

CC5212-1

PROCESAMIENTO MASIVO DE DATOS

OTOÑO 2016

Lecture 8: Information Retrieval II

Aidan Hogan

aidhog@gmail.com

How does Google crawl the Web?



Inverted Indexing

1



Fruitvale Station

From Wikipedia, the free encyclopedia



1 10 18 21 23 28 37 43 47 55 59 68 71 76
Fruitvale Station is a 2013 American [drama film](#) written and directed by [Ryan Coogler](#).

Inverted index:

Term List	Posting Lists
a	(1,[21,96,103,...]), (2,[...]), ...
american	(1,[28,123]), (5,[...]), ...
and	(1,[57,139,...]), (2,[...]), ...
by	(1,[70,157,...]), (2,[...]), ...
directed	(1,[61,212,...]), (4,[...]), ...
drama	(1,[38,87,...]), (16,[...]), ...
...	...

INFORMATION RETRIEVAL: RANKING

How Does Google Get Such Good Results?

Google  

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

About 1,150,000 results (0.28 seconds)

Dr. Aidan Hogan | DERI
www.deri.ie/users/aidan-hogan ▾
Tel: +353 91 495723. [aidan \[dot\] hogan \[at\] deri \[dot\] org](mailto:aidan@deri.org). Homepage. Aidan worked with DERI Galway as Postdoctoral Researcher from to ...

dblp: Aidan Hogan
www.informatik.uni-trier.de/~ley/pers/hy/h/HoganAidan.html ▾
Mar 7, 2014 - Emir Muñoz, Aidan Hogan, Alessandra Mileo: Using linked data to ...
Patrick O'Byrne, Aidan Hogan: Exploring the Dynamics of Linked Data.

Aidan Hogan's Homepage
aidanhogan.com/ ▾
Aidan Hogan, Antoine Zimmermann, Jürgen Umbrich, Axel Polleres and Stefan Decker. "Scalable and Distributed Methods for Entity Matching, Consolidation ...
[journal](#) - [book-chapter](#) - [conference](#) - [workshop](#)



Two Sides to Ranking: Relevance

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 \(disambiguation\)](#) - [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 Area, the committee organized a site visit to the O...

**RANKING:
RELEVANCE**

Example Query



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 screenshot shows the Wikipedia page for 'Braveheart'. The browser's address bar displays 'en.wikipedia.org/wiki/Braveheart'. A search bar at the top right contains the text 'freedom', and a red box highlights the search results '3 of 7'. The page content includes the Wikipedia logo, navigation links (Main page, Contents, Featured content, Current events, Random article, Donate to Wikipedia, Wikimedia Shop), and a sidebar with 'Interaction' and 'Tools' sections. The main article text describes 'Braveheart' as a 1995 epic historical drama war film directed by and starring Mel Gibson. A movie poster for 'Braveheart' is also visible.

freedom

- 7 occurrences

Matches in a Document

← → ↻ en.wikipedia.org/wiki/Braveheart 🔍 ☆ 🔒 ☰

movie 1 of 16 ^ v x

Article **Talk** Read Edit View history Search 🔍

Braveheart

From Wikipedia, the free encyclopedia

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

Braveheart is a 1995 [epic historical drama 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 based on [Blind Harry's epic poem *The Actes and Deidis of the Illustre and Vallyeant*](#)



Every man dies, but every man truly lives.

freedom

- 7 occurrences

movie

- 16 occurrences

Matches in a Document



The screenshot shows the Wikipedia article for "Braveheart". The browser address bar displays "en.wikipedia.org/wiki/Braveheart". A search bar at the top right shows the search term "wallace" with "43 of 88" matches. The article title "Braveheart" is prominently displayed. The text describes it as a 1995 epic historical drama war film directed by and starring Mel Gibson. It mentions that 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 based on Blind Harry's epic poem "The Actes and Deidis of the Illustre and Vallyeant". To the right of the text is a movie poster for "Braveheart" featuring Mel Gibson. The left sidebar contains navigation links such as "Main page", "Contents", "Featured content", "Current events", "Random article", "Donate to Wikipedia", and "Wikimedia Shop".

freedom

- 7 occurrences

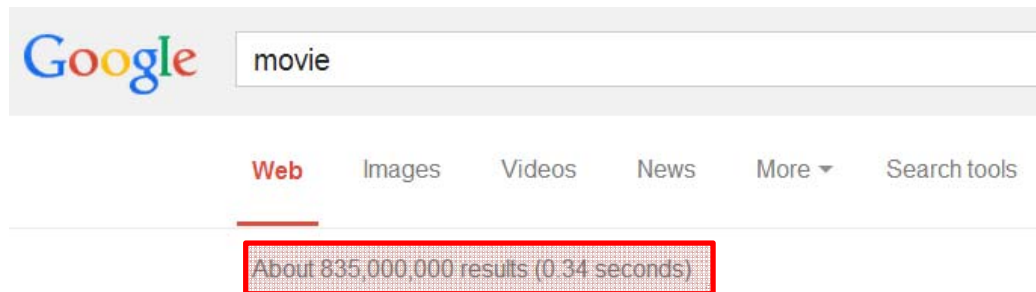
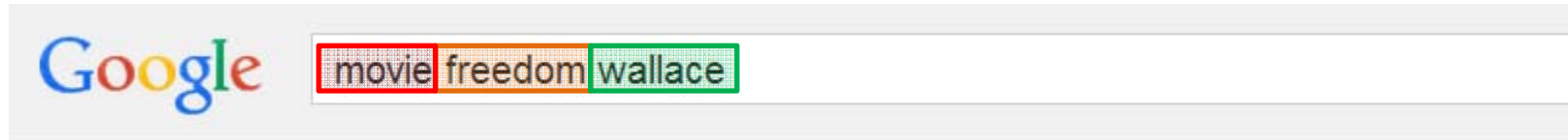
movie

- 16 occurrences

wallace

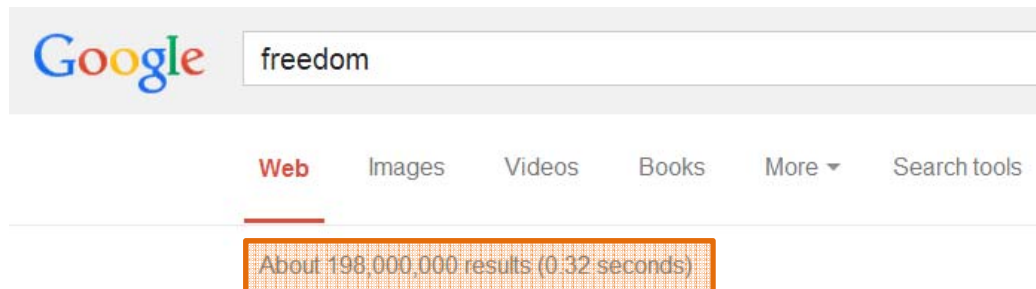
- 88 occurrences

Usefulness of Words



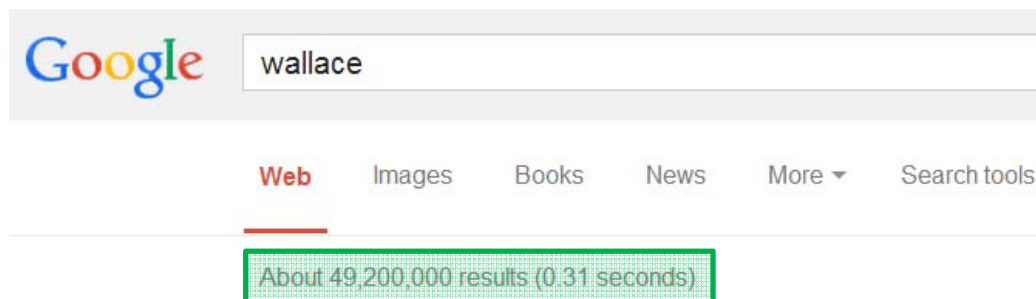
movie

- occurs very frequently



freedom

- occurs frequently



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*

- $\text{tf}(t, d)$... various options

- Raw count of occurrences

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

- Logarithmically scaled

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

- Normalised by document length

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

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

- A combination / something else ☺

Relevance Measure: TF–IDF

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

$$\text{idf}(t, D) = \log\left(\frac{|D|}{|\{d' \in D : t \in d'\}| + 1}\right)$$

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

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

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

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	
freedom	7	198,000,000	
wallace	43	49,200,000	

The image shows a Google search bar with the text 'the'. Below the search bar, the 'Web' tab is selected, and the search results are displayed. The search results show 'About 11,410,000,000 results (0.27 seconds)'. To the right of the search results, the value $|D| = 11,410,000,000$ is displayed.

