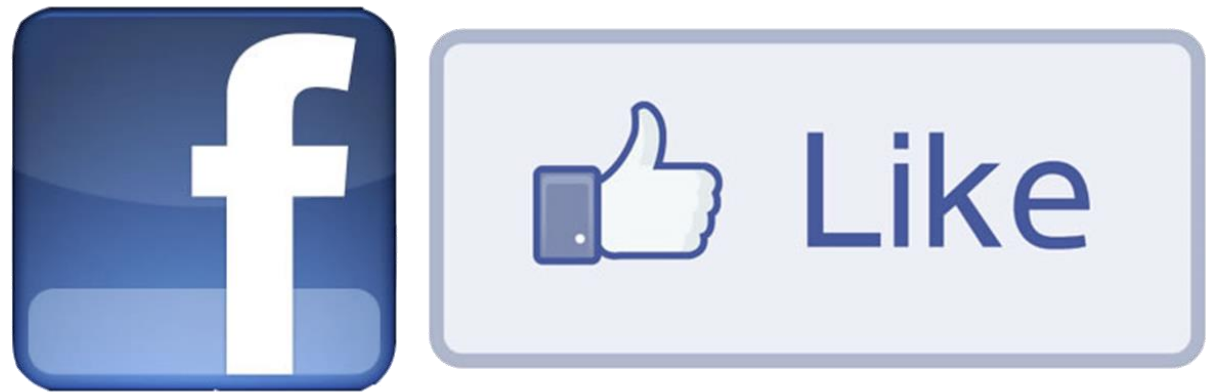
Link Analysis

Which will have more links from other pages?
The Wikipedia article for Mount Obama?
The Wikipedia article for Barack Obama?



Link Analysis

- Consider links as votes of confidence in a page
- A hyperlink is the open Web's version of ...



(... even if the page is linked in a negative way.)

Link Analysis

So if we just count links to a page we can determine its importance and we are done?



Link Spamming



semanticweb.com™

The Voice of Semantic Technology Business
Big Data, Linked Data, Smart Data

Home

Events

Media

Industry Verticals

Answers

Jobs

Questions

Tags

Users

Badges

[Claritin](#) [Clomid](#) [Combivent](#) [Confido](#) [Copegus](#) [Cordarone](#) [Coreg](#) [Coumadin](#) [Cozaar](#) [Crestor](#) [Cyklokapron](#) [Cymbalta](#) [Cystone](#) [Cytotec](#) [Danazol](#) [Deltason](#) [Depakote](#) [Desyrel](#) [Detrol](#) [Diabecor](#) [Diakof](#) [Diarex](#) [Didronel](#) [Differin](#) [Dilantin](#) [Diovan](#) [Dostinex](#) [Elavil](#) [Elimite](#) [Emsam](#) [Endep](#) [Eurax](#) [Evecare](#) [Evista](#) [Exelon](#) [Famvir](#) [Feldene](#) [Femara](#) [Femcare](#) [Flomax](#) [Flonase](#) [Flovent](#) [Fosamax](#) [Gasex](#) [Geodon](#) [Geriforte](#) [Herbolax](#) [High Love](#) [Himcocid](#) [Himcolin](#) [Himcospaz](#) [Himplasia](#) [Hoodia](#) [Hytrin](#) [Hyzaar](#) [Imdur](#) [Imitrex](#) [Inderal](#) [Ismo](#) [Isoptin](#) [Isordil](#) [Kamagra](#) [Karela](#) [Keftab](#) [Koflet](#) [Kytril](#) [Lamictal](#) [Lamisil](#) [Lanoxin](#) [Lariam](#) [Lasix](#) [Lasuna](#) [Leukeran](#) [Levaquin](#) [Levlen](#) [Levothroid](#) [Lincocin](#) [Lioresal](#) [Lisinopril](#) [Liv.52](#) [Lopid](#) [Lopressor](#) [Loprox](#) [Lotensin](#) [Lotrisone](#) [Loxitane](#) [Lozol](#) [Lukol](#) [Lynoral](#) [Maxaquin](#) [Menosan](#) [Mentat](#) [Mentax](#) [Mevacor](#) [Mexitil](#) [Miacalcin](#) [Micardis](#) [Mobic](#) [Monoket](#) [Motrin](#) [Myambutol](#) [Mycelex-G](#) [Mysoline](#) [Naprosyn](#) [Neurontin](#) [Nicotinell](#) [Nimotop](#) [Nirdosh](#) [Nizoral](#) [Nolvadex](#) [Nonoxinol](#) [Noroxin](#) [Omnicef](#) [Ophthalmicare](#) [Oxytrol](#) [Pamelor](#) [Parlodel](#) [Paxil](#) [Penisole](#) [Phentermine](#) [Pilex](#) [Plan B](#) [Plavix](#) [Plendil](#) [Pletal](#) [Prandin](#) [Pravachol](#) [Prednisone](#) [Prenarmin](#) [Prevacid](#) [Prilosec](#) [Prinivil](#) [Procardia](#) [Prograf](#) [Prometrium](#) [Propecia](#) [Proscar](#) [Protonix](#) [Proventil](#) [Prozac](#) [Purin](#) [Purinethol](#) [Quibron-T](#) [Relafen](#) [Renalka](#) [Reosto](#) [Requip](#) [Retin-A](#) [Revvia](#) [Rhinocort](#) [Rimonabant](#) [Risperdal](#) [Rocaltrol](#) [Rogaine](#) [Rumalaya](#) [Sarafem](#) [Septilin](#) [Serevent](#) [Serophepe](#) [Seroquel](#) [Shallaki](#) [Shoot](#) [Sinequan](#) [Singular](#) [Snoroff](#) [Sorbitrate](#) [Speman](#) [Starlix](#) [StretchNil](#) [Stromectol](#) [Styplon](#) [Sumycin](#) [Superman](#) [Sustiva](#) [Synthroid](#) [Tenormin](#) [Topamax](#) [Trandate](#) [Tricor](#) [Trimox](#) [Triphala](#) [Tulasi](#) [Urispas](#) [V-Gel](#) [Vantin](#) [Vasodilan](#) [Vasotec](#) [Ventolin](#) [Viramune](#) [Vytorin](#) [Xeloda](#) [Xenacore](#) [Zanaflex](#) [Zantac](#) [Zebeta](#) [Zelnorm](#) [Zerit](#) [Yerba Diet](#) [Wellbutrin SR](#) [Women Attracting Pheromones](#) [Women's Intimacy Enhancer](#) [Women's Intimacy Enhancer Cream](#) [Virility Gum](#) [Vitamin A & D](#) [Viagra + Cialis](#) [Viagra + Cialis + Levitra](#) [Viagra Jelly](#) [Viagra Soft + Cialis Soft](#) [Viagra Soft Tabs](#) [Ultimate Male Enhancer](#) [Toprol XL](#) [Touch-Up Kit](#) [Tentex Royal](#) [Tentex Forte](#) [Tiberius Erectus](#) [Zero Nicotine 2 Complete Professional Whitening Kits 2 Sets Of Moldable Mouth Trays 36 Beauty Acne-n-Pimple Cream ActoPlus Met Superloss Multi SleepWell \(Herbal XANAX\) Shuddha Guggulu Rythmol SR Rumalaya Forte Pulmicort Inhaler Professional Plasma Tooth Whitening Kit Premium Diet Patch Penis Growth Oil Penis Growth Pack Penis Growth Patch Penis Growth Pills Orgasm Enhancer Norpace CR Mental Booster Men Attracting Pheromones Menopause Gum Male Enhancement Oil Male Enhancement Patch Male Enhancement Pills Male Sexual Tonic InnoPran XL Hoodia Weight Loss Gum Hoodia Weight Loss Patch Human Growth Hormone Agent Glucotrol XL Green Tea Grifulvin V Gyne-Lotrimin Hair Loss Cream Herbal Maxx Herbal Phentermine Flagyl ER Female Sexual Tonic Female Viagra Epivir-HBV Diet Maxx Deluxe Handheld Plasma Whitening Tool Deluxe Whitening System With Plasma Maxx Coral Calcium Cialis Jelly Cialis Soft Tabs Calcium Carbonate Bust Enhancer Beconase AQ Anatriam Diet Pills Advair Diskus Advanced Gain Pro Breast Augmentation Breast Enhancement Breast Enhancement Gel Breast Enhancement Gum Breast Intense Buy Trazodone Buy Celebrex Buy Alprazolam Buy Tramadol Buy Fioricet Buy Soma Buy Cialis Buy Carisoprodol Buy Levitra Buy Ultram Buy Ambien Buy Viagra Buy Xanax Buy Phentermine Buy Valium Buy Diazepam Generic Celebrex Generic Alprazolam Generic Tramadol Generic Fioricet Generic Soma Generic Cialis Generic Carisoprodol Generic Levitra Generic Ultram Generic Ambien](#)

