

# CC5212-1

PROCESAMIENTO MASIVO DE DATOS

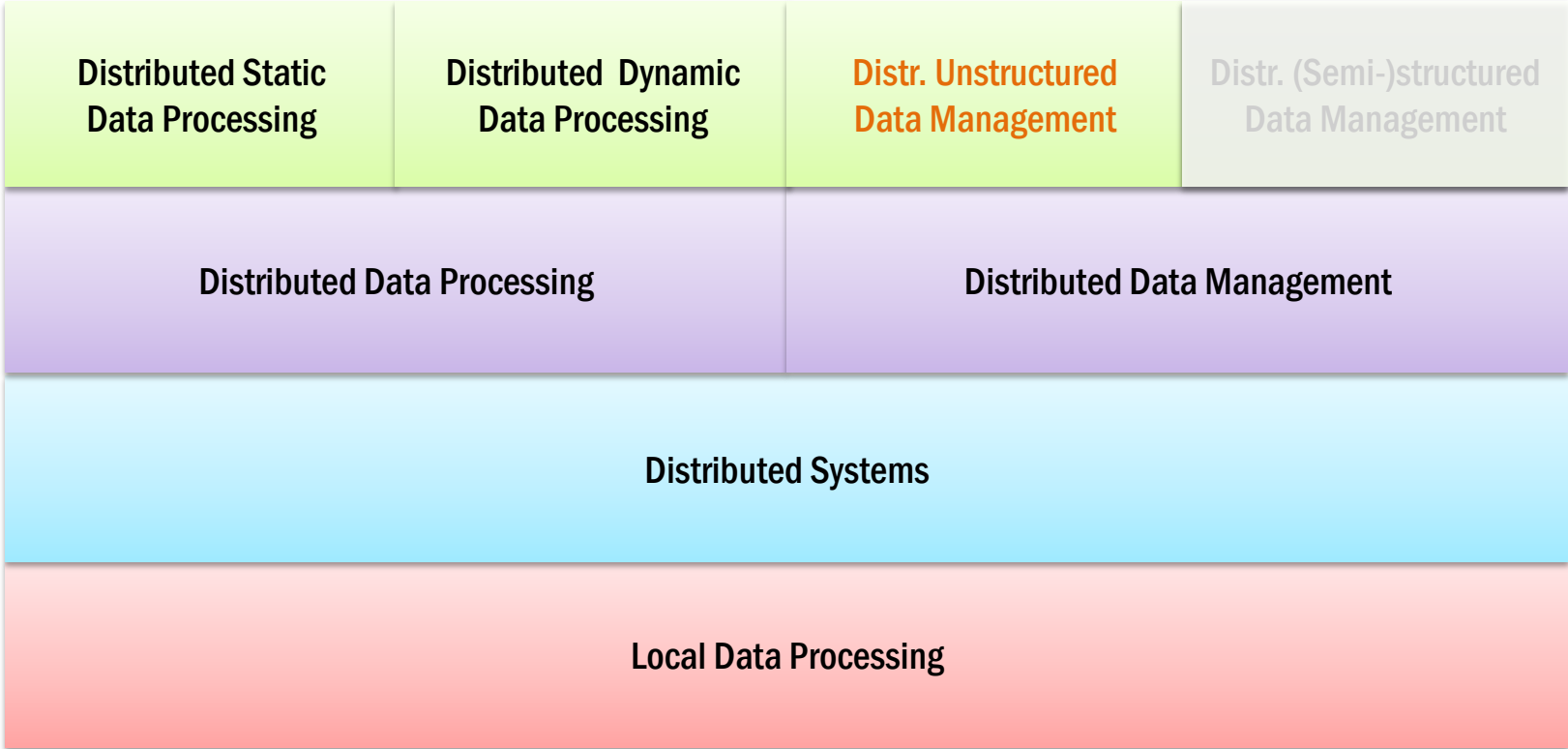
OTOÑO 2023

## Lecture 8

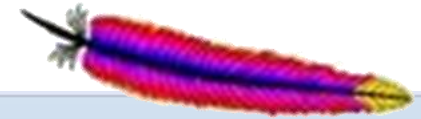
### Information Retrieval: Ranking

Aidan Hogan

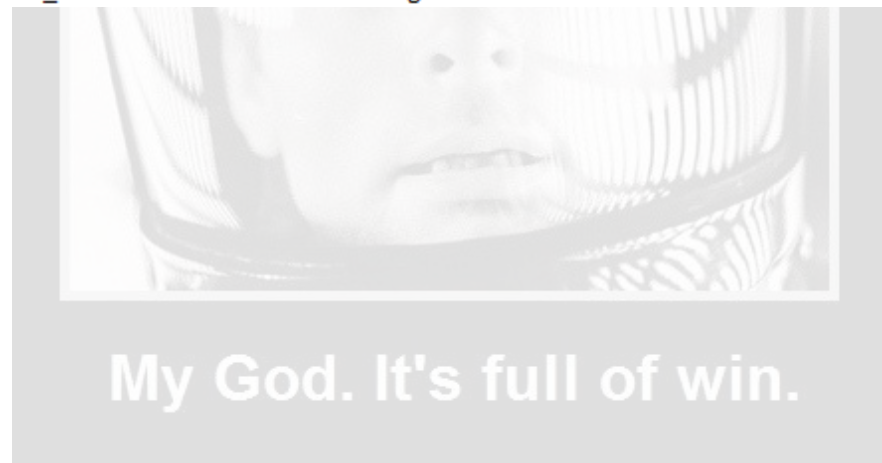
[aidhog@gmail.com](mailto:aidhog@gmail.com)