Relevance Measure: TF-IDF



Term Frequency

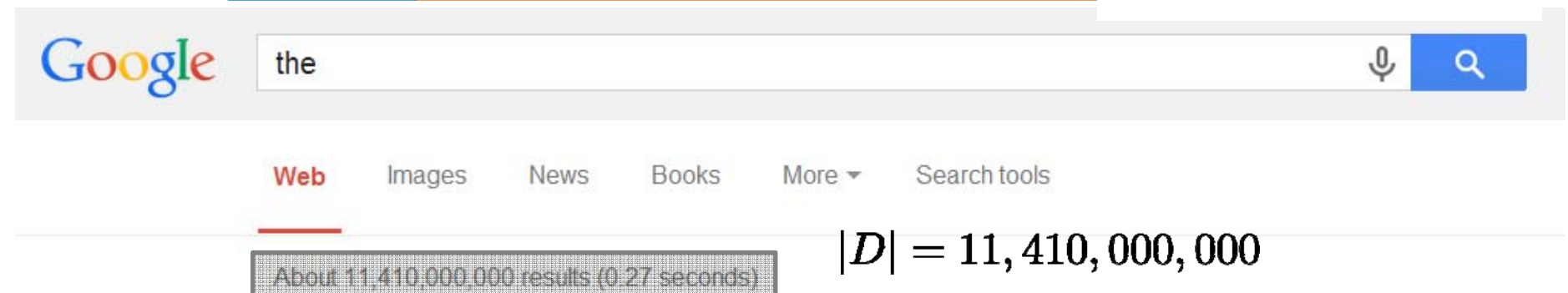
$$\text{tf}(t, d) = \text{count}(t, d)$$

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

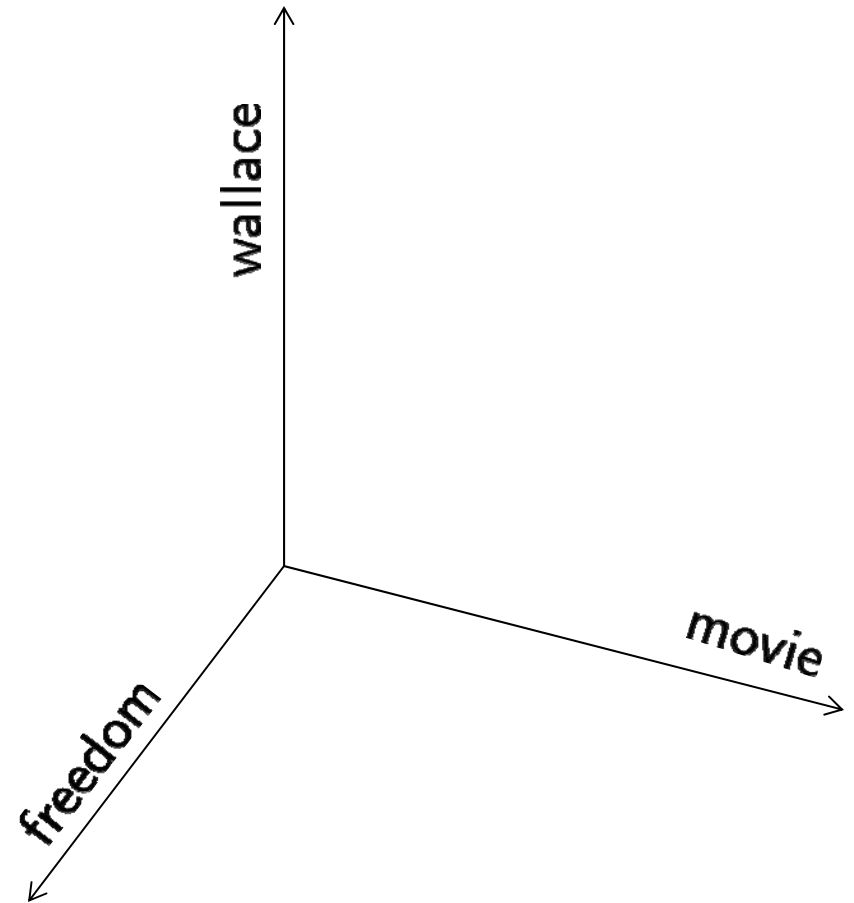
$$\text{score}(q, d) = \sum_{t \in q} \text{tf-idf}(t, d)$$

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

Vector Space Model (a mention)

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

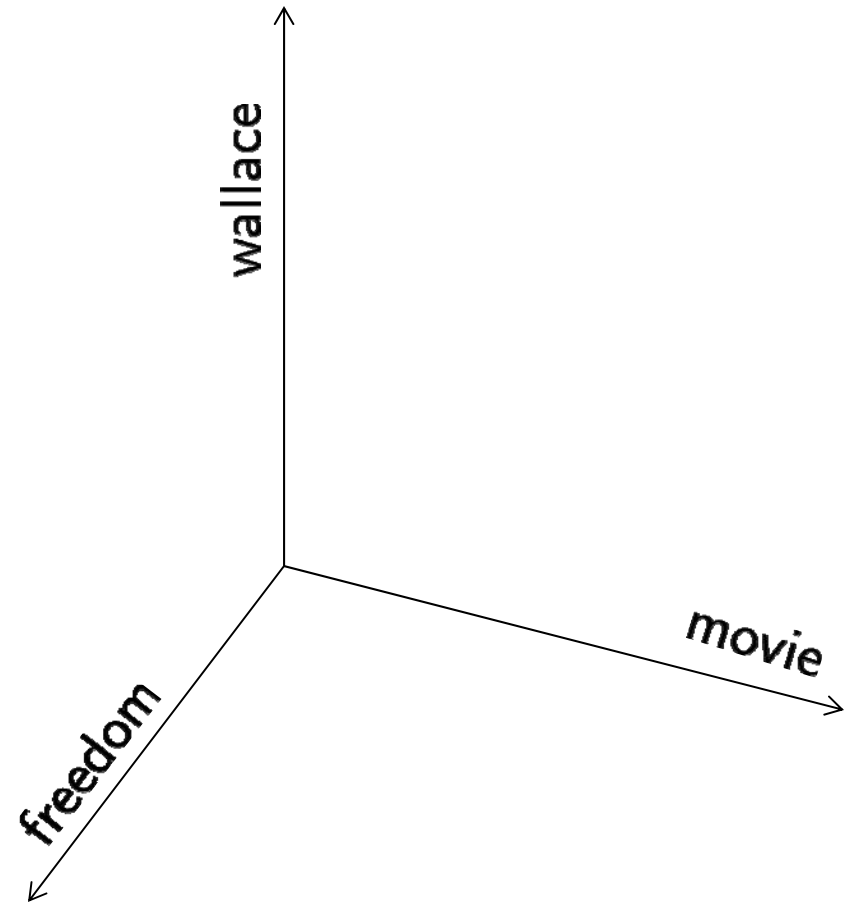
$$l = \sqrt{\sum_{t \in q} \text{tf}(t, d)^2}$$



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

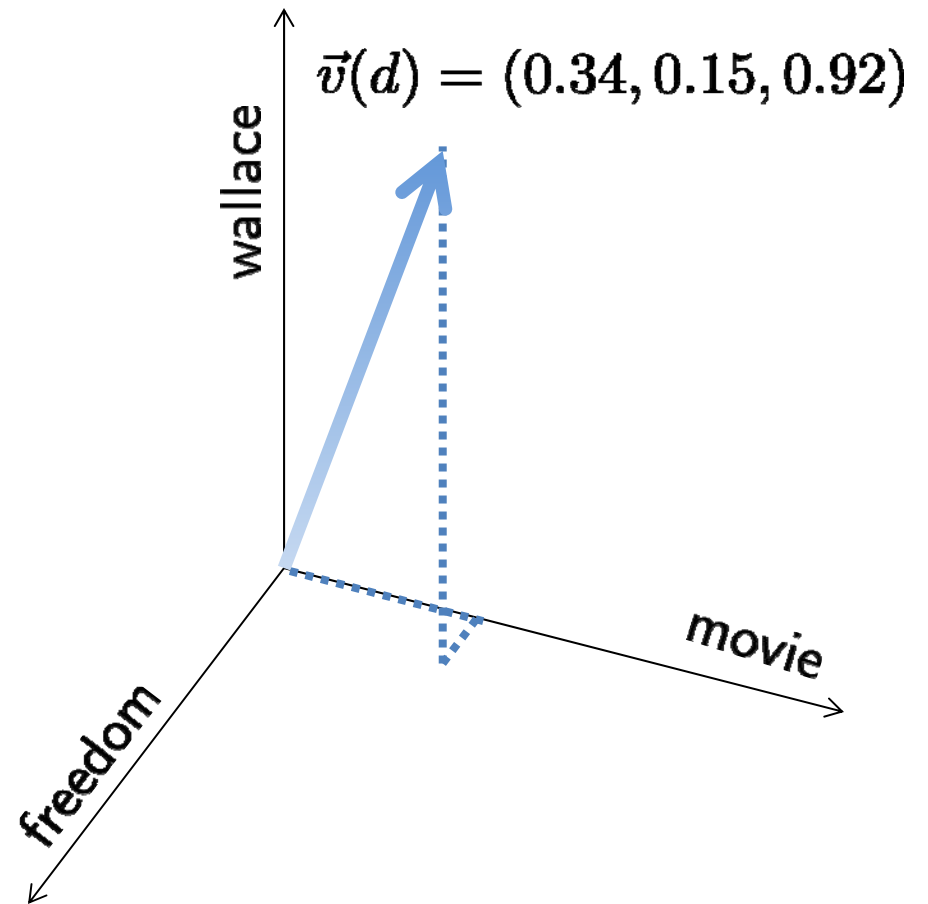
$$l = \sqrt{\sum_{t \in q} \text{tf}(t, d)^2}$$