[deleted] Kala Jadu Specialist +91961



black magic specialist baba ji call now +919610897260



<http://www.blackmagicspecialist.net.in>



java

edit | close | undelete | more ▼

Link Importance

So which should count for more?

A link from http://en.wikipedia.org/wiki/Barack_Obama?

Or a link from <http://blackmagicspecialist.net.in>?



Link Importance

Maybe we could consider links from some domains as having more “vote”?



The screenshot shows the Digg website interface. At the top, there is a navigation bar with the Digg logo and links for 'My News', 'Top News', and 'Upcoming'. A search bar on the right says 'Submit a story to Digg...' with a 'GO' button. Below this is a secondary navigation bar with categories: 'All Topics', 'Business', 'Entertainment', 'Gaming', 'Lifestyle' (which is highlighted), 'Offbeat', 'Politics', 'Science', and 'Sports'. The main content area features a link from 'rxpills.host-sc.com' with a Digg score of 385. The link title is 'RxPILLS - Best discounts for all Pharmacy! Check out or men's health category!'. The link text is a long list of keywords related to Viagra, such as 'viagra toll free number', 'viagra tonytigeraz', 'viagra top ten', 'viagra toronto', 'viagra torrent', 'viagra total knee', 'viagra toung', 'viagra travel', 'viagra treat childhood pulmonary hypertension', 'viagra treatment', 'viagra treatment for crohn's', 'viagra treatment for pe', 'viagra treatment for women', 'viagra treatment hape', 'viagra treatment impotence', 'viagra treats children s lethal hypertension', 'viagra trh pharmacy', 'viagra trial', 'viagra trial coupon', 'viagra trial pack', 'viagra trial pack canada', 'viagra trial packs', 'viagra trial pak', 'viagra trial sample', 'viagra trials for shrinking cancerous tumors', 'viagra triangle', 'viagra triangle chicago', 'viagra triangle chicago illinois', 'viagra triangle cleveland', 'viagra triangle cleveland ohio', 'viagra tricks', 'viagra trip in thailand', 'viagra trivia', 'viagra tune', 'viagra tupperware', 'viagra tupperware pool', 'viagra tutorial', 'viagra tv ad', 'viagra tv ad 1999', 'viagra tv advertising', 'viagra tv commercial', 'viagra tv commercial girl', 'viagra tv commercial viva las vegas', 'viagra tv commercials', 'viagra tv girl', 'viagra type drugs', 'viagra type medications', 'viagra type products', 'viagra types', 'viagra u', 'viagra u s pharmacies', 'viagra uden receipt', 'viagra uk', 'viagra uk 32', 'viagra uk buy online', 'viagra uk cheap purchase', 'viagra uk cost pill', 'viagra uk delivery', 'viagra uk forum', 'viagra uk kamagra', 'viagra uk news', 'viagra uk online a href', 'viagra uk online', 'viagra uk viagra', 'viagra uk purchase', 'viagra ghuk retail price', 'viagra ukasdasd retffsRxPILLS - Best discounts to all Pharmacysadasdasddailers viasd 2 days ago'. To the right of the link text is a small thumbnail image showing a couple.