# Apache Lucene

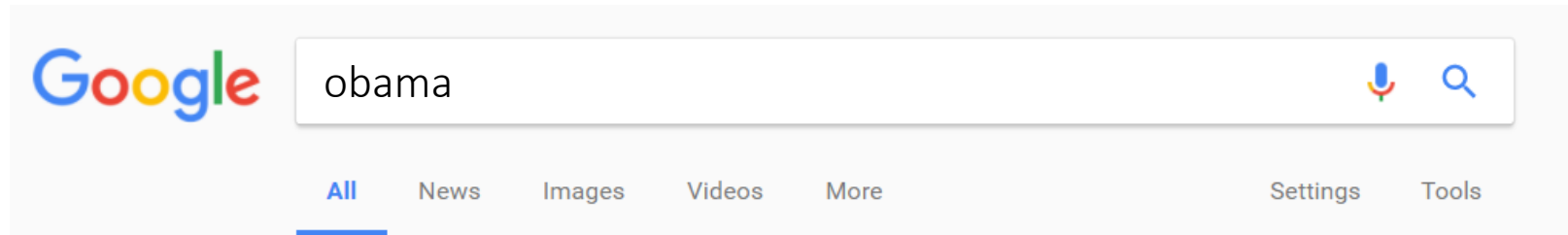


```
Tasks Console
SearchWikiIndex [Java Application] C:\Program Files\Java\jre1.8.0_65\bin\javaw.exe (03-05-2017 12:45:22 a. m.)
Opening directory at lucene
Enter a keyword search phrase:
obama
Running query: obama
Parsed query: TITLE:obam^5.0 ABSTRACT:obam
Matching documents: 255
Showing top 10 results
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](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

Google™

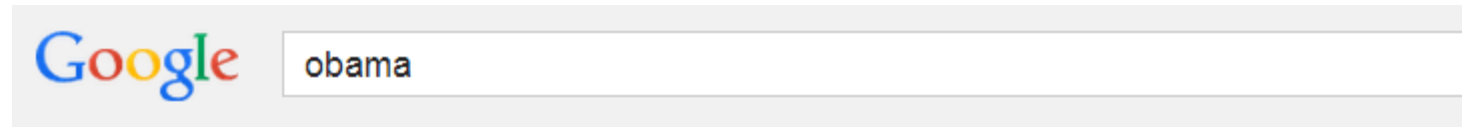
sometimes when i'm

- 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](http://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](http://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](http://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 ...

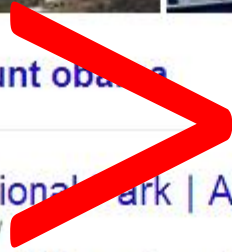
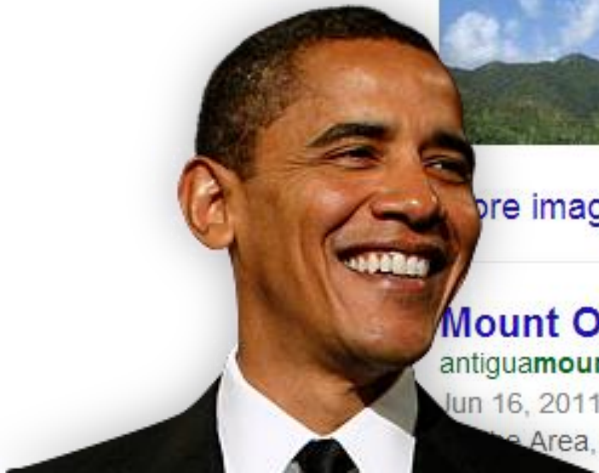
**Images for mount obama** [Report images](#)



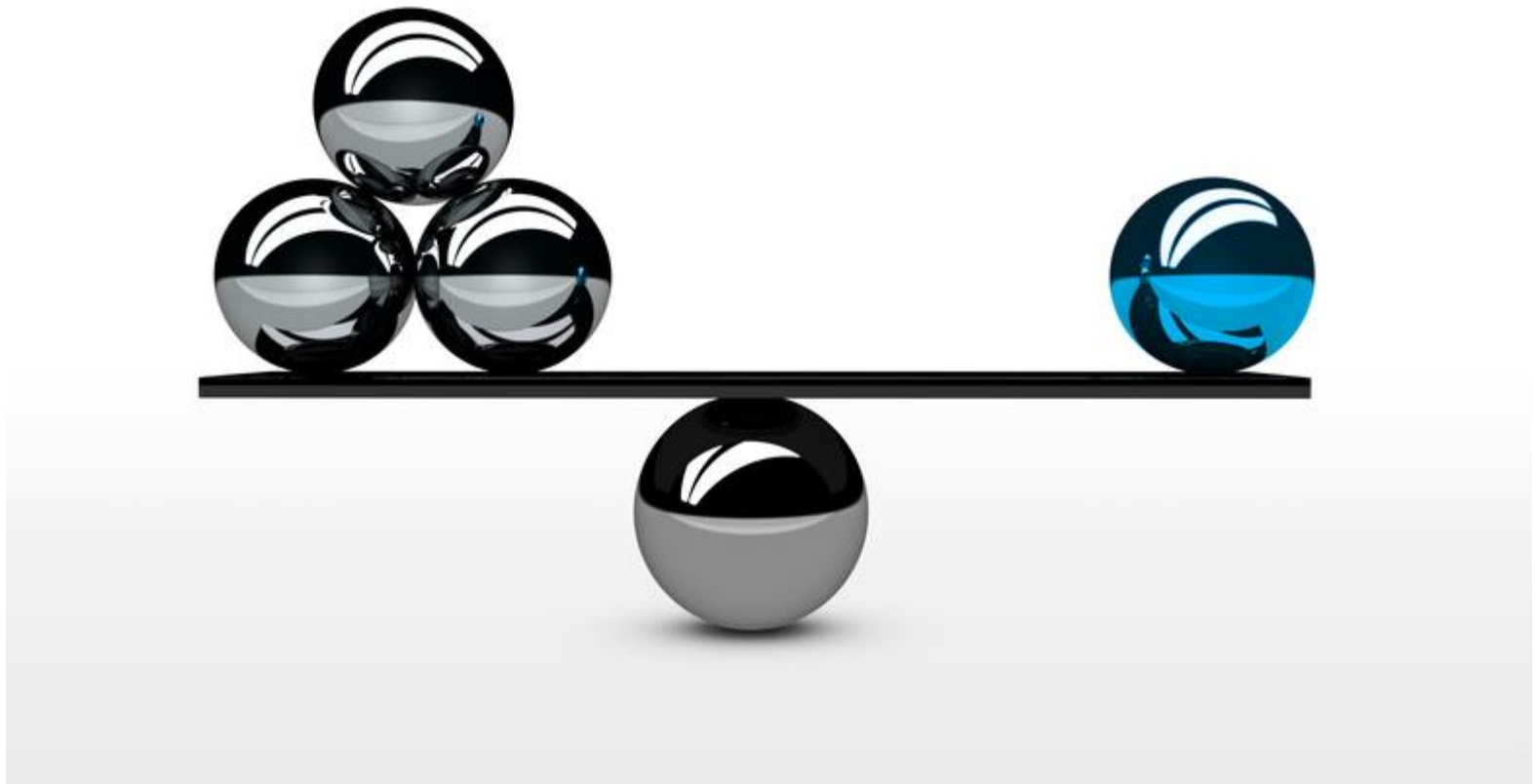
More images for mount obama

**Mount Obama National Park | Antigua and Barbuda**  
[antiguamountobama.com/](http://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](http://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](http://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](http://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, a quote from the film, and a movie poster. Two search results are highlighted: 'freedom' (7 occurrences) and 'movie' (16 occurrences). The search results are displayed in a box at the bottom of the page.

W Braveheart - Wikipedia

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](#).

**Braveheart**

MEL · GIBSON

*Every man dies,  
not every man  
really lives.*

BRAVEHEART

freedom

- 7 occurrences

movie

- 16 occurrences

# 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 content includes the Wikipedia logo, navigation tabs for 'Article' and 'Talk', and a search bar. The main text of the article describes the 1995 film 'Braveheart' directed by Mel Gibson, starring Mel Gibson as William Wallace. The text mentions that the film was nominated for ten Academy Awards and won five. A movie poster for 'Braveheart' is visible on the right side of the page. Three search results are highlighted in colored boxes: 'freedom' (7 occurrences) in a red box, 'movie' (16 occurrences) in an orange box, and 'wallace' (88 occurrences) in a green box. The search results for 'wallace' are also visible in the browser's search bar at the top right, showing 'wallace' and '44 de 88'.

W Braveheart - Wikipedia

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

1 Plot

2 Cast

3 Production

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

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

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

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

# Relevance Measure: TF-IDF



## Term Frequency

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

## Inverse Document Frequency

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

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

| $t$     | $\text{tf}(t, d)$ | $ \{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

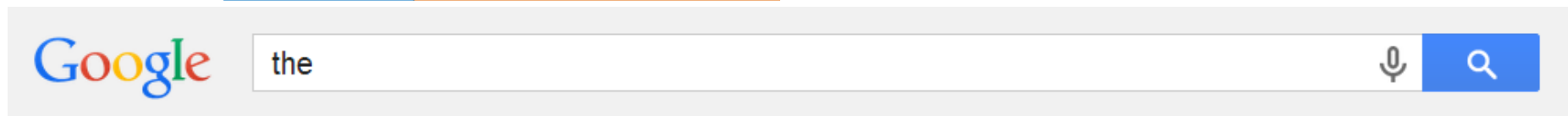
$$tf(t, d) = \text{count}(t, d)$$

## Inverse Document Frequency

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

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

| $t$     | $tf(t, d)$ | $ \{d \in D : t \in d\} $ | $\frac{ D +1}{ \{d \in D : t \in d\} +1}$ |
|---------|------------|---------------------------|---|
| 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

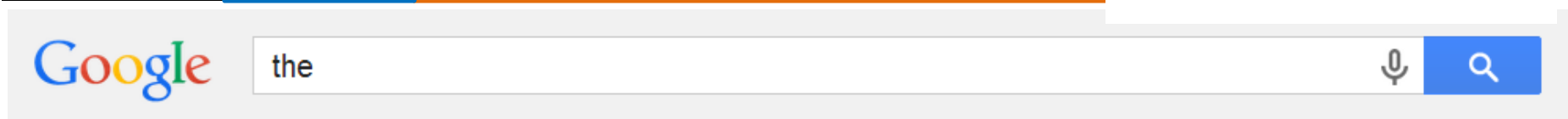
$$tf(t, d) = \text{count}(t, d)$$

## Inverse Document Frequency

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

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

| $t$     | $tf(t, d)$ | $ \{d \in D : t \in d\} $ | $\frac{ D +1}{ \{d \in D : t \in d\} +1}$ |
|---------|------------|---------------------------|---|
| movie   | 16         | 835,000,000               | 13.66                                     |
| freedom | 7          | 198,000,000               | 57.63                                     |
| wallace | 43         | 49,200,000                | 231.91                                    |



Web Images News Books More Search tools

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

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

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

| $t$     | $\text{tf}(t, d)$ | $ \{d \in D : t \in d\} $ | $\frac{ D +1}{ \{d \in D : t \in d\} +1}$ | $\text{idf}(t, d)$ |
|---------|-------------------|---------------------------|---|--------------------|