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 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

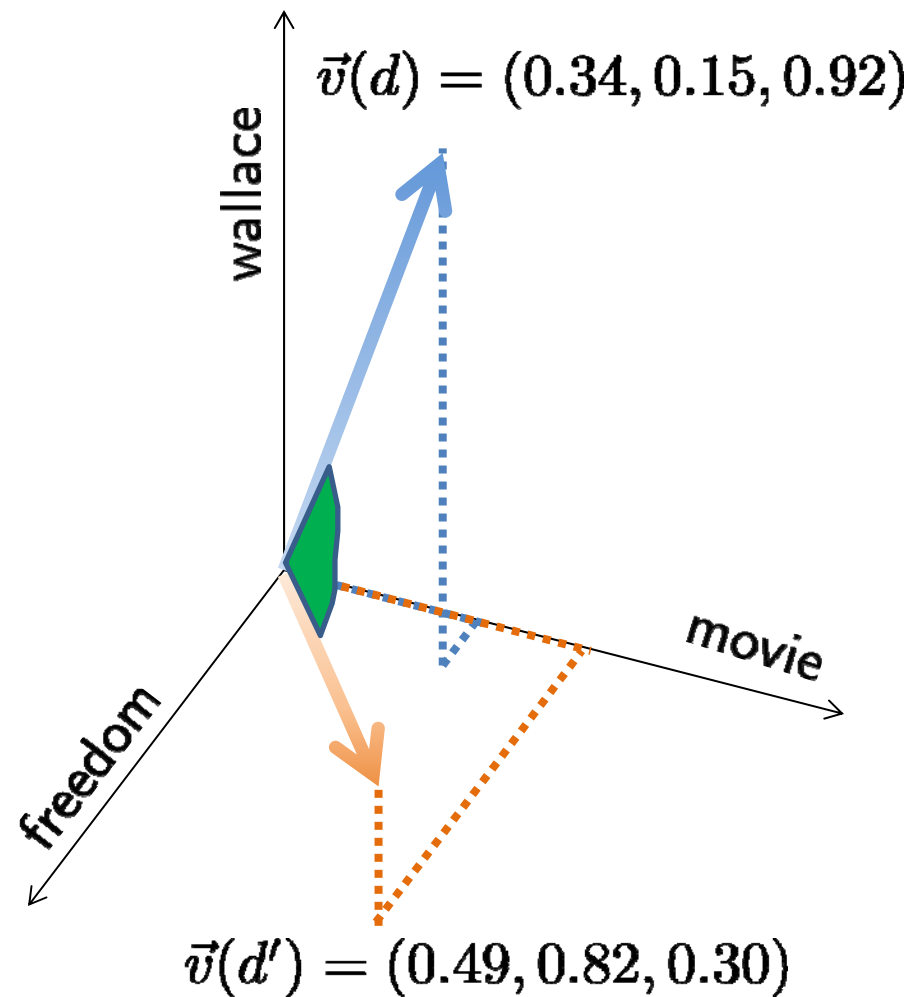
$$\text{sim}(d, d') \approx 0.57 \quad \Sigma$$

- 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



Two Sides to Ranking: Relevance

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 \(disambiguation\)](#) - [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 ...



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>
```



WIKIPEDIA
The Free Encyclopedia

[Main page](#)
[Contents](#)
[Featured content](#)
[Current events](#)
[Random article](#)
[Donate to Wikipedia](#)
[Wikimedia Shop](#)

▼ Interaction
[Help](#)
[About Wikipedia](#)
[Community portal](#)

[Article](#) [Talk](#) [Read](#) [View source](#) [View history](#)

Create account [Log in](#)

Barack Obama

From Wikipedia, the free encyclopedia

"Obama" redirects here. For other uses, see [Obama \(disambiguation\)](#).

This article is about the 44th president of the United States. For his father, see [Barack Obama, Sr.](#)

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](#)

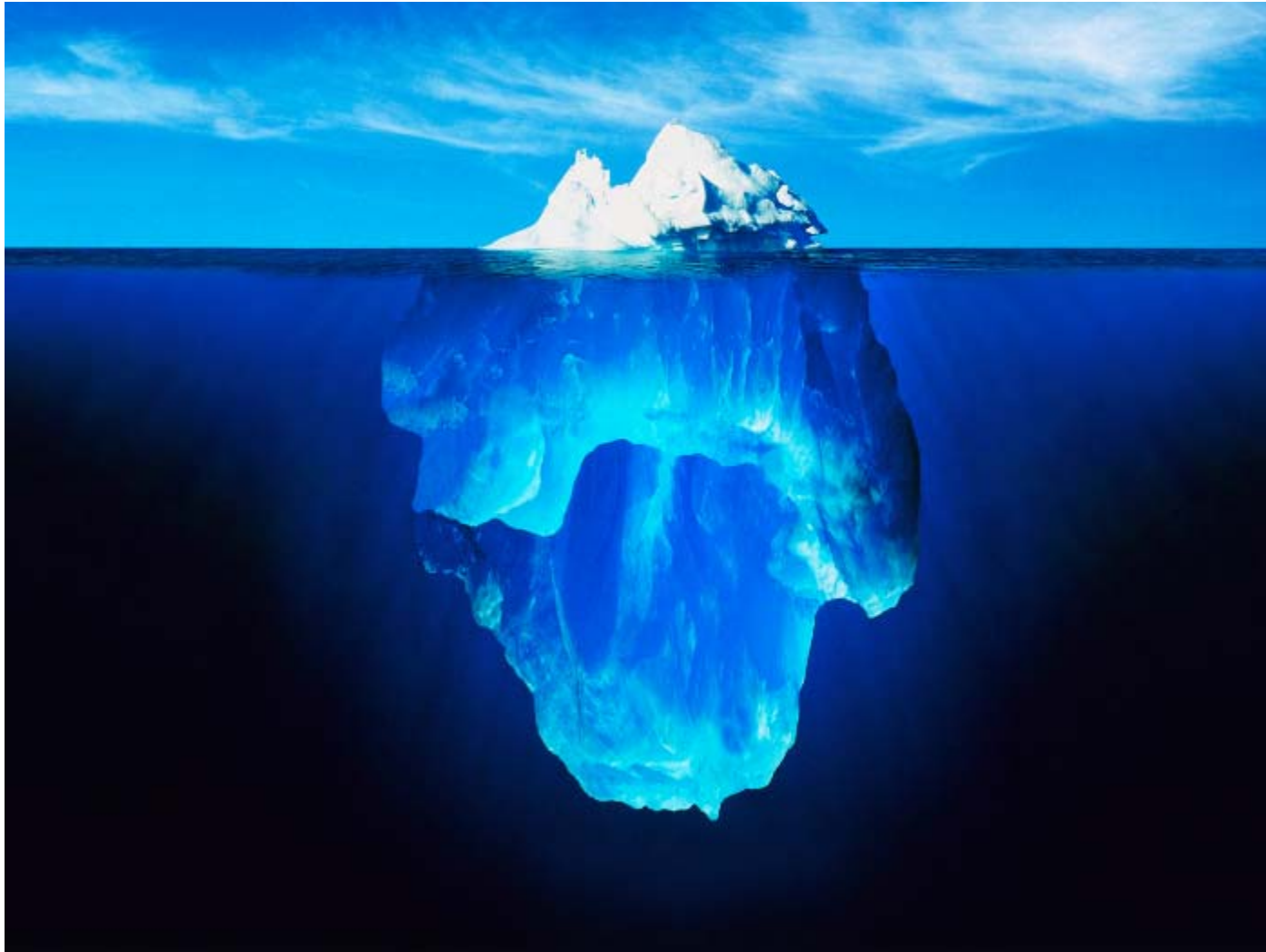


Anchor Text

- 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>
```

Information Retrieval & Relevance



**RANKING:
IMPORTANCE**

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 Area, the committee organized a site visit to the O

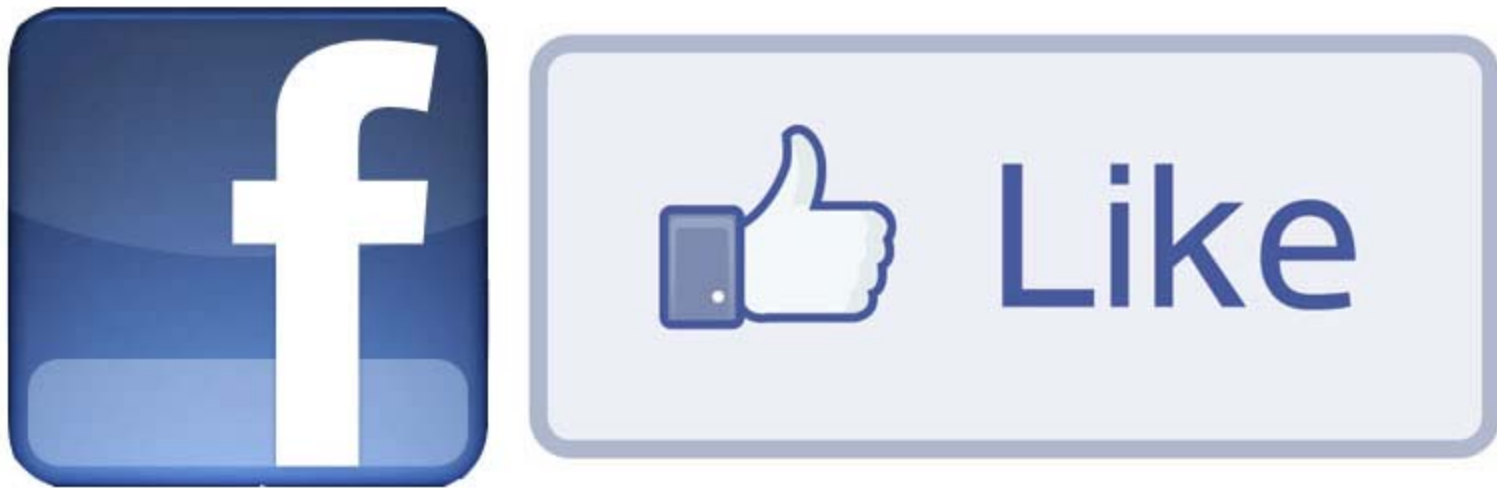
Link Analysis

Which will have more links:
Barack Obama's Wikipedia Page or
Mount Obama's Wikipedia Page?