PageRank



PageRank

- Not just a count of inlinks
 - A link from a more important page is more important
 - A link from a page with fewer links is more important
- ∴ A page with lots of inlinks from important pages (which have few outlinks) is more important

PageRank is Recursive

- Not just a count of inlinks
 - A link from a more important page is more important

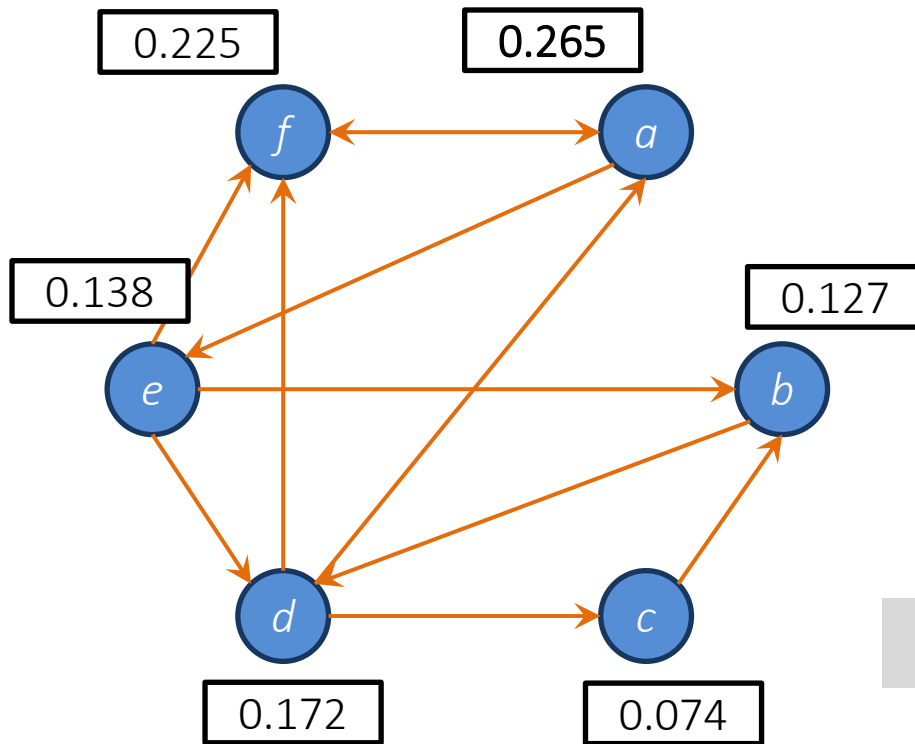
– A link from a page

∴ A page with lots of
have few outlin



PageRank Model

- The Web: a directed graph



$$G = \boxed{V}, \boxed{E}$$

Vertices
(pages)

Edges
(links)

Which vertex is most important?

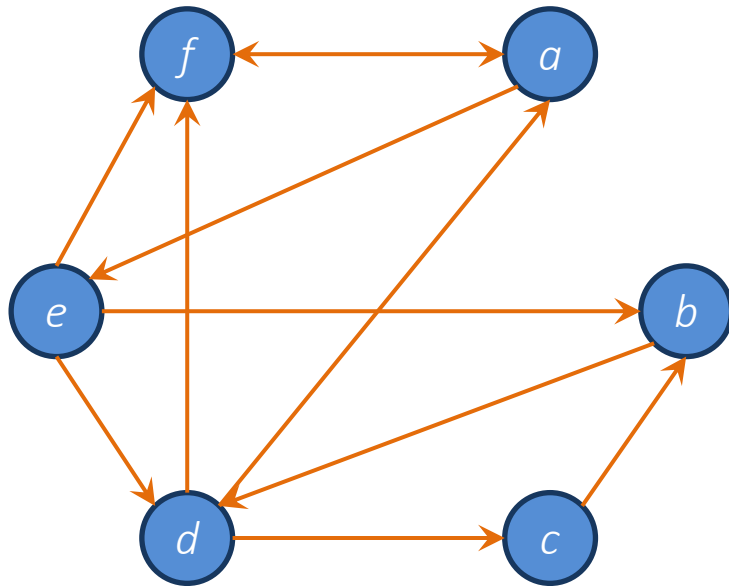


$$V = \{a, b, c, d, e, f\}$$

$$E = \{(a, e), (a, f), (b, d), (c, b), (d, a), (d, c), (d, f), (e, b), (e, d), (e, f), (f, a)\}$$

PageRank Model

- The Web: a directed graph



$$G = \boxed{V} \boxed{E}$$

Vertices
(pages)

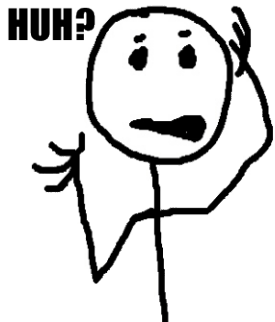
Edges
(links)

$$\text{out}(v) := \{v' \in V \mid (v, v') \in E\}$$

$$\text{in}(v) := \{v' \in V \mid (v', v) \in E\}$$

$$\text{rank}_0(v) := \frac{1}{|V|}$$

$$\text{rank}_i(v) := \sum_{v' \in \text{in}(v)} \frac{\text{rank}_{i-1}(v')}{|\text{out}(v')|}$$



PageRank Model

$$G = [V, E]$$

Vertices
(pages)

Edges
(links)

$$\text{rank}_1(f) = \frac{1}{6} \times \frac{1}{3}$$

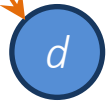
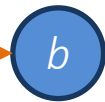


$$\text{rank}_0(e) = \frac{1}{6}$$

$$|\text{out}(e)| = 3$$



$$\text{rank}_1(b) = \frac{1}{6} \times \frac{1}{3}$$



$$\text{rank}_1(d) = \frac{1}{6} \times \frac{1}{3}$$

$$\text{out}(v) := \{v' \in V \mid (v, v') \in E\}$$

$$\text{in}(v) := \{v' \in V \mid (v', v) \in E\}$$

$$\text{rank}_0(v) := \frac{1}{|V|}$$

$$\text{rank}_i(v) := \sum_{v' \in \text{in}(v)} \frac{\text{rank}_{i-1}(v')}{|\text{out}(v')|}$$