| movie   | 16                | 835,000,000               | 13.66                                     | 3.77               |
| freedom | 7                 | 198,000,000               | 57.63                                     | 5.85               |
| wallace | 43                | 49,200,000                | 231.91                                    | 7.86               |

# Relevance Measure: TF-IDF



## Term Frequency

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

## Inverse Document Frequency

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

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

| $t$     | $\text{tf}(t, d)$ | $ \{d \in D : t \in d\} $ | $\frac{ D +1}{ \{d \in D : t \in d\} +1}$ | $\text{idf}(t, d)$ | $\text{tf-idf}(t, d)$ |
|---------|-------------------|---------------------------|---|--------------------|-----------------------|
| movie   | 16                | 835,000,000               | 13.66                                     | 3.77               | 60.36                 |
| freedom | 7                 | 198,000,000               | 57.63                                     | 5.85               | 40.94                 |
| wallace | 43                | 49,200,000                | 231.91                                    | 7.86               | 337.87                |

# Relevance Measure: TF-IDF



## Term Frequency

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

## Inverse Document Frequency

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

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

| $t$     | $\text{tf}(t, d)$ | $ \{d \in D : t \in d\} $ | $\frac{ D +1}{ \{d \in D : t \in d\} +1}$ | $\text{idf}(t, d)$ | $\text{tf-idf}(t, d)$ |
|---------|-------------------|---------------------------|---|--------------------|-----------------------|
| movie   | 16                | 835,000,000               | 13.66                                     | 3.77               | 60.36                 |
| freedom | 7                 | 198,000,000               | 57.63                                     | 5.85               | 40.94                 |
| wallace | 43                | 49,200,000                | 231.91                                    | 7.86               | 337.87                |

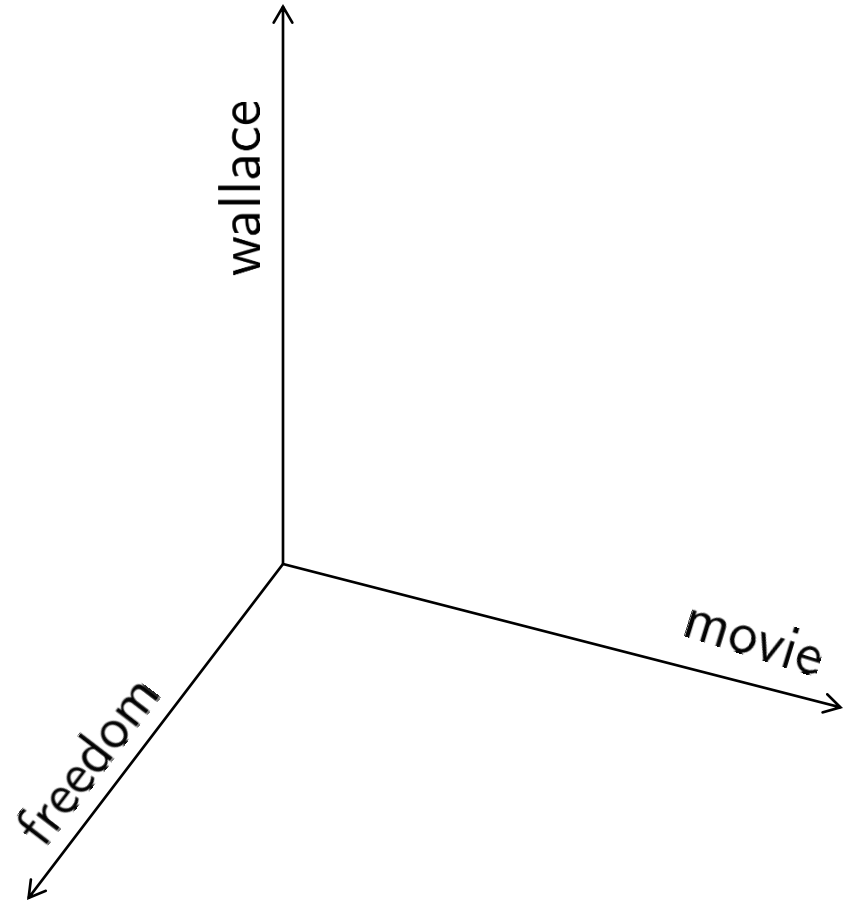
$$\text{score}(q, d) = \sum_{t \in q} \text{tf-idf}(t, d)$$

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

# Vector Space Model (a mention)

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

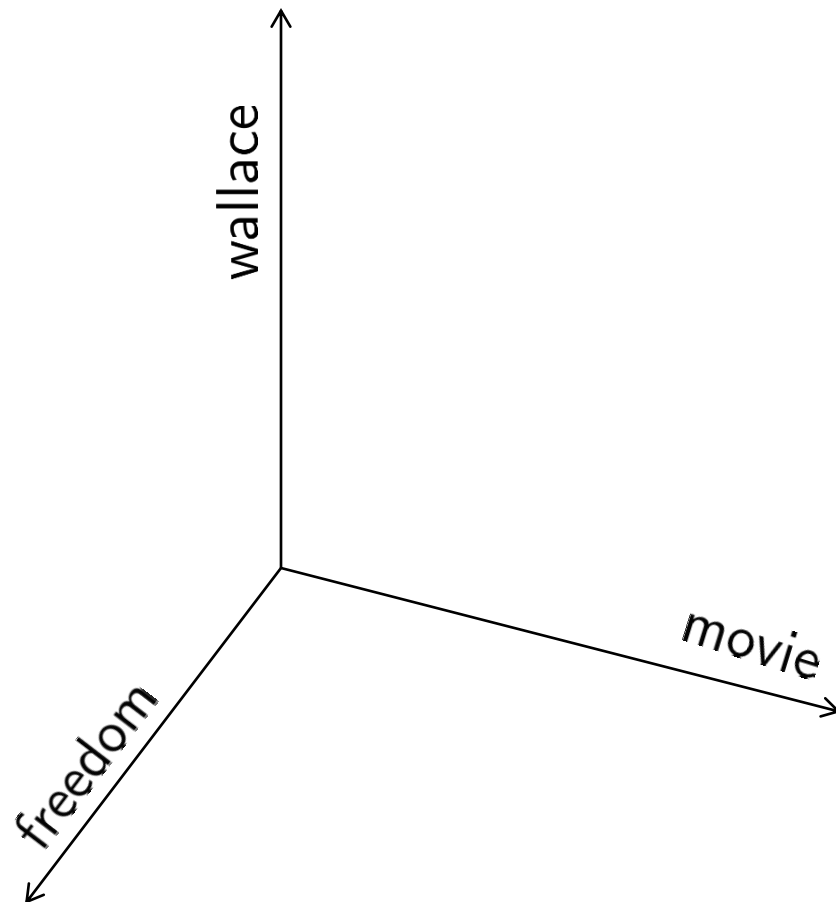
$$l = \sqrt{\sum_{t \in q} \text{tf}(t, d)^2}$$



# 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               |

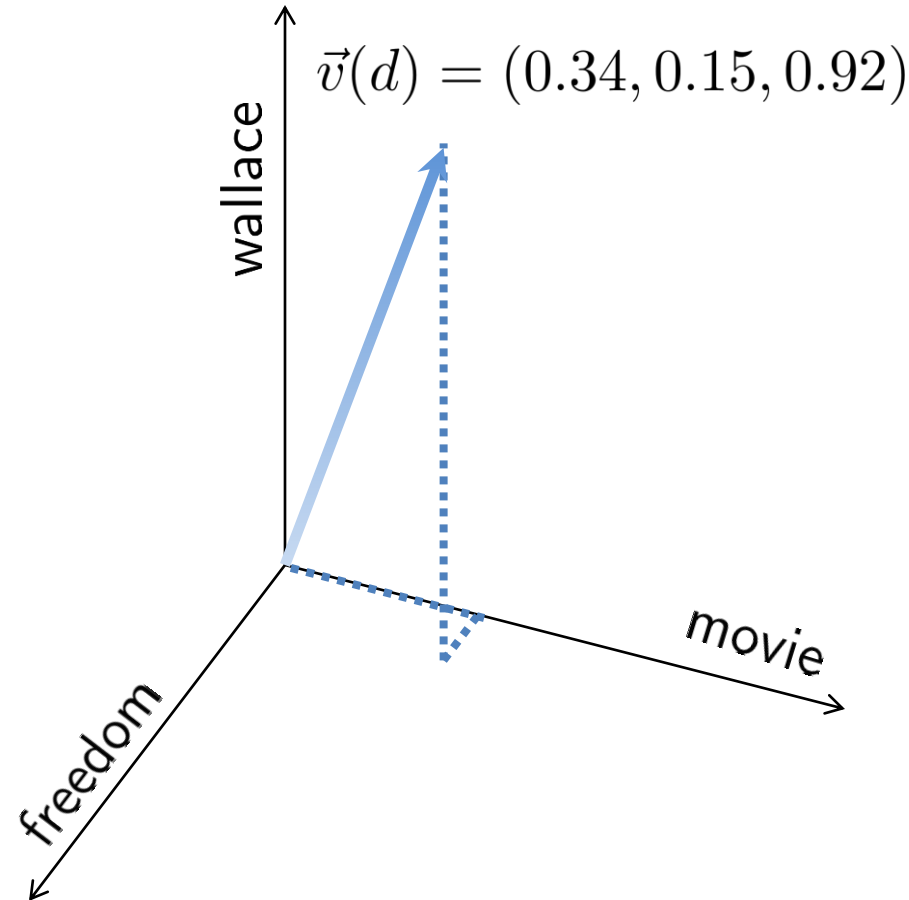
$$l = \sqrt{\sum_{t \in q} \text{tf}(t, d)^2}$$



# 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     |

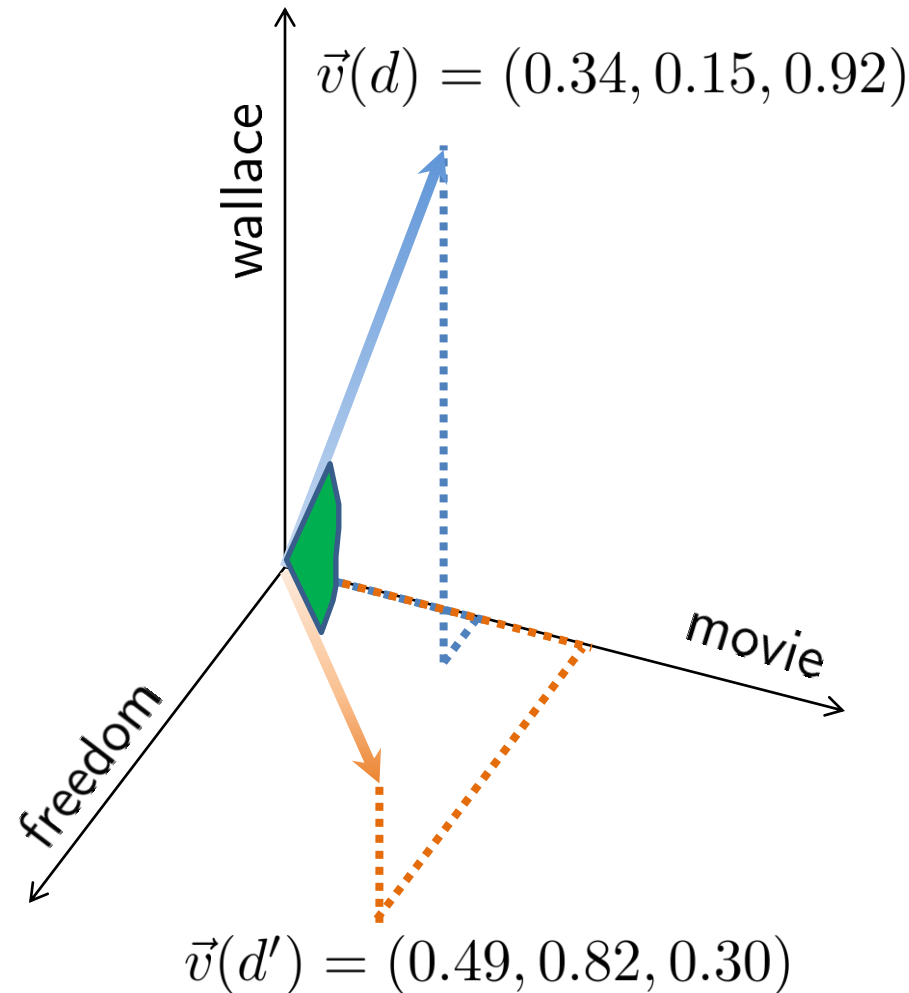
$$\text{sim}(d, d') \approx 0.57$$

$\Sigma$

- 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



A screenshot of a Google search interface. The search bar contains the text "obama". Below the search bar, there are tabs for "Web", "Images", "News", "Videos", "More", and "Search tools". The "Web" tab is selected. Below the tabs, it says "About 16,700,000 results (0.23 seconds)". The search results are as follows:

- 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 ...

In the bottom left corner, there is a photograph of Barack Obama smiling. In the bottom right corner, there is a photograph of a head of broccoli. A large red "not equal" symbol ( $\neq$ ) is overlaid on the search results, positioned between the "News for broccoli" and "Mistakes We All Make With Spaghetti, Steak And E" results.

# 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 prominently displayed in a large, bold font, enclosed in an orange box. To the right of the title are icons for a lock, a speaker, and a star. Below the title, the text reads "From Wikipedia, the free encyclopedia". A redaction notice states: "Obama" redirects here. For other uses, see Obama (disambiguation). Below this, a note says: "This article is about the 44th president of the United States. For his father, see Barack Obama, Sr." The main body of text begins with "Barack Hussein Obama II" followed by a pronunciation guide and a detailed biography. On the right side, there is a portrait of Barack Obama with the caption "Barack Obama". The left sidebar contains navigation links such as "Main page", "Contents", "Featured content", "Current events", "Random article", "Donate to Wikipedia", and "Wikimedia Shop". At the top right, there are links for "Create account" and "Log in".

# 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
```