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 the number of inlinks a web-page receives we know its importance, right?



Link Spamming



semanticweb.com™

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

Home Events Media Industry Verticals Answers

Questions Tags Users Badges

[deleted] Kala Jadu Specialist +9196



black magic specialist baba ji call now +919610897260



http://www.blackmagicspecialist.net.in



java

edit | close | undelete | more ▼

Claritin Clomid Combivent Confido Copegus Cordarone Coreg Coumadin Cozaar Crestor
Cyklokapron Cymbalta Cystone Cytotec Danazol Deltasone Depakote Desyrel Detrol Diabecon
Diakof Diarex Didronel Differin Dilantin Diovan Dostinex Elavil Elimate Emsam Endep Eurax
Evecare Evista Exelon Famvir Feldene Femara Femcare Flomax Flonase Flovent Fosamax Gasex
Geodon Geriforte Herbolax High Love Hincocid Himcolin Hincospaz 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 Mycelelex-G Mysoline Naprosyn Neurontin Nicotinell Nimotop Nirdosh Nizoral
Nolvadex Nonoxinol Noroxin Omnicef Ophthacare Oxytrol Pamelor Parlodel Paxil Penisole
Pheniramine Pilex Plan B Plavix Plendil Pletal Prandin Pravachol Prednisone Premarin Prevacid
Prilosec Prinivil Procardia Prograf Prometrium Propecia Proscar Protonix Proventil Prozac Purim
Purinethol Quibron-T Relafen Renalka Reosto Requip Retin-A Revia 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 Rhythmol 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 Dehux
Handheld Plasma Whitening Tool Dehux Whitening System With Plasma Lamp Coral Calcium
Cialis Jelly Cialis Soft Tabs Calcium Carbonate Bust Enhancer Beconase AQ Anatrium 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

Link Importance

Which is more “important”: a link from Barack Obama’s Wikipedia page or a link from buyv1agra.com?



PageRank



PageRank

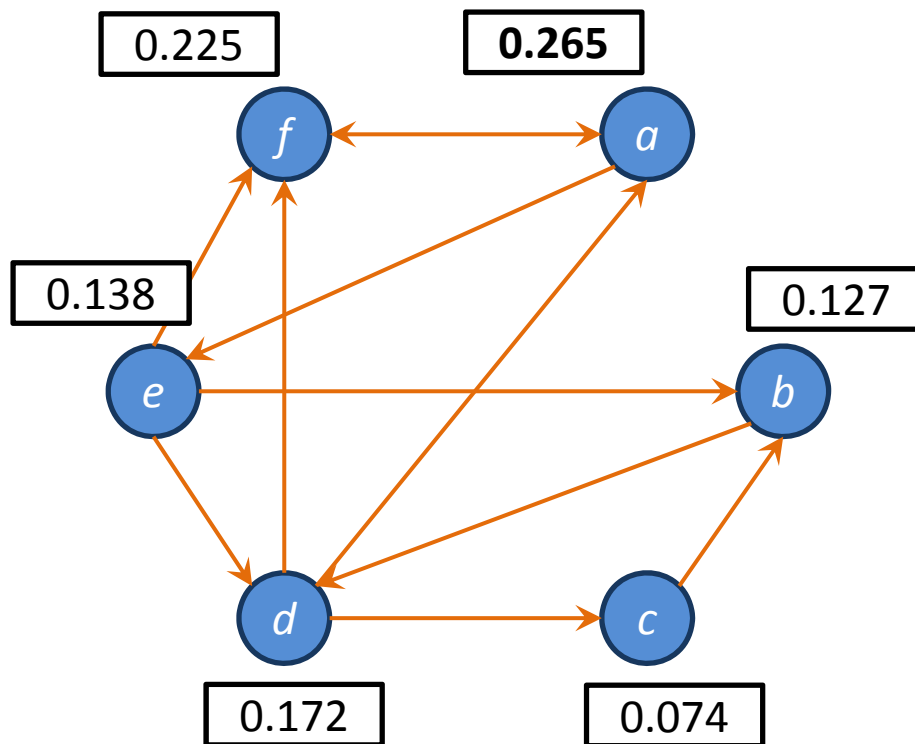
- Not just a count of inlinks
 - A link from a more important page is more important
 - A link from a page with fewer links is more important
 - ∴ A page with lots of inlinks from important pages (which have few outlinks) is more important

PageRank is Recursive



PageRank Model

- The Web: a directed graph



$$G = (V, E)$$

Vertices
(pages)

Edges
(links)

Which is the most
“important” vertex?

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

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

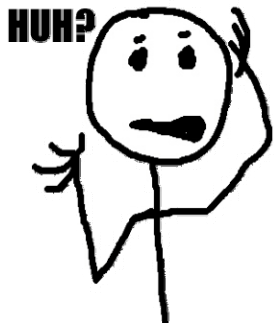
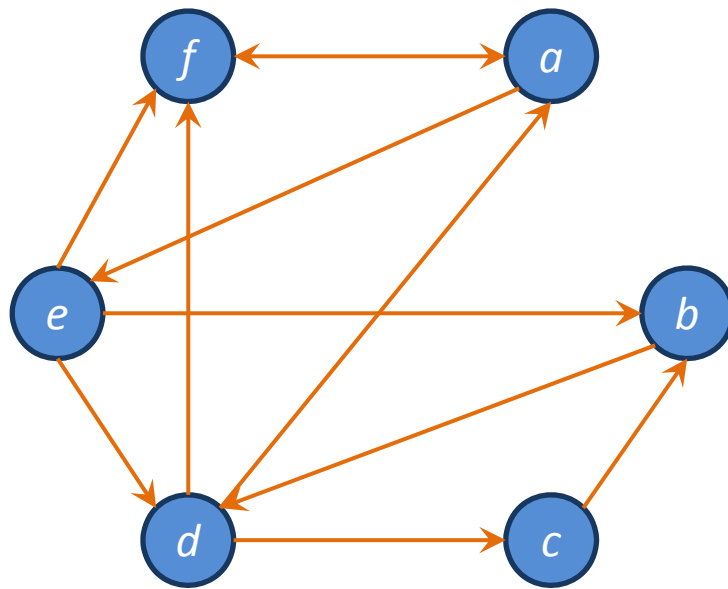
PageRank Model

- The Web: a directed graph

$$G = [V, E]$$

Vertices
(pages)

Edges
(links)



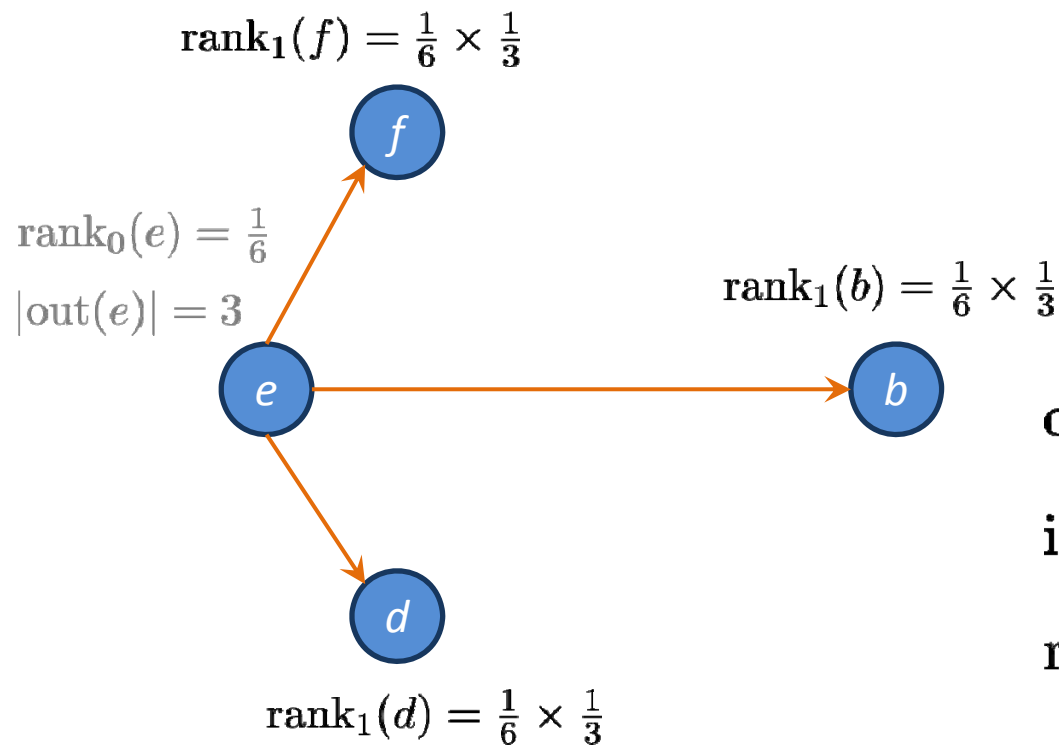
$$\text{out}(v) \doteq \{v' \in V \mid (v, v') \in E\}$$