PageRank Model

$$G = [V, E]$$

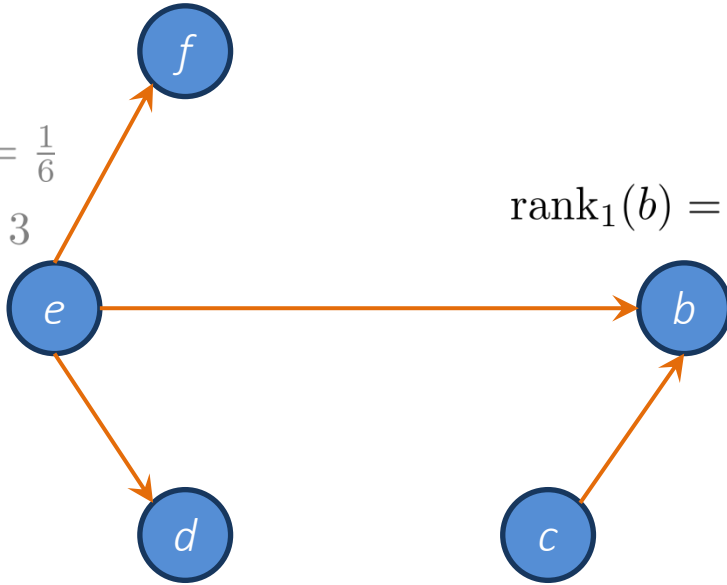
Vertices
(pages)

Edges
(links)

$$\text{rank}_1(f) = \frac{1}{6} \times \frac{1}{3}$$

$$\text{rank}_0(e) = \frac{1}{6}$$

$$|\text{out}(e)| = 3$$



$$\text{rank}_1(b) = \frac{1}{6} \times \frac{1}{3} + 1 \times \frac{1}{6}$$

$$\text{rank}_1(d) = \frac{1}{6} \times \frac{1}{3}$$

$$\text{rank}_0(c) = \frac{1}{6}$$

$$|\text{out}(c)| = 1$$

• • •

$$\text{out}(v) := \{v' \in V \mid (v, v') \in E\}$$

$$\text{in}(v) := \{v' \in V \mid (v', v) \in E\}$$

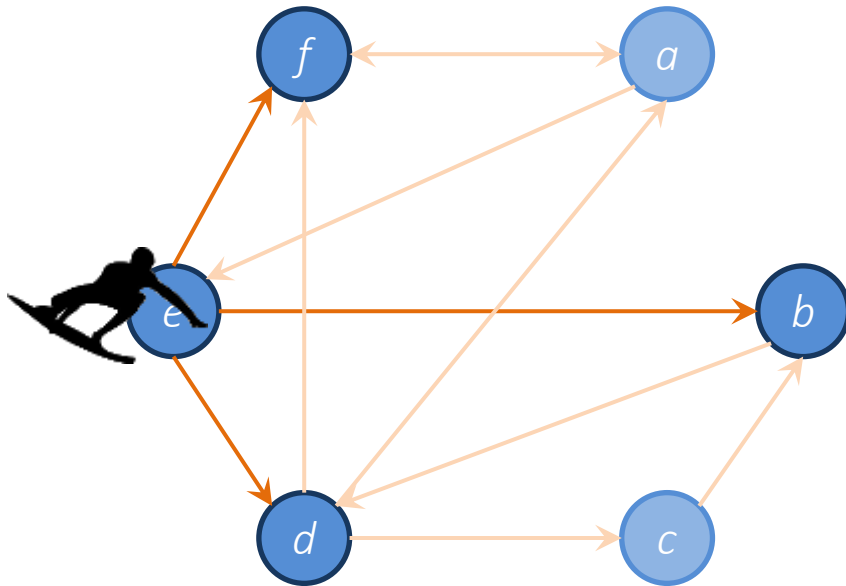
$$\text{rank}_0(v) := \frac{1}{|V|}$$

$$\text{rank}_i(v) := \sum_{v' \in \text{in}(v)} \frac{\text{rank}_{i-1}(v')}{|\text{out}(v')|}$$

PageRank: Random Surfer Model



= someone surfing the web,
clicking links randomly

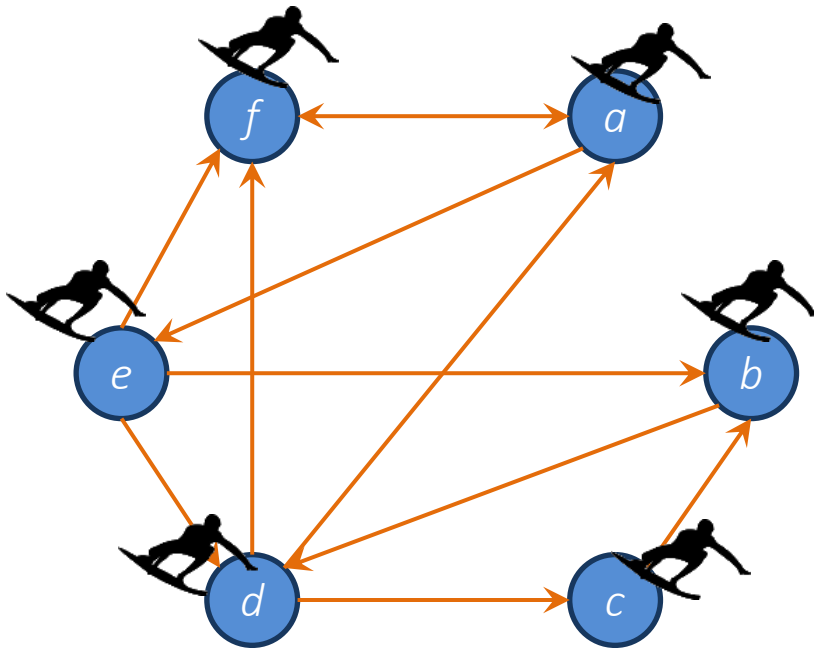


- What is the probability of being at page x after n hops?