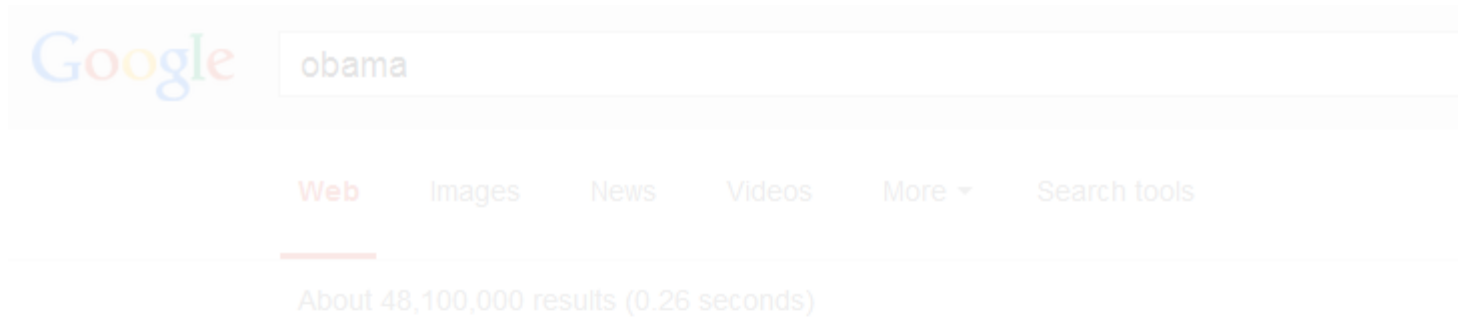


... and Elasticsearch uses Lucene

RANKING:

IMPORTANCE

# Two Sides to Ranking: Importance

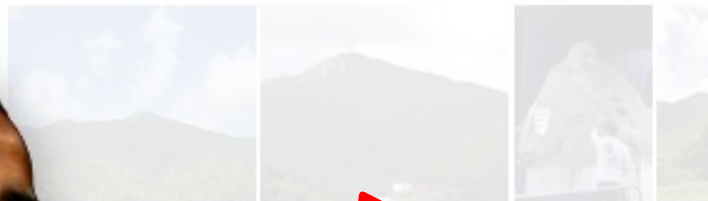


How could we determine that Barack Obama is more important than Mount Obama as a search result for "obama" on the Web?



Images for mount obama

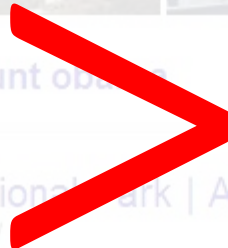
Report images



More images for mount obama

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

Jun 16, 2011 - As the Mount Obama Committee continues its work in the Area, the committee organized a site visit to the C



# Link Analysis

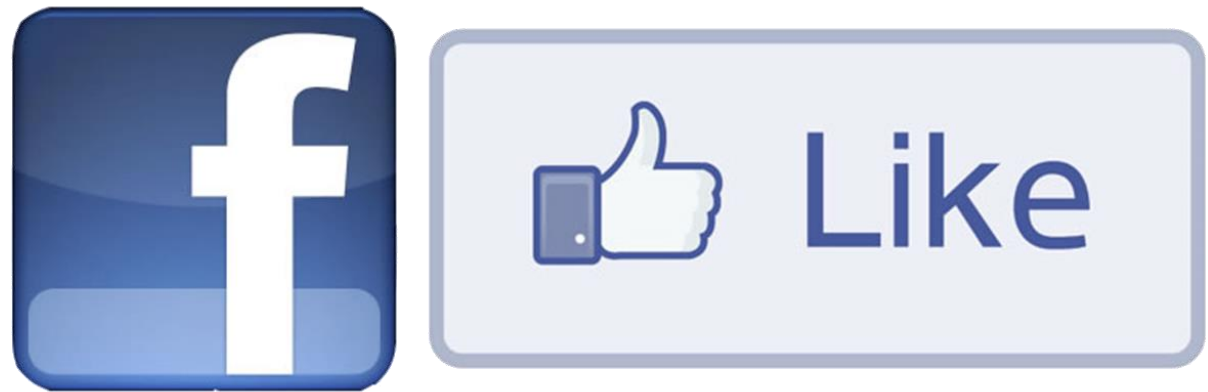
Which will have more links from other pages?  
The Wikipedia article for Mount Obama?  
The Wikipedia article for Barack Obama?





# 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](#) [Miacardis](#) [Mobic](#) [Monoket](#) [Motrin](#) [Myambutol](#) [Mycelex-G](#) [Mysoline](#) [Naprosyn](#) [Neurontin](#) [Nicotinell](#) [Nimotop](#) [Nirdosh](#) [Nizoral](#) [Nolvadex](#) [Nonoxinol](#) [Noroxin](#) [Omnicef](#) [Ophthalmicare](#) [Oxytrol](#) [Pamelor](#) [Parlodol](#) [Paxil](#) [Penisole](#) [Phentermine](#) [Pilex](#) [Plan B](#) [Plavix](#) [Plendil](#) [Pletal](#) [Prandin](#) [Pravachol](#) [Prednisone](#) [Prenamin](#) [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](#) [Serophene](#) [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](http://en.wikipedia.org/wiki/Barack_Obama)?

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



PageRank

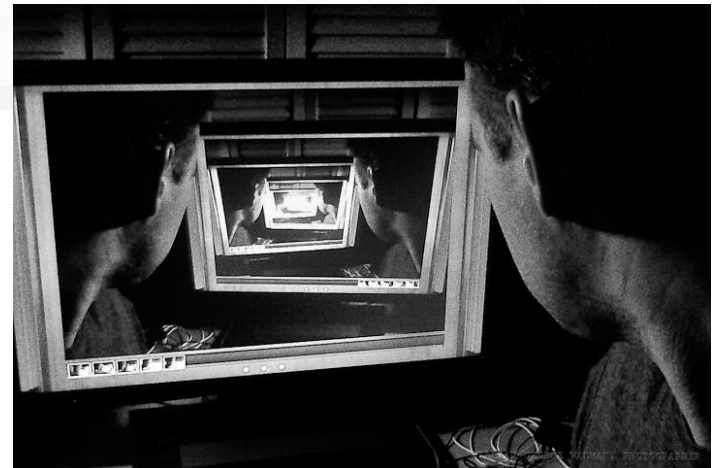


# PageRank: Central Assumption

A page with **lots** of inlinks **from important pages** with **few outlinks** is more important

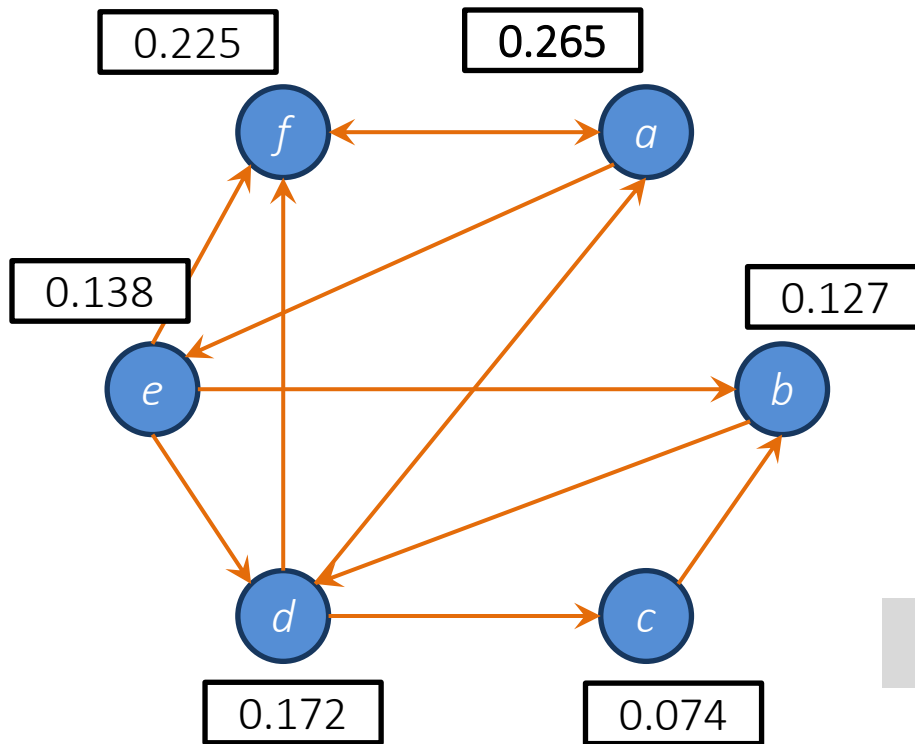
# PageRank: Recursive Definition

A page with **lots** of inlinks from important pages with few outlinks is more important



# PageRank Model

- The Web: a directed graph



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

Vertices  
(pages)

Edges  
(links)

Which vertex is most important?