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

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

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

PageRank Model



$$G = [V, E]$$

Vertices
(pages)

Edges
(links)

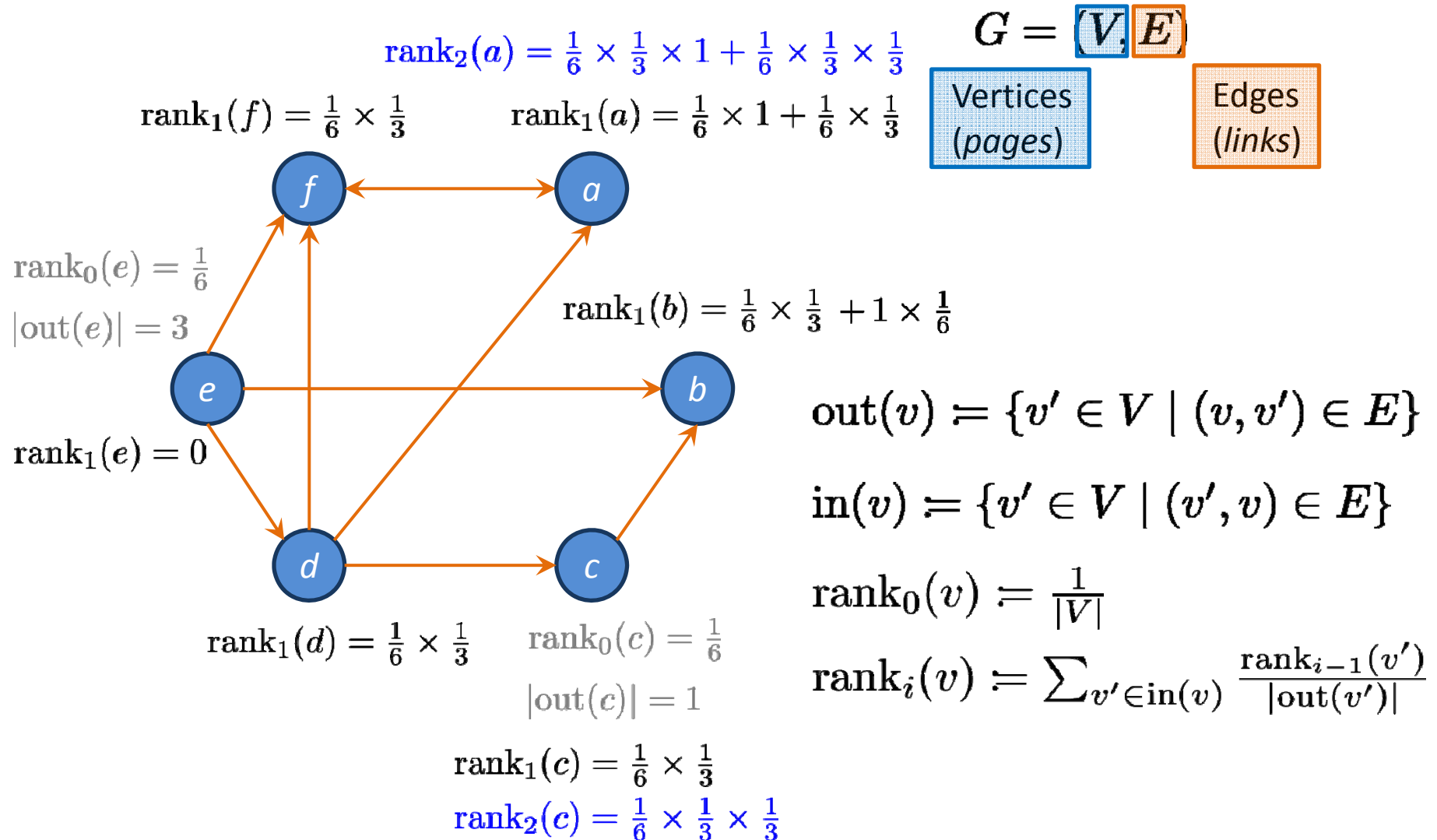
$$\text{out}(v) \doteq \{v' \in V \mid (v, v') \in E\}$$

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

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

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

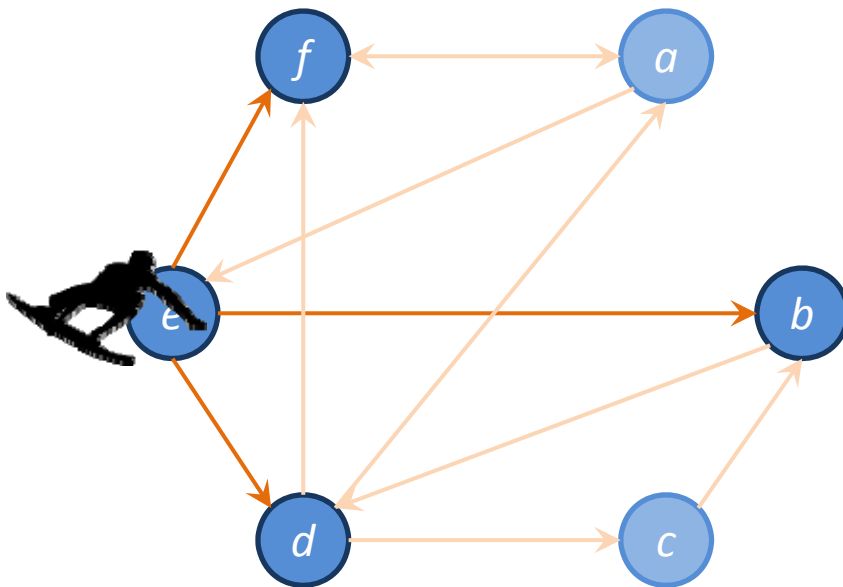
PageRank Model



PageRank: Random Surfer Model



= someone surfing the web,
clicking links randomly

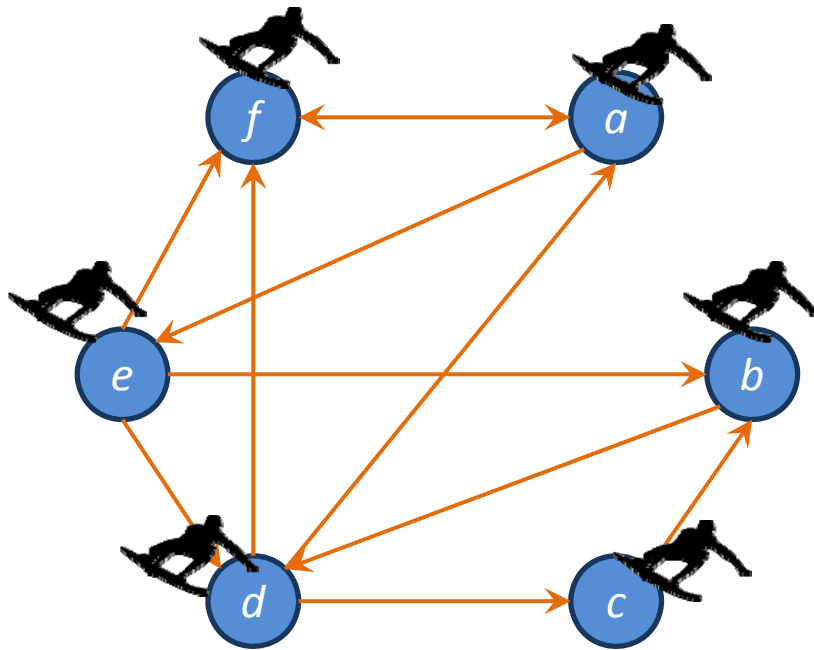


- What is the probability of being at page *x* after *n* hops?