PageRank: Random Surfer Model



= someone surfing the web,
clicking links randomly

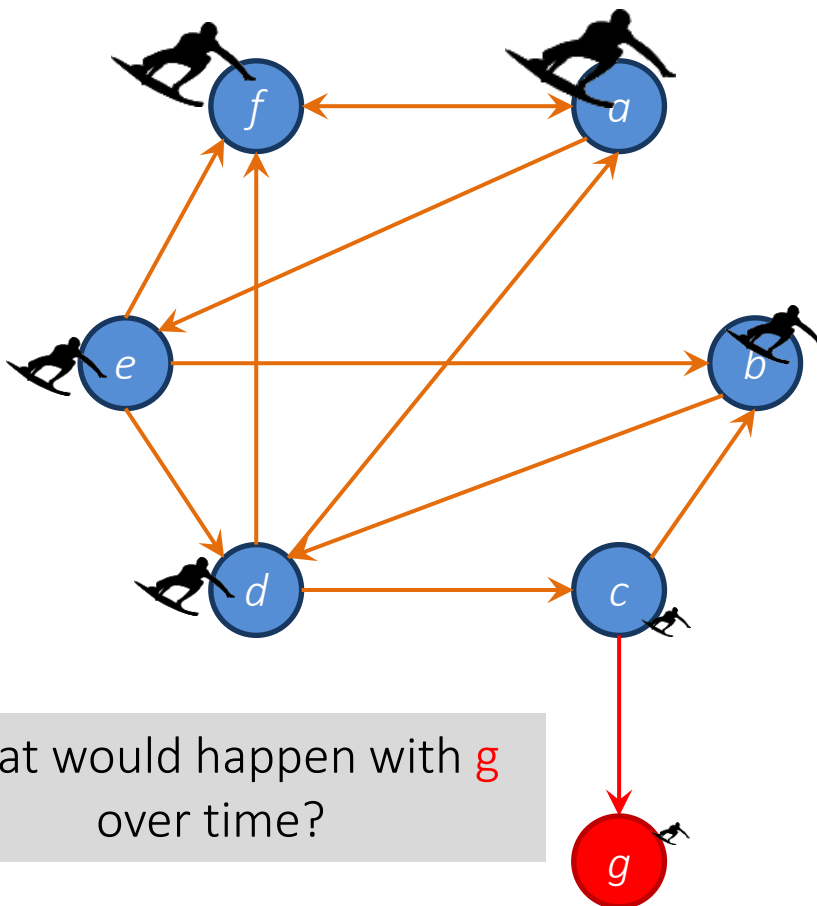


- What is the probability of being at page x after n hops?
- *Initial state*: surfer equally likely to start at any node

PageRank: Random Surfer Model



= someone surfing the web,
clicking links randomly



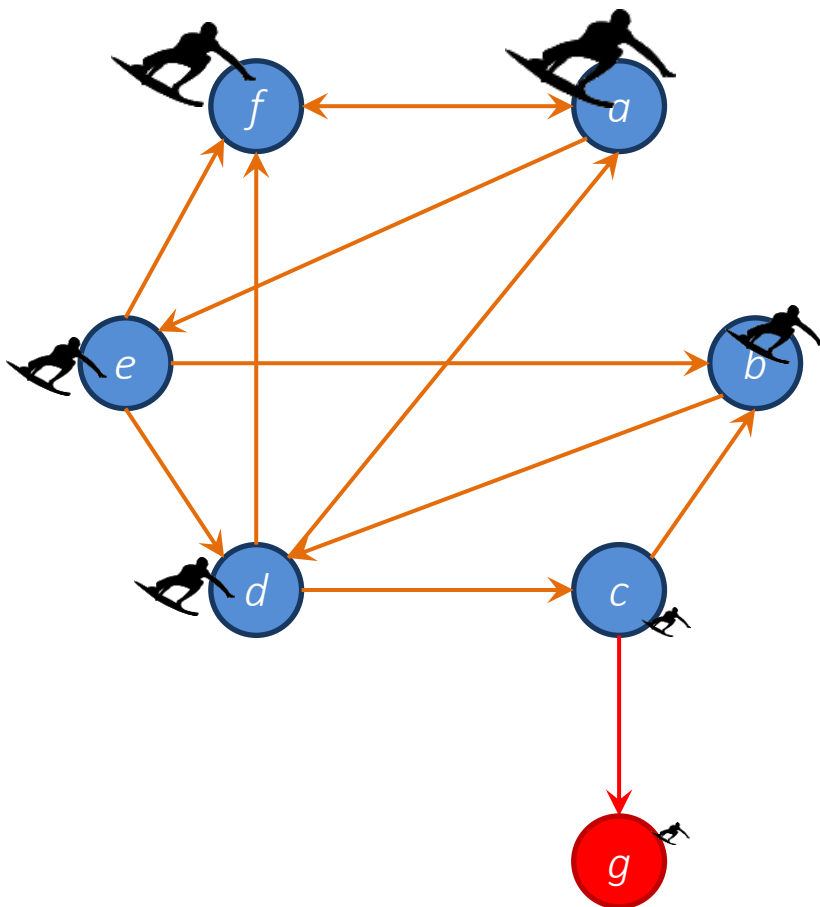
What would happen with **g**
over time?