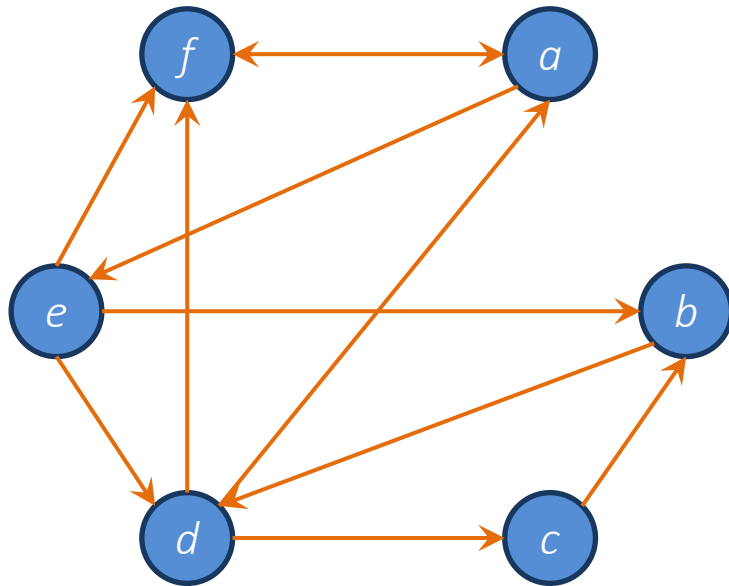
$$V = \{a, b, c, d, e, f\}$$

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



# PageRank Model

- The Web: a directed graph



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

Vertices  
(pages)

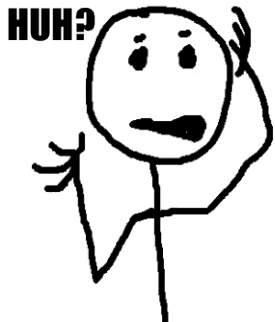
Edges  
(links)

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

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

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

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



# PageRank Model

$$G = [V, E]$$

Vertices  
(pages)

Edges  
(links)

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

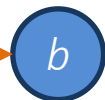


$$\text{rank}_0(e) = \frac{1}{6}$$

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



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



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

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

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

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

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

# PageRank Model

$$G = [V, E]$$

Vertices  
(pages)

Edges  
(links)

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



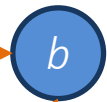
$$\text{rank}_0(e) = \frac{1}{6}$$

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



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

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



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

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

• • •

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

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

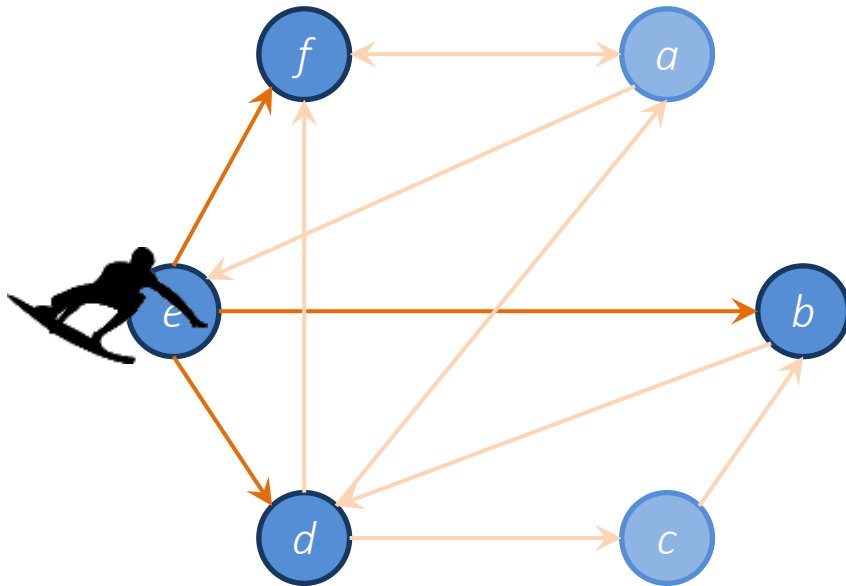
$$\text{rank}_0(v) := \frac{1}{|V|}$$

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

# PageRank: Random Surfer Model



= someone surfing the web,  
clicking links randomly

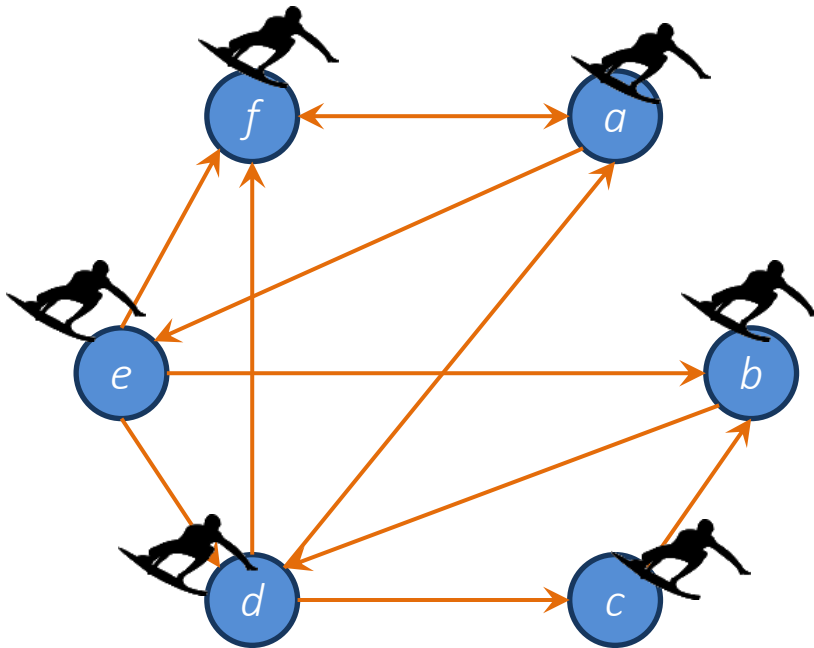


- What is the probability of being at page *x* after *n* hops?

# PageRank: Random Surfer Model



= someone surfing the web,  
clicking links randomly

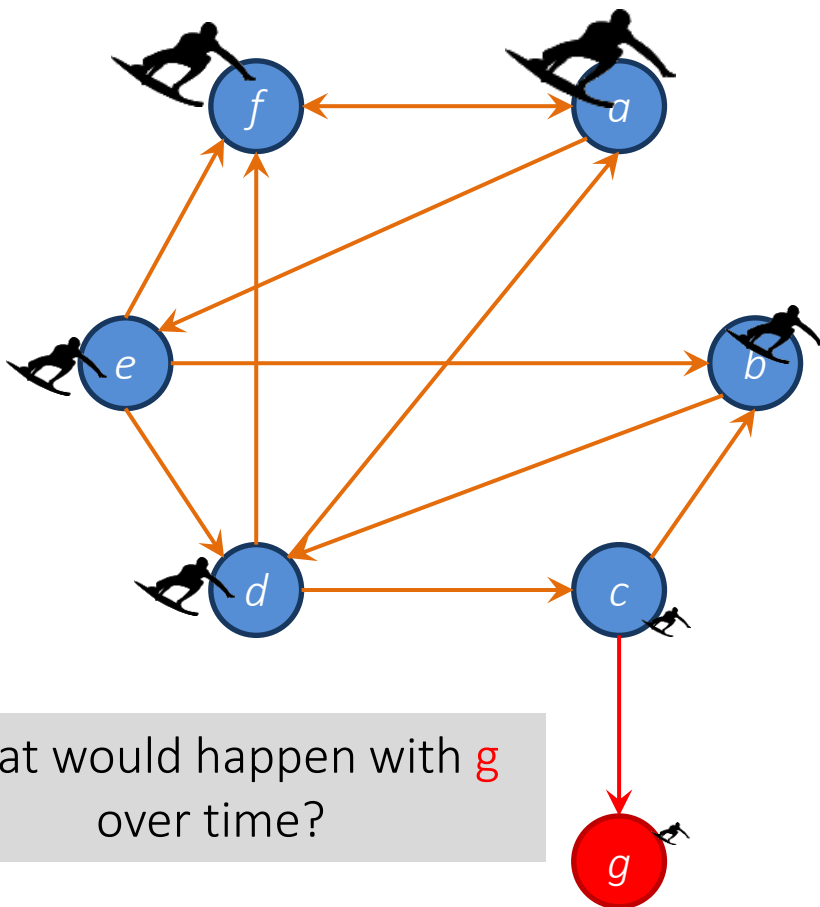


- What is the probability of being at page *x* after *n* hops?
- *Initial state*: surfer equally likely to start at any node

# PageRank: Random Surfer Model



= someone surfing the web,  
clicking links randomly



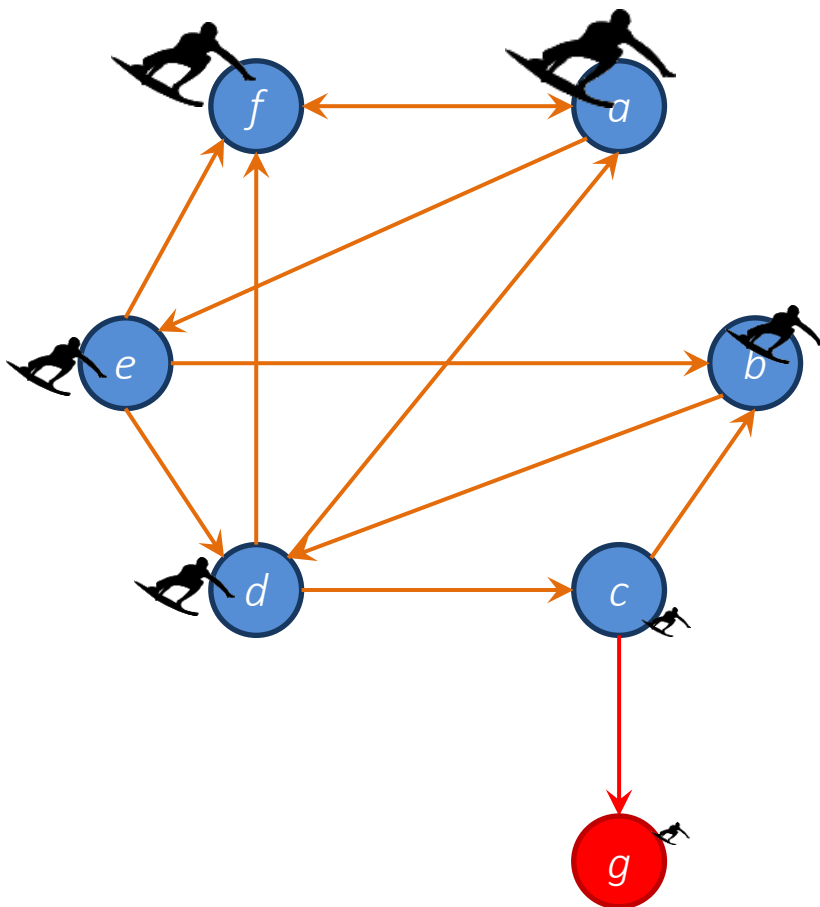
What would happen with **g**  
over time?