PageRank: Random Surfer Model



= someone surfing the web,
clicking links randomly

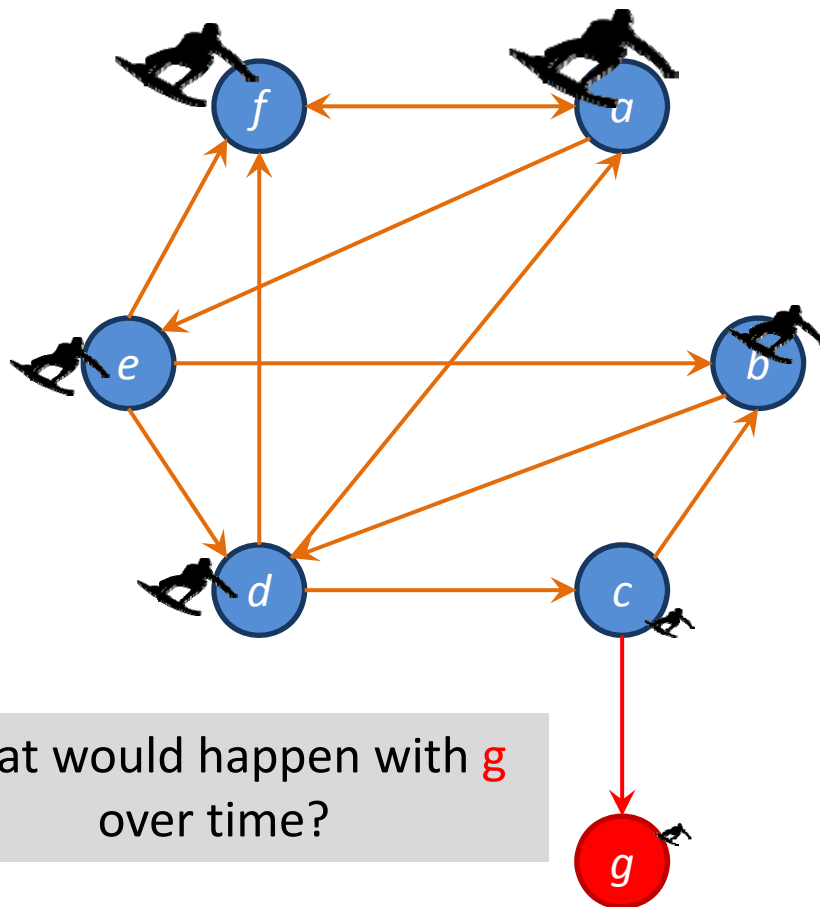


- What is the probability of being at page x after n hops?
- *Initial state:* surfer equally likely to start at any node

PageRank: Random Surfer Model



= someone surfing the web,
clicking links randomly



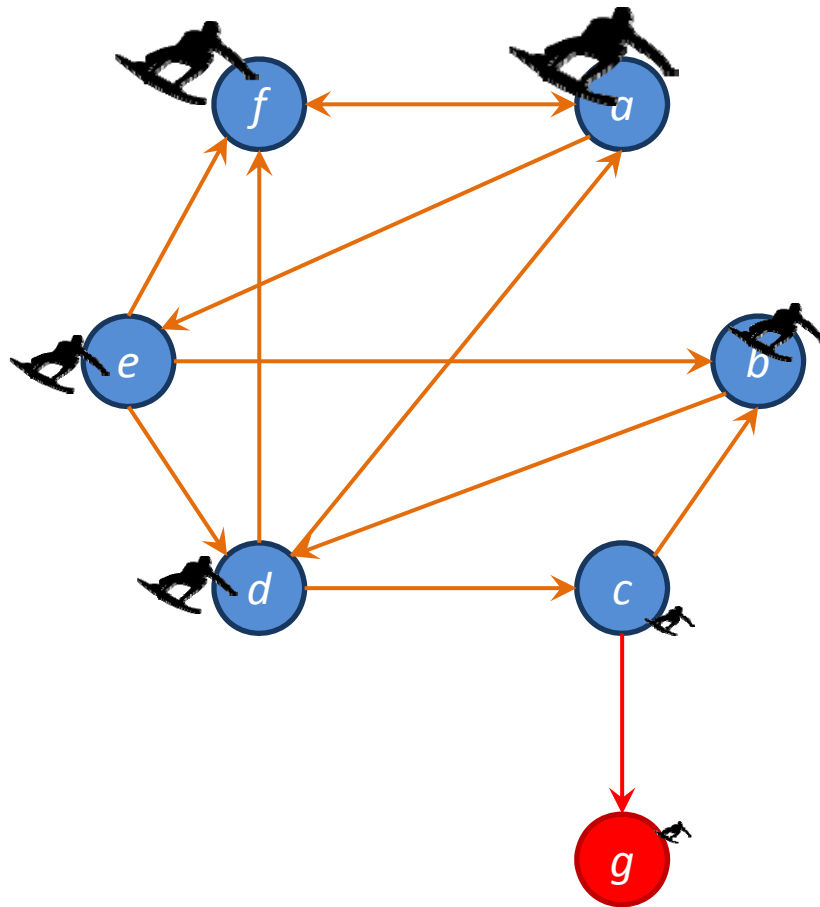
What would happen with **g**
over time?

- 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

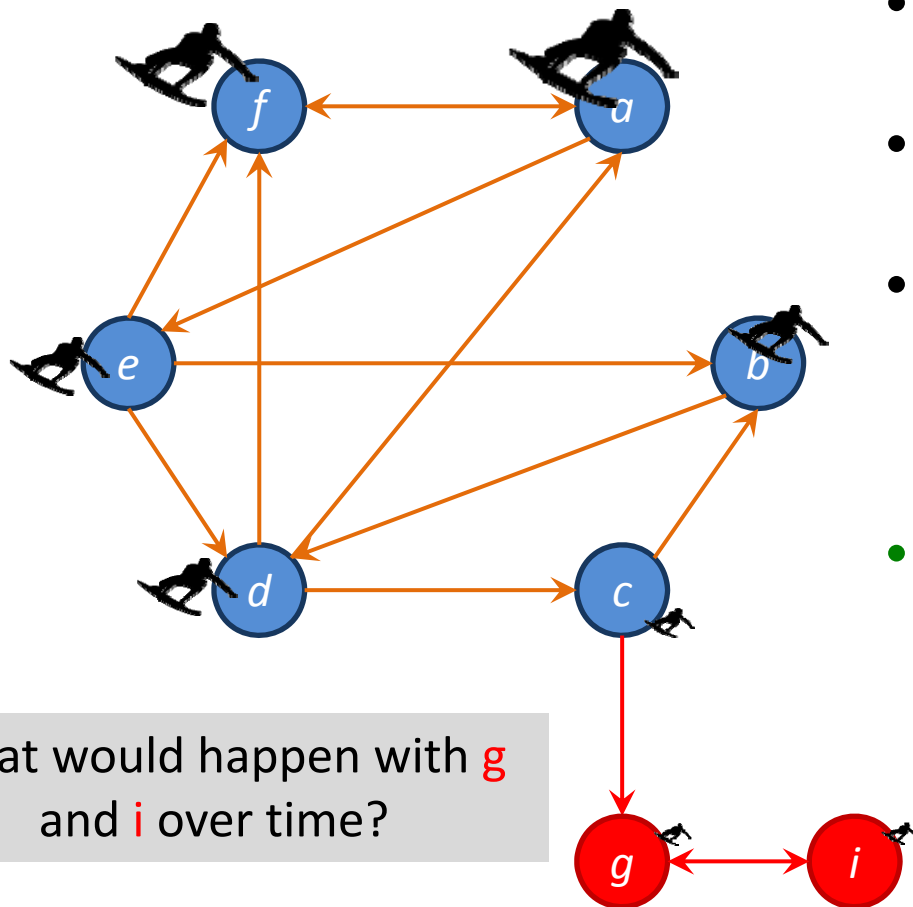


- What is the probability of being at page x after n hops?
- *Initial state*: surfer equally likely to start at any node
- PageRank applied iteratively for each hop: score indicates probability of being at that page after than many hops
- If the surfer reaches a page without links, the surfer randomly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

PageRank: Random Surfer Model



= someone surfing the web,
clicking links randomly



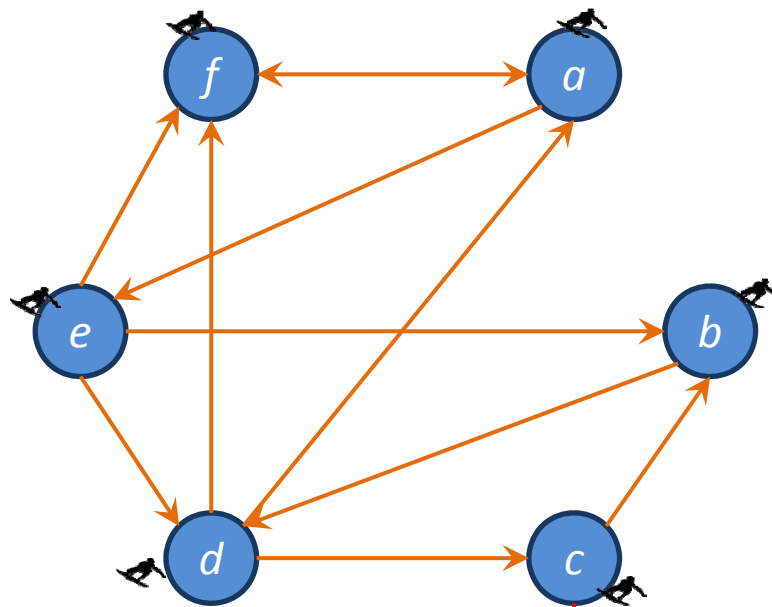
What would happen with **g**
and **i** over time?