- What is the probability of being at page x after n hops?
- *Initial state*: surfer equally likely to start at any node
- PageRank applied iteratively for each hop: score indicates probability of being at that page after that many hops

PageRank: Random Surfer Model



= someone surfing the web,
clicking links randomly

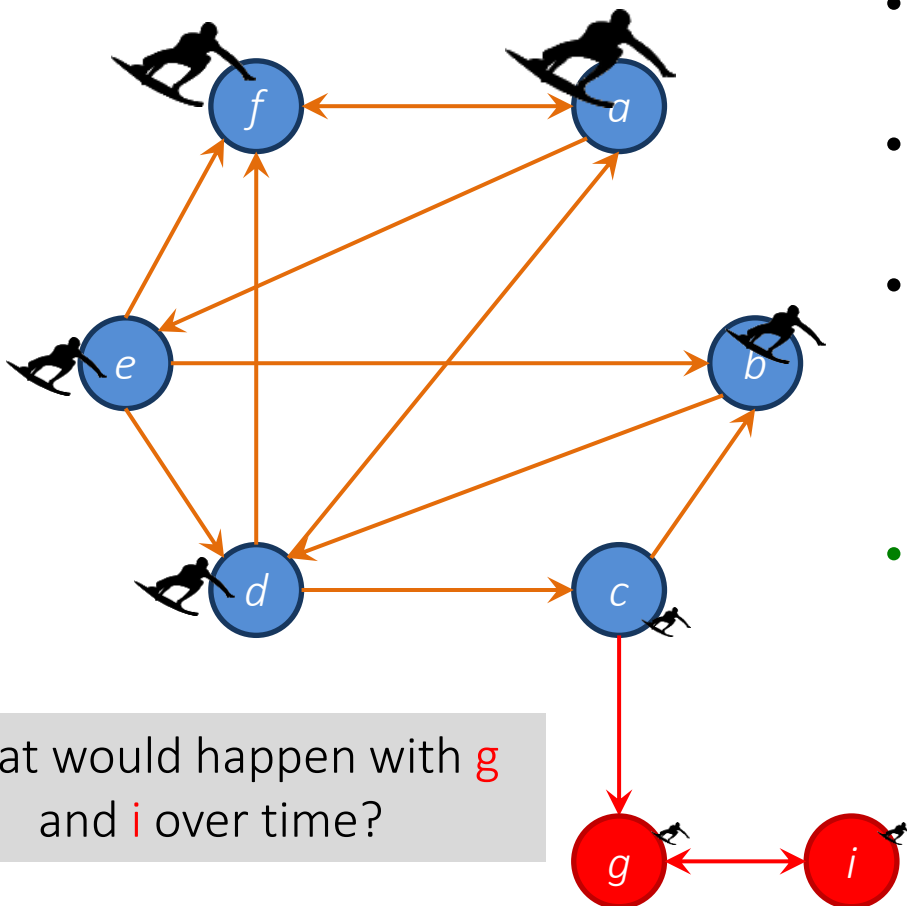


- What is the probability of being at page x after n hops?
- *Initial state*: surfer equally likely to start at any node
- PageRank applied iteratively for each hop: score indicates probability of being at that page after that many hops
- If the surfer reaches a page without links, the surfer randomly jumps to another page

PageRank: Random Surfer Model



= someone surfing the web, clicking links randomly



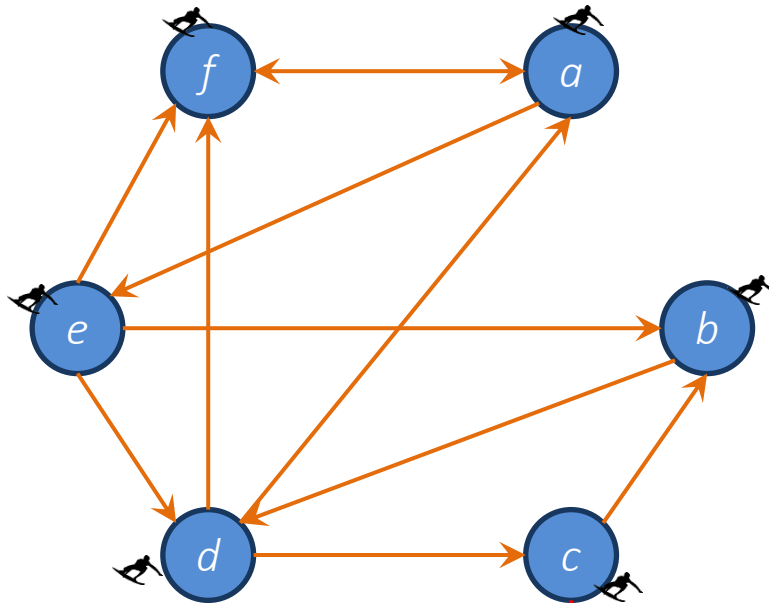
What would happen with **g** and **i** over time?

- What is the probability of being at page x after n hops?
- *Initial state*: surfer equally likely to start at any node
- PageRank applied iteratively for each hop: score indicates probability of being at that page after that many hops
- If the surfer reaches a page without links, the surfer randomly jumps to another page

PageRank: Random Surfer Model

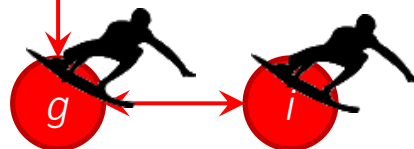


= someone surfing the web, clicking links randomly



- What is the probability of being at page x after n hops?
- *Initial state*: surfer equally likely to start at any node
- PageRank applied iteratively for each hop: score indicates probability of being at that page after than many hops
- If the surfer reaches a page without out-links, the surfer randomly jumps to another page