- 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

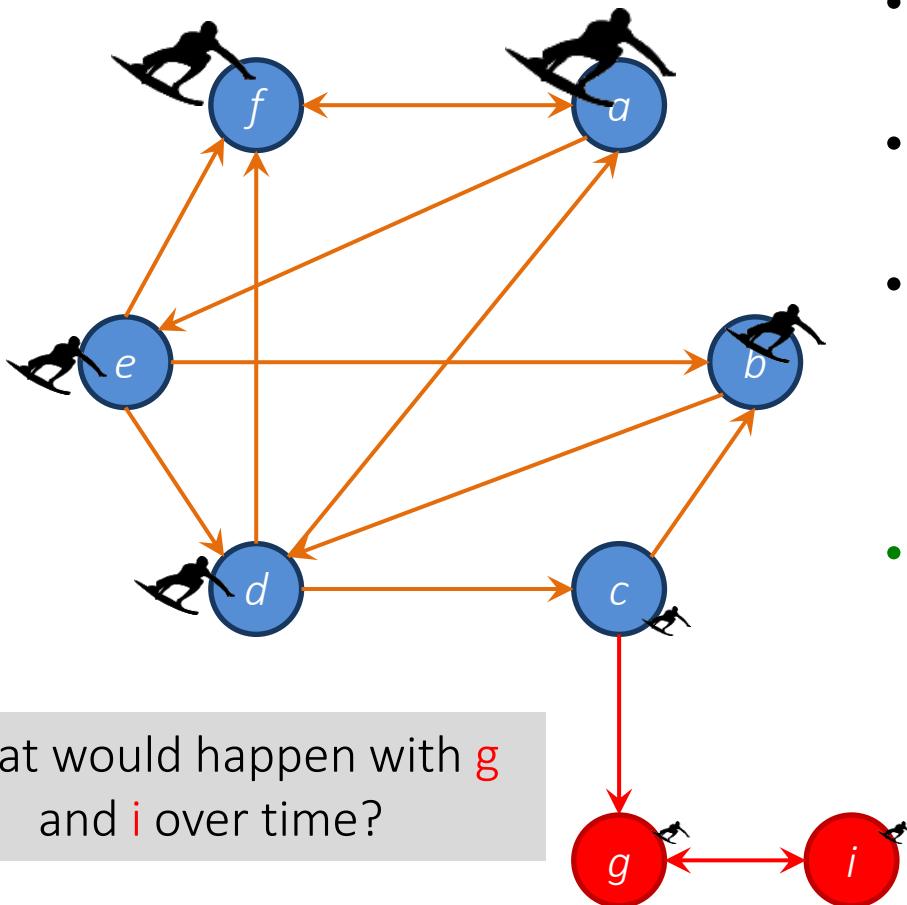


- What is the probability of being at page  $x$  after  $n$  hops?
- *Initial state*: surfer equally likely to start at any node
- PageRank applied iteratively for each hop: score indicates probability of being at that page after that many hops
- If the surfer reaches a page without links, the surfer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

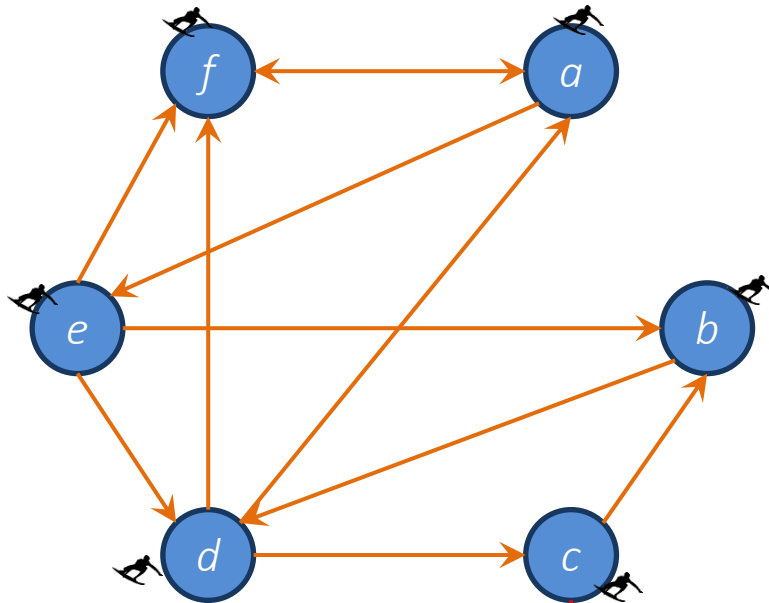
- What is the probability of being at page  $x$  after  $n$  hops?
- *Initial state*: surfer equally likely to start at any node
- PageRank applied iteratively for each hop: score indicates probability of being at that page after that many hops
- If the surfer reaches a page without links, the surfer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



# PageRank: Random Surfer Model

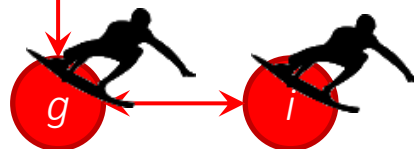


= someone surfing the web,  
clicking links randomly



- 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

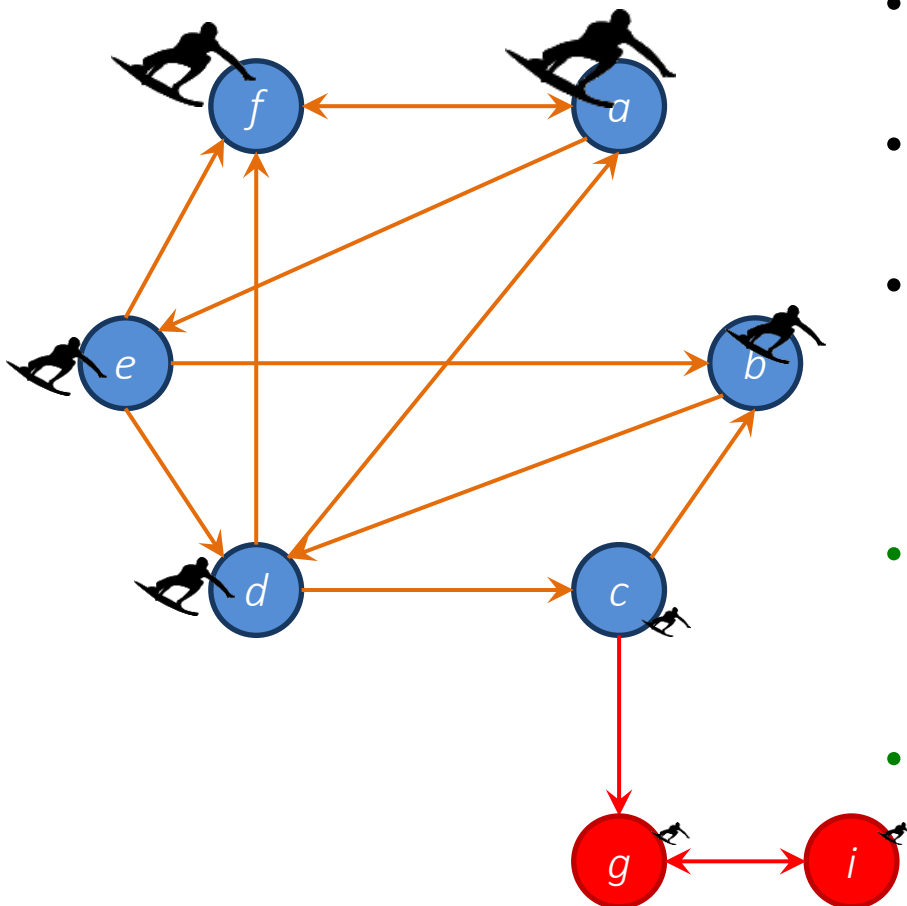
What would happen with  $g$   
and  $i$  over time?