- 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 than many hops
- If the surfer reaches a page without links, the surfer randomly jumps to another page

PageRank: Random Surfer Model

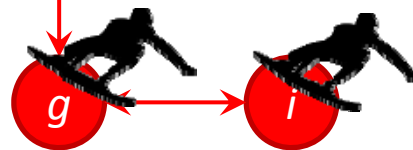


= someone surfing the web,
clicking links randomly



- 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 links, the surfer randomly jumps to another page

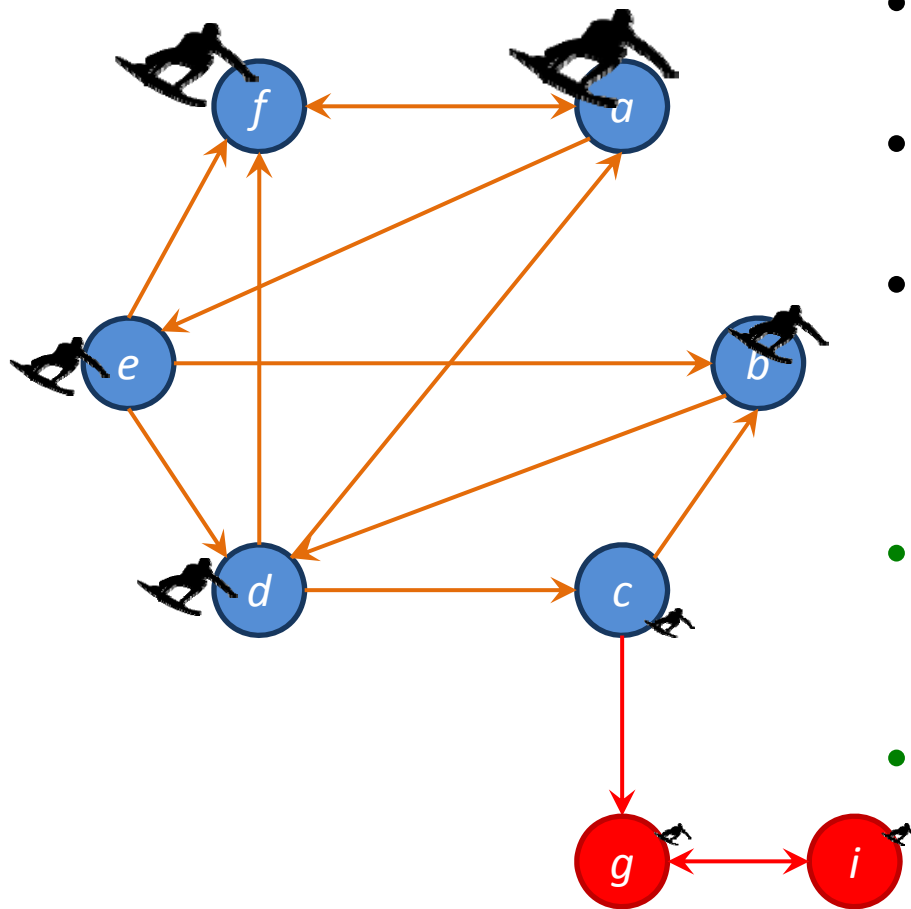
What would happen with **g** and **i** over time?



PageRank: Random Surfer Model



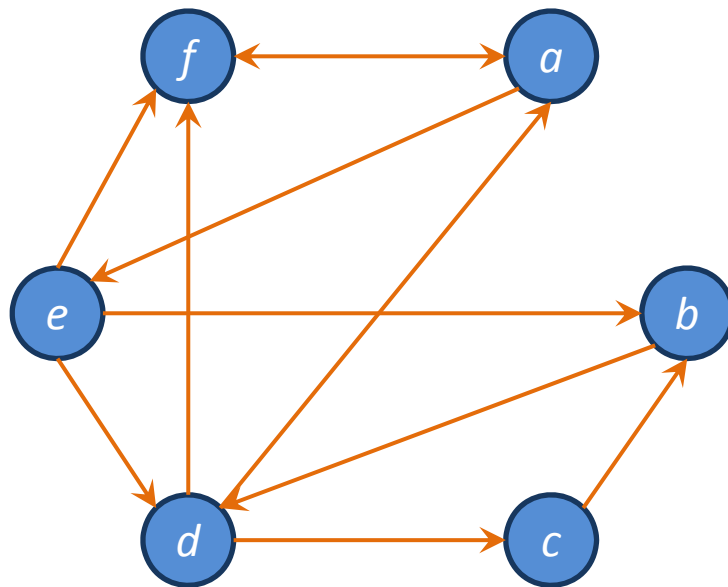
= someone surfing the web,
clicking links randomly



- 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 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



$$G = [V, E]$$

Vertices
(pages)

Edges
(links)

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

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

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

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

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

d is the dampening factor

typically ($d = 0.85$)

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

PageRank: Benefits

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



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 Area, the committee organized a site visit to the O...

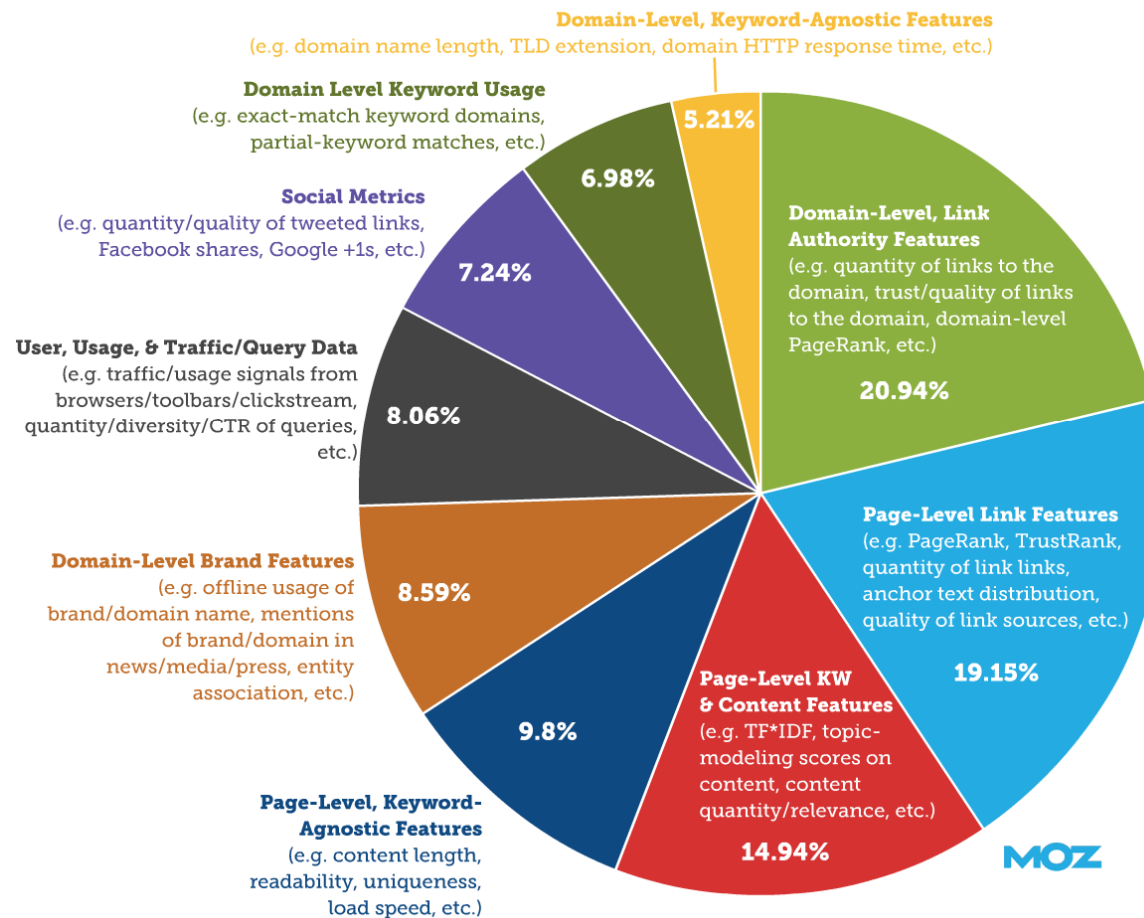


GOOGLE: A GUESS

How modern Google ranks (maybe, who knows, I don't)

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

INFORMATION RETRIEVAL: RECAP

How Does Google Get Such Good Results?

Google  

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

About 1,150,000 results (0.28 seconds)

Dr. Aidan Hogan | DERI
www.deri.ie/users/aidan-hogan ▾
Tel: +353 91 495723. [aidan \[dot\] hogan \[at\] deri \[dot\] org](mailto:aidan@deri.org). Homepage. Aidan worked with DERI Galway as Postdoctoral Researcher from to ...

dblp: Aidan Hogan
www.informatik.uni-trier.de/~ley/pers/hy/h/HoganAidan.html ▾
Mar 7, 2014 - Emir Muñoz, Aidan Hogan, Alessandra Mileo: Using linked data to ...
Patrick O'Byrne, Aidan Hogan: Exploring the Dynamics of Linked Data.