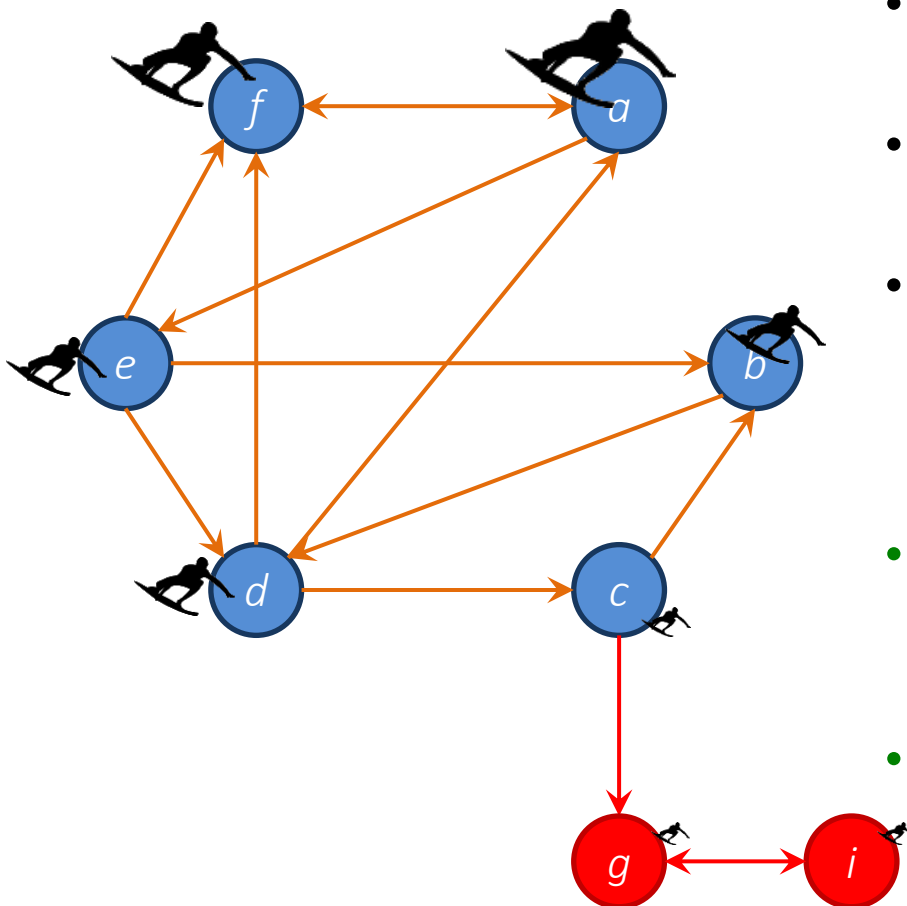
What would happen with g and i over time?



PageRank: Random Surfer Model



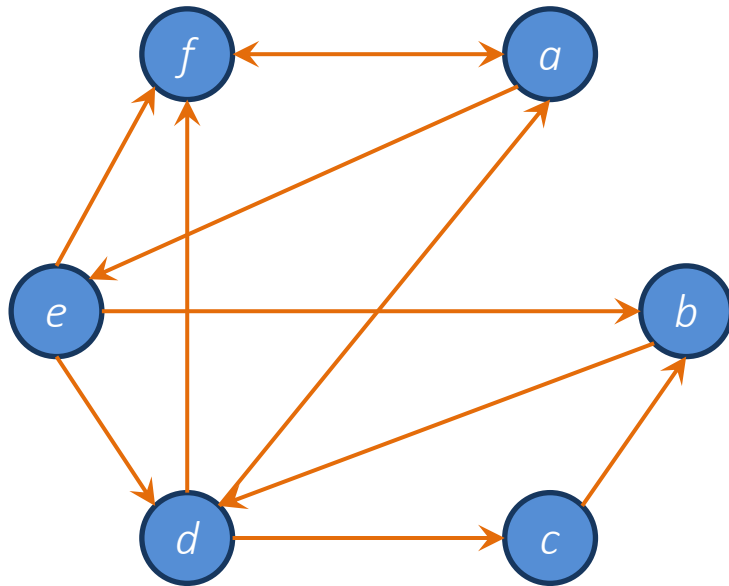
= someone surfing the web,
clicking links randomly



- What is the probability of being at page x after n hops?
- *Initial state*: surfer equally likely to start at any node
- PageRank applied iteratively for each hop: score indicates probability of being at that page after than many hops
- If the surfer reaches a page without out-links, the surfer randomly jumps to another page
- The surfer will jump to a random page at any time with a probability $1 - d$... *this avoids traps and ensures convergence!*

PageRank Model: Final Version

- The Web: a directed graph



$$G = \boxed{V} \boxed{E}$$

Vertices
(pages)

Edges
(links)

$$\text{out}(v) := \{v' \in V \mid (v, v') \in E\}$$

$$\text{in}(v) := \{v' \in V \mid (v', v) \in E\}$$

$$\text{rank}_0(v) := \frac{1}{|V|}$$

$$V' := \{v \in V : |\text{out}(v)| = 0\}$$

$$V'' := \{v \in V : |\text{out}(v)| \neq 0\}$$

d is the follow-a-link probability
typically ($d = 0.85$)

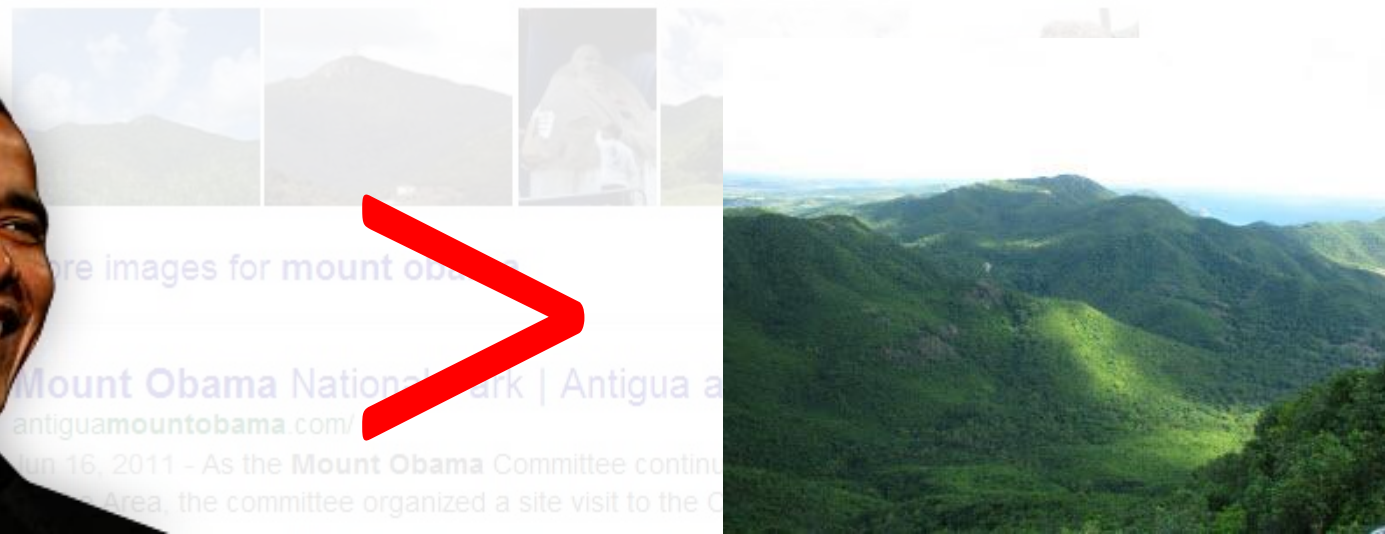
$$\text{rank}_i(v) := d \times \sum_{u \in \text{in}(v)} \frac{\text{rank}_{i-1}(u)}{|\text{out}(u)|} + \sum_{v' \in V'} \frac{\text{rank}_{i-1}(v')}{|V|} + (1-d) \times \sum_{v'' \in V''} \frac{\text{rank}_{i-1}(v'')}{|V|}$$