# PageRank: Random Surfer Model



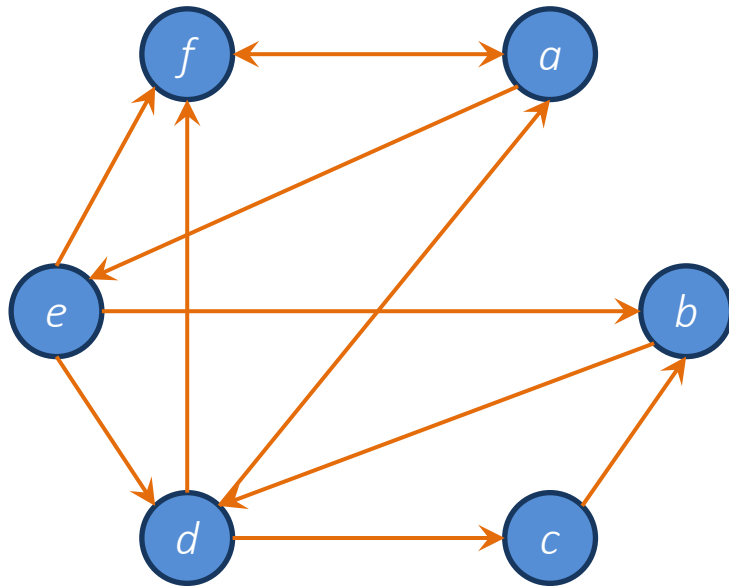
= someone surfing the web, clicking links randomly



- 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 ( $d = 0.85$ )

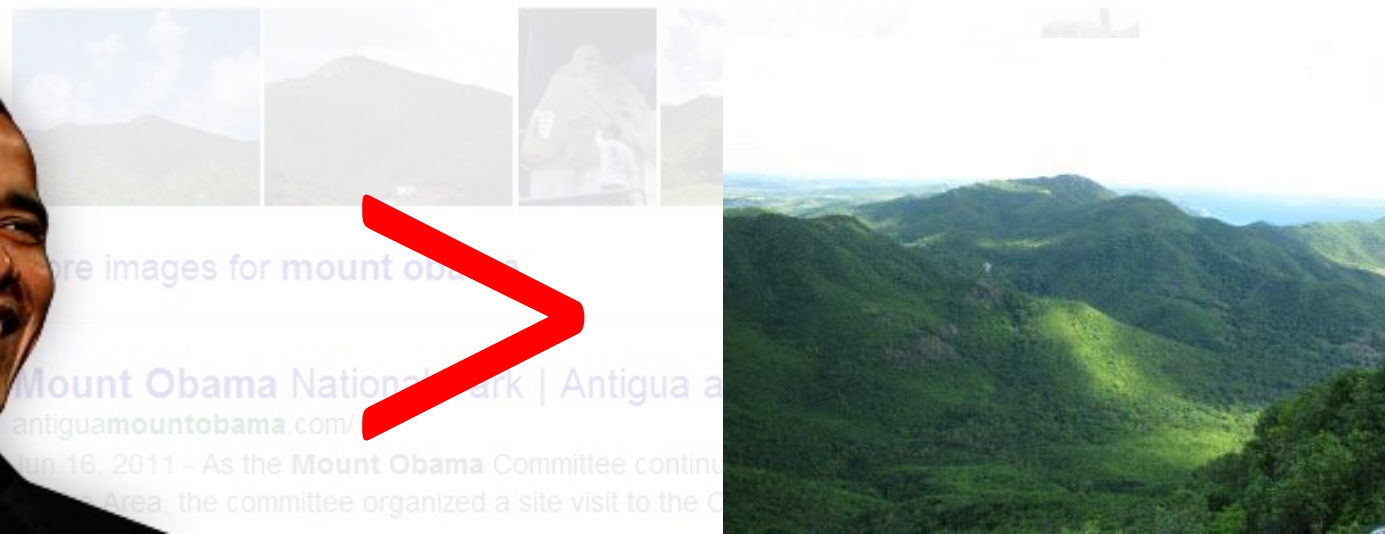
$$\text{rank}_i(v) := d \times \sum_{u \in \text{in}(v)} \frac{\text{rank}_{i-1}(u)}{|\text{out}(u)|} + \sum_{v' \in V'} \frac{\text{rank}_{i-1}(v')}{|V|} + (1-d) \times \sum_{v'' \in V''} \frac{\text{rank}_{i-1}(v'')}{|V|}$$

# 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



# COMPUTING PAGERANK AT SCALE

**Distributed Static  
Data Processing**

**Distributed Dynamic  
Data Processing**

**Distr. Unstructured  
Data Management**

**Distr. (Semi-)structured  
Data Management**

**Distributed Data Processing**

**Distributed Data Management**

**Distributed Systems**

**Local Data Processing**

# Graph Parallel Frameworks: Pregel

## Pregel: A System for Large-Scale Graph Processing

Grzegorz Malewicz, Matthew H. Austern, Aart J. C. Bik, James C. Dehnert, Ilan Horn,  
Naty Leiser, and Grzegorz Czajkowski  
Google, Inc.  
{malewicz,austern,ajcbik,dehnert,ilan,naty,gczaj}@google.com

### ABSTRACT

Many practical computing problems concern large graphs. Standard examples include the Web graph and various social networks. The scale of these graphs—in some cases billions of vertices, trillions of edges—poses challenges to their efficient processing. In this paper we present a computational model suitable for this task. Programs are expressed as a sequence of iterations, in each of which a vertex can receive messages sent in the previous iteration, send messages to other vertices, and modify its own state and that of its outgoing edges or mutate graph topology. This vertex-centric approach is flexible enough to express a broad set of algorithms. The model has been designed for efficient, scalable and fault-tolerant implementation on clusters of thousands of commodity computers, and its implied synchronicity makes reasoning about programs easier. Distribution-related details are hidden behind an abstract API. The result is a framework for processing large graphs that is expressive and easy to program.

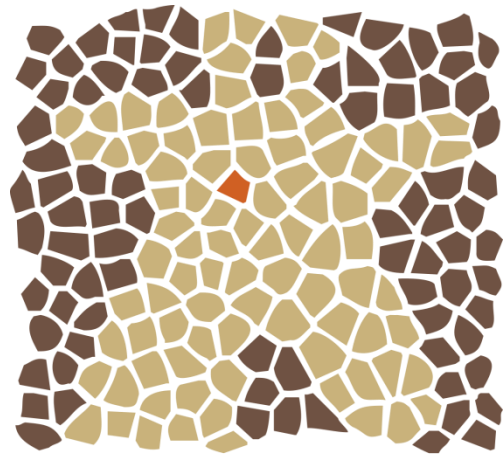
disease outbreaks, or citation relationships among published scientific work—have been processed for decades. Frequently applied algorithms include shortest paths computations, different flavors of clustering, and variations on the page rank theme. There are many other graph computing problems of practical value, *e.g.*, minimum cut and connected components.

Efficient processing of large graphs is challenging. Graph algorithms often exhibit poor locality of memory access, very little work per vertex, and a changing degree of parallelism over the course of execution [31, 39]. Distribution over many machines exacerbates the locality issue, and increases the probability that a machine will fail during computation. Despite the ubiquity of large graphs and their commercial importance, we know of no scalable general-purpose system for implementing arbitrary graph algorithms over arbitrary graph representations in a large-scale distributed environment.

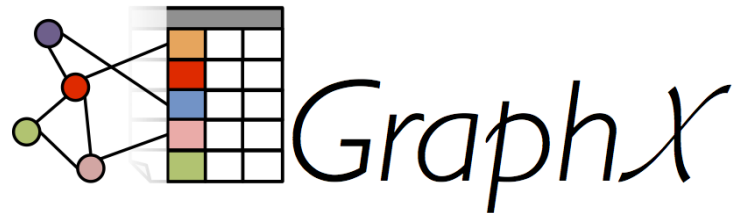
Implementing an algorithm to process a large graph typically means choosing among the following options:



# Graph Parallel Frameworks: Open Source

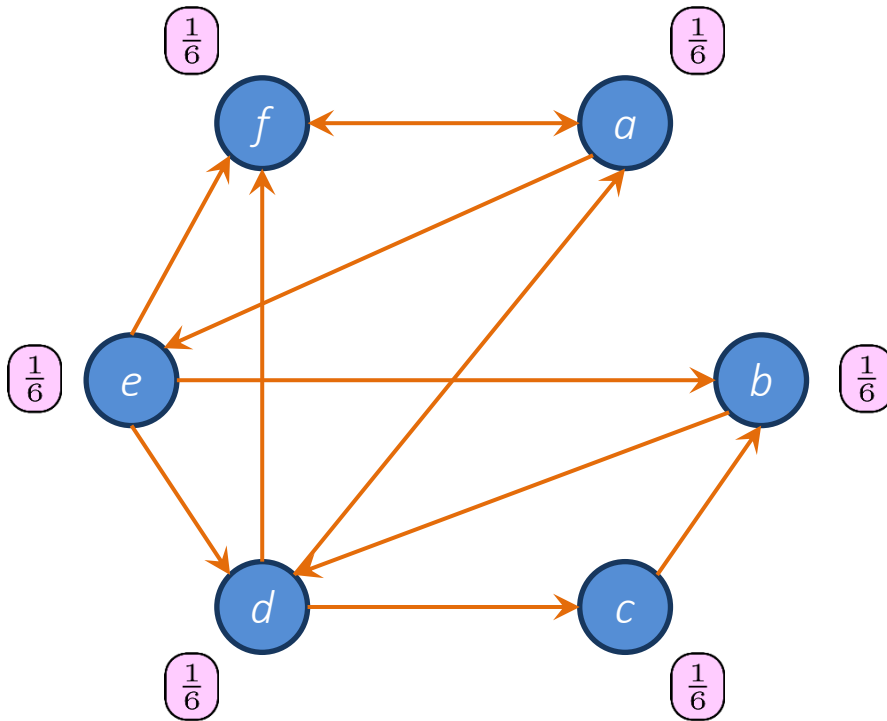


A P A C H E  
G I R A P H



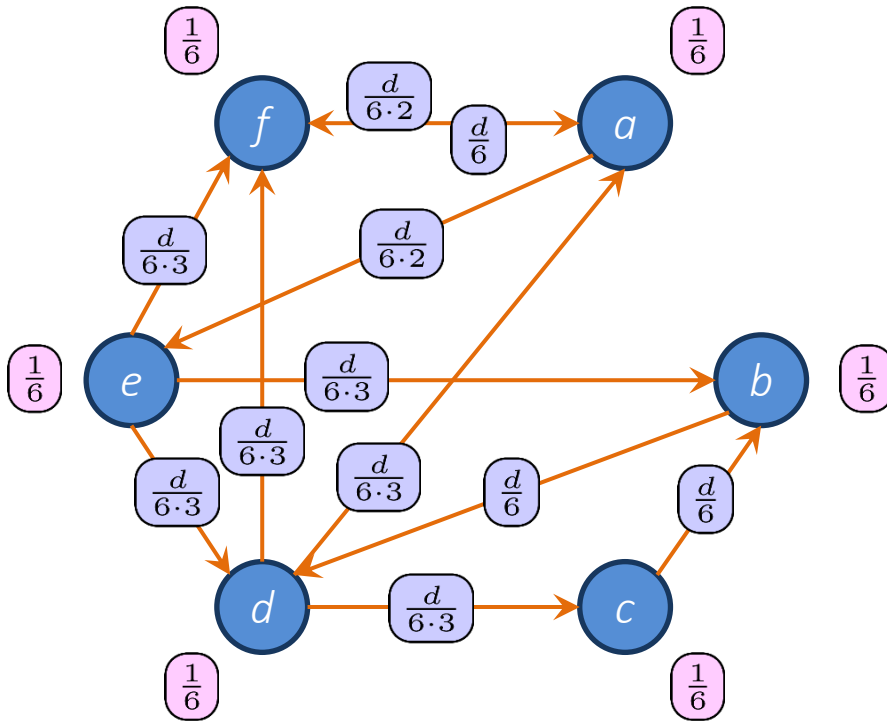
# Vertex-Centric Computation

1. Nodes assigned **state**

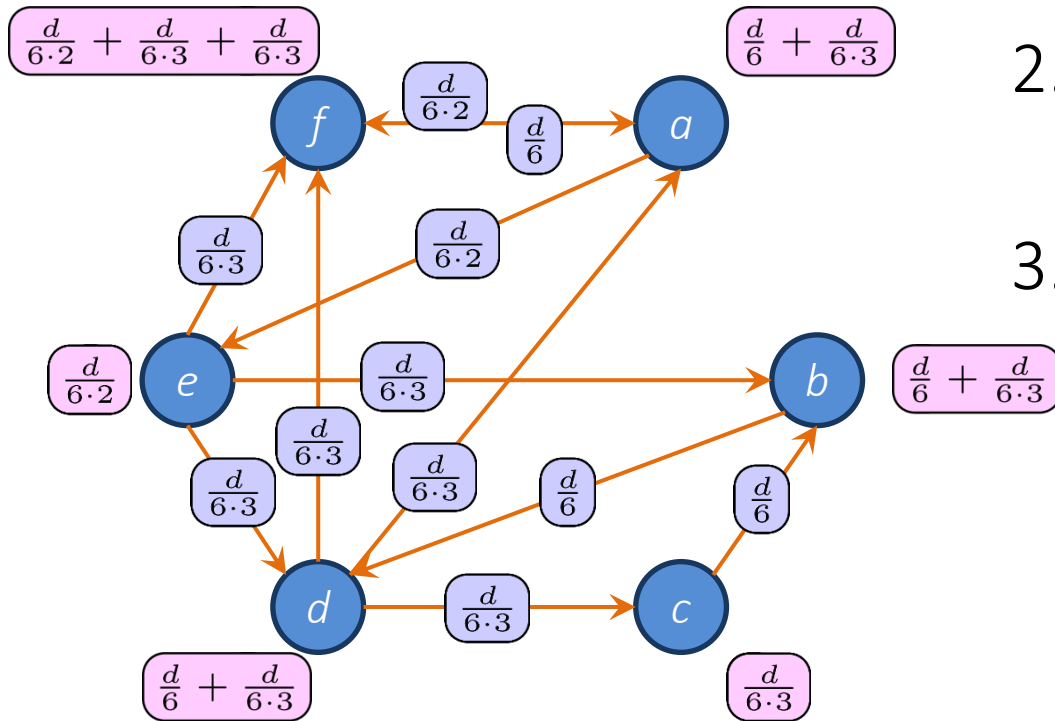


# Vertex-Centric Computation

1. Nodes assigned state
2. Nodes pass messages (typically along edges)

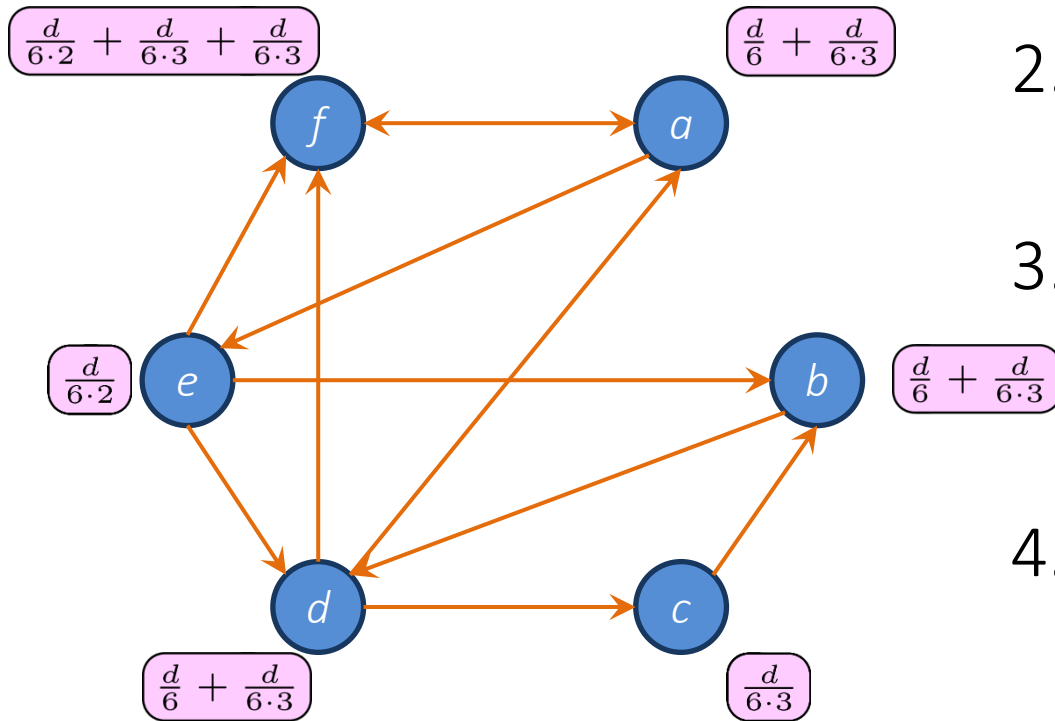


# Vertex-Centric Computation



1. Nodes assigned state
2. Nodes pass messages (typically along edges)
3. Nodes aggregate messages received and update state

# Vertex-Centric Computation



1. Nodes assigned state
2. Nodes pass messages (typically along edges)
3. Nodes aggregate messages received and update state
4. GOTO 2. until some termination criteria are reached

# Vertex-Centric Computation: Other Features

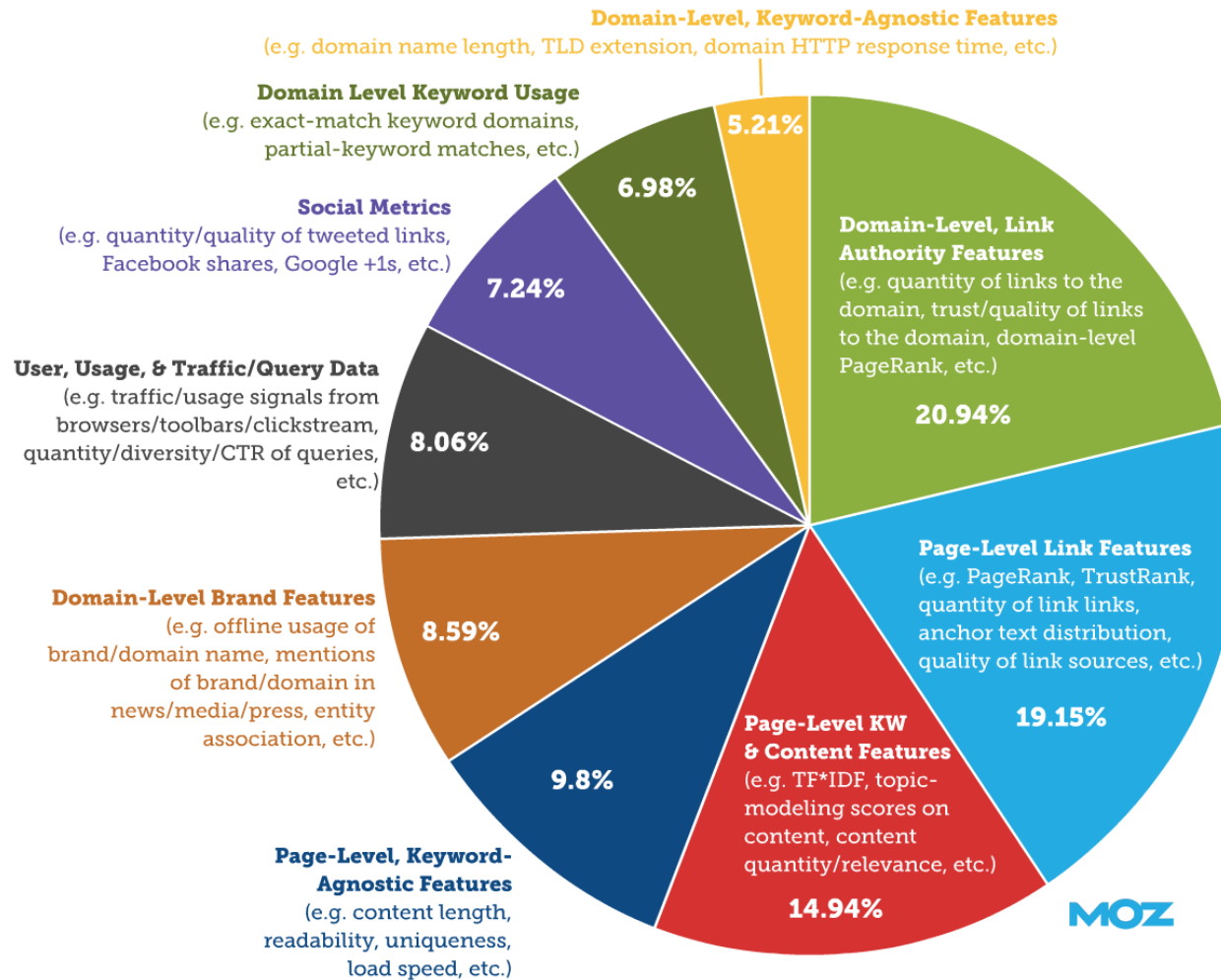
- Message passing and aggregation done in parallel
- Optional message passing to non-neighbours
- Optional global “aggregation” phase
- Optional changes to the graph topology

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)



*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%



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?