Aidan Hogan's Homepage
aidanhogan.com/ ▾
Aidan Hogan, Antoine Zimmermann, Jürgen Umbrich, Axel Polleres and Stefan Decker. "Scalable and Distributed Methods for Entity Matching, Consolidation ...
[journal](#) - [book-chapter](#) - [conference](#) - [workshop](#)



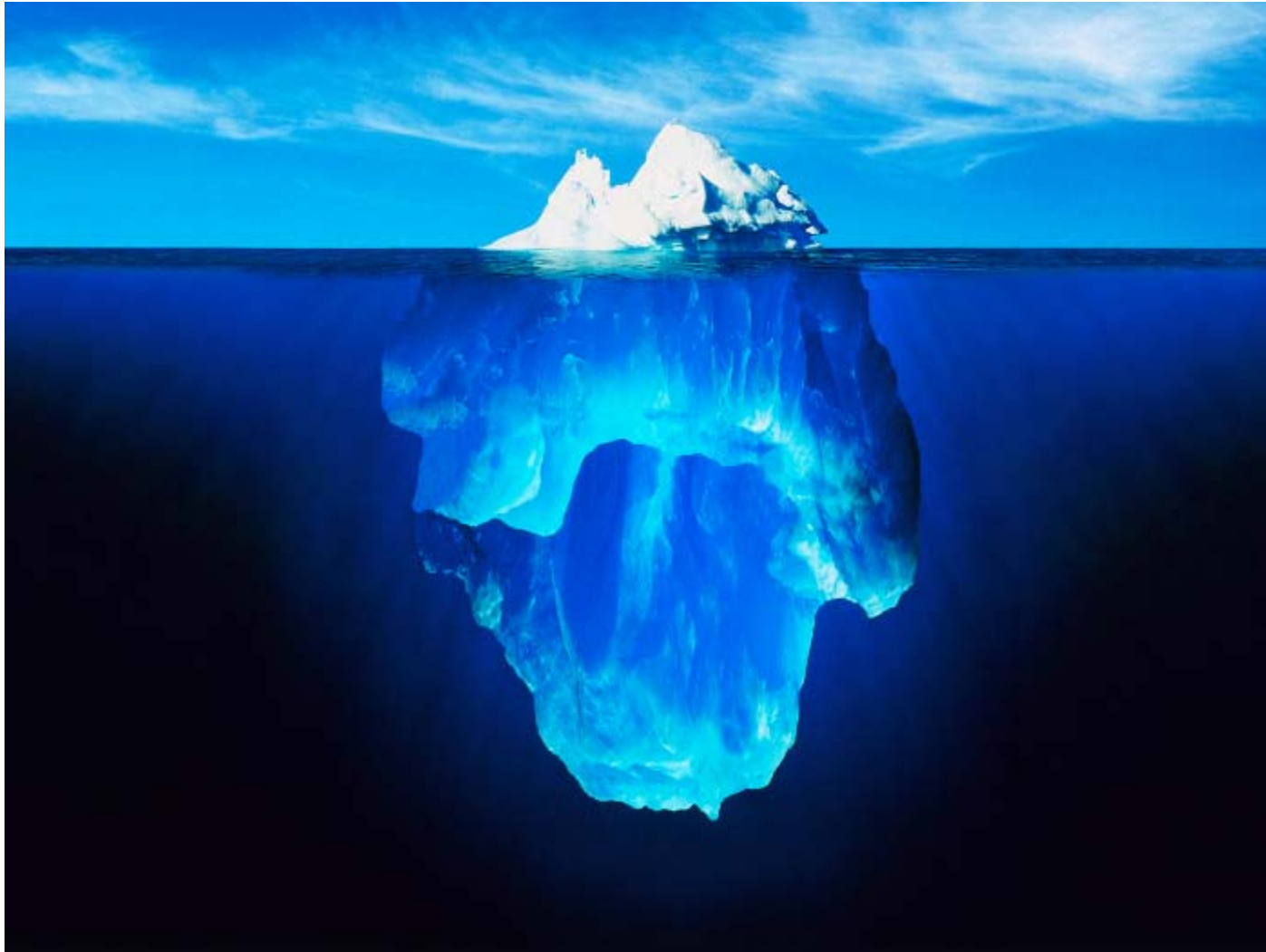
Ranking in Information Retrieval

- **Relevance**: Is the document relevant for the query?
 - Term Frequency * Inverse Document Frequency
 - Touched on Cosine similarity
- **Importance**: Is the document an important/prominent one?
 - Links analysis
 - PageRank

Ranking: Science or Art?



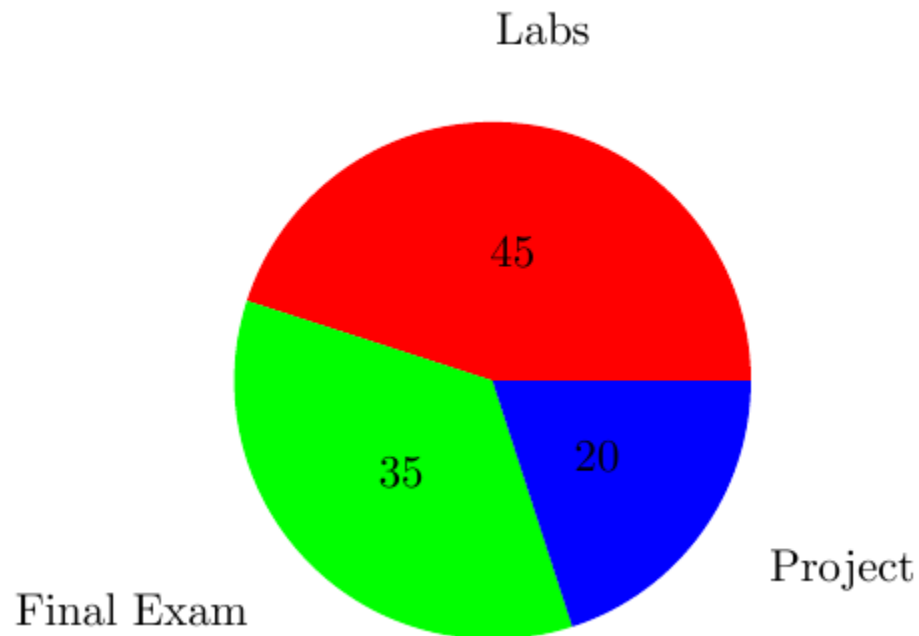
Information Retrieval & Relevance



CLASS PROJECTS

Course Marking

- 45% for Weekly Labs (~3% a lab!)
- 35% for Final Exam
- 20% for Small Class Project



Class Project

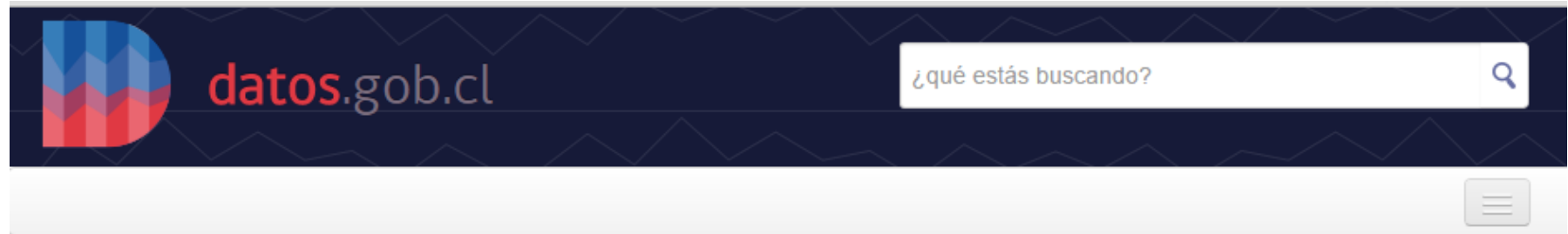



- Done in pairs (typically)
- Goal: Use what you've learned to do something cool (basically)
- Expected difficulty: A bit more than a lab's worth
 - But without guidance (can extend lab code)
- Marked on: Difficulty, appropriateness, scale, good use of techniques, presentation, coolness
 - Ambition is appreciated, even if you don't succeed: **feel free to bite off more than you can chew!**
- Process:
 - Pair up (default random) **by Wednesday, the end of the lab**
 - Start thinking up topics
 - If you need data or get stuck, I will (try to) help out
- Deliverables: 5 minute presentation & 3-page report

Datasets to play with

- Wikipedia information
- IMDb (including ratings, directors, etc.)
- ArnetMiner (CS research papers w/ citations)
- Wikidata (like Wikipedia for data!)
- Twitter
- World Bank
- **Find others, e.g., at <http://datahub.io/>**

Open Government Data Chile



 Inicio / Catálogo

 **CATÁLOGO DE DATOS** DATASETS PUBLICADOS: 1.195

 **CATEGORÍAS** [ver todas](#)

Instituciones publicadoras

Presidencia de la República (2)

➤ Fundación Integra (2)

Ministerio del Interior (314)

- Subsecretaría del Interior (5)
- Subsecretaría de Desarrollo Regional (23)
- Oficina Nacional de Emergencia (1)
- Intendencia Arica y Parinacota (3)

Ministerio de Relaciones Exteriores (41)

- Subsecretaría de Relaciones Exteriores (8)



- Gobierno (600)
- Salud (245)
- Comunidad (177)
- General (169)
- Geografía (164)
- Sociedad (140)
- Finanzas (132)
- Educación (117)
- Planificación (116)
- Negocios (109)
- Industria (89)
- Seguridad (80)
- Empleo (79)
- Turismo (76)
- Comunicaciones (73)

Questions