PageRank: Benefits



- ✓ More robust than a simple link count
- ✓ Fewer ties than link counting
- ✓ Scalable to approximate (for sparse graphs)
- ✓ Convergence guaranteed

Two Sides to Ranking: Importance

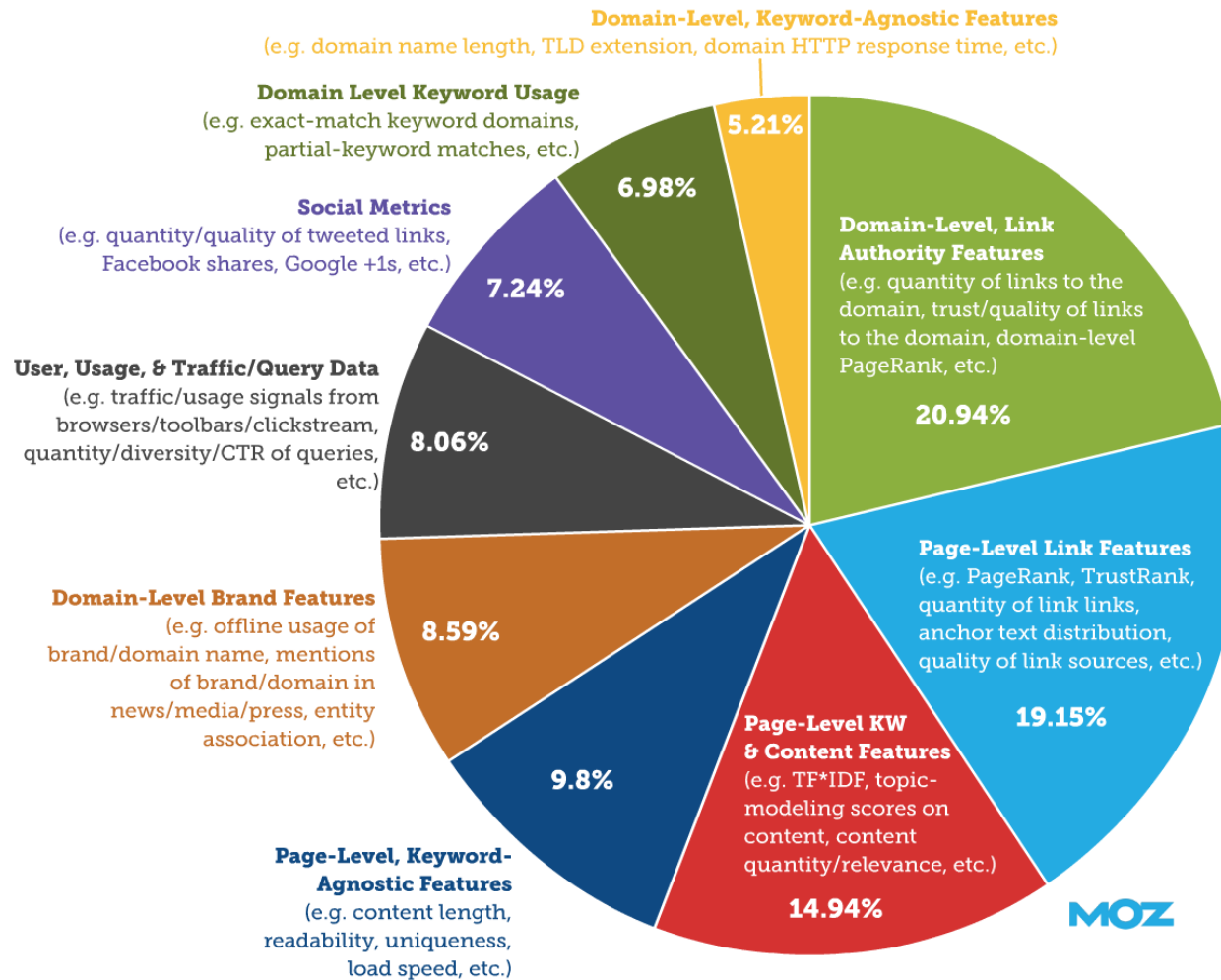


HOW DOES GOOGLE REALLY RANK?
AN EDUCATED GUESS

How Modern Google ranks results (maybe)

Weighting of Thematic Clusters of Ranking Factors in Google

(based on survey responses by 128 SEO professionals in June 2013)



According to survey of SEO experts, not people in Google

How Modern Google ranks results (maybe)

Weighting of Thematic Clusters of Ranking Factors in Google

(based on survey responses by 128 SEO professionals in June 2013)

Domain-Level, Keyword-Agnostic Features
(e.g. domain name length, TLD extension, domain HTTP response time, etc.)

Why so secretive?



partial-keyword matches, etc.)

6.98%

User

Engage
Quality

D

b



Page-Level, Keyword-Agnostic Features
(e.g. content length, readability, uniqueness, load speed, etc.)

quantity/relevance, etc.)

14.94%

MOZ

According to survey of SEO experts, not people in Google

Ranking: Science or Art?





Questions?