

CC5212-1
PROCESAMIENTO MASIVO DE DATOS
OTOÑO 2015

Lecture 8: Information Retrieval II

Aidan Hogan
aidhog@gmail.com

How does Google crawl the Web?



Inverted Indexing

WIKIPEDIA The Free Encyclopedia

en.wikipedia.org/wiki/Fruitvale_Station

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](#).

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,[...]), ...
...	...

Inverted index:

**INFORMATION RETRIEVAL:
RANKING**

How Does Google Get Such Good Results?

Google

aidan hogan

Web Images News Videos More Search tools

Dr. Aidan Hogan | DERI
[www.deri.ie/~aidan-hogan](#)
 Tel: +353 91 407723. Aidan (do) Hogan is the Director, Homepage: Aidan.Hogan@deri.ie
 with 5296 CiteSpace, 46 Pruning, 46 Nodes, 46 Edges, 46 Weights, 46 Labels, 46 Modularity, 46 Q, 46 P, 46 S, 46 T, 46 U, 46 V, 46 W, 46 X, 46 Y, 46 Z, 46 AA, 46 AB, 46 AC, 46 AD, 46 AE, 46 AF, 46 AG, 46 AH, 46 AI, 46 AJ, 46 AK, 46 AL, 46 AM, 46 AN, 46 AO, 46 AP, 46 AQ, 46 AR, 46 AS, 46 AT, 46 AU, 46 AV, 46 AW, 46 AX, 46 AY, 46 AZ, 46 BA, 46 BB, 46 BC, 46 BD, 46 BE, 46 BF, 46 BG, 46 BH, 46 BI, 46 BJ, 46 BK, 46 BL, 46 BM, 46 BN, 46 BO, 46 BP, 46 BQ, 46 BR, 46 BS, 46 BT, 46 BU, 46 BV, 46 BW, 46 BX, 46 BY, 46 BZ, 46 CA, 46 CB, 46 CC, 46 CD, 46 CE, 46 CF, 46 CG, 46 CH, 46 CI, 46 CJ, 46 CK, 46 CL, 46 CM, 46 CN, 46 CO, 46 CP, 46 CQ, 46 CR, 46 CS, 46 CT, 46 CU, 46 CV, 46 CW, 46 CX, 46 CY, 46 CZ, 46 DA, 46 DB, 46 DC, 46 DD, 46 DE, 46 DF, 46 DG, 46 DH, 46 DI, 46 DJ, 46 DK, 46 DL, 46 DM, 46 DN, 46 DO, 46 DP, 46 DQ, 46 DR, 46 DS, 46 DT, 46 DU, 46 DV, 46 DW, 46 DX, 46 DY, 46 DZ, 46 EA, 46 EB, 46 EC, 46 ED, 46 EE, 46 EF, 46 EG, 46 EH, 46 EI, 46 EJ, 46 EK, 46 EL, 46 EM, 46 EN, 46 EO, 46 EP, 46 EQ, 46 ER, 46 ES, 46 ET, 46 EU, 46 EV, 46 EW, 46 EX, 46 EY, 46 EZ, 46 FA, 46 FB, 46 FC, 46 FD, 46 FE, 46 FF, 46 FG, 46 FH, 46 FI, 46 FJ, 46 FK, 46 FL, 46 FM, 46 FN, 46 FO, 46 FP, 46 FQ, 46 FR, 46 FS, 46 FT, 46 FU, 46 FV, 46 FW, 46 FX, 46 FY, 46 FZ, 46 GA, 46 GB, 46 GC, 46 GD, 46 GE, 46 GF, 46 GG, 46 GH, 46 GI, 46 GJ, 46 GK, 46 GL, 46 GM, 46 GN, 46 GO, 46 GP, 46 GQ, 46 GR, 46 GS, 46 GT, 46 GU, 46 GV, 46 GW, 46 GX, 46 GY, 46 GZ, 46 HA, 46 HB, 46 HC, 46 HD, 46 HE, 46 HF, 46 HG, 46 HH, 46 HI, 46 HJ, 46 HK, 46 HL, 46 HM, 46 HN, 46 HO, 46 HP, 46 HQ, 46 HR, 46 HS, 46 HT, 46 HU, 46 HV, 46 HW, 46 HX, 46 HY, 46 HZ, 46 IA, 46 IB, 46 IC, 46 ID, 46 IE, 46 IF, 46 IG, 46 IH, 46 II, 46 IJ, 46 IK, 46 IL, 46 IM, 46 IN, 46 IO, 46 IP, 46 IQ, 46 IR, 46 IS, 46 IT, 46 IU, 46 IV, 46 IW, 46 IX, 46 IY, 46 IZ, 46 JA, 46 JB, 46 JC, 46 JD, 46 JE, 46 JF, 46 JG, 46 JH, 46 JI, 46 JJ, 46 JK, 46 JL, 46 JM, 46 JN, 46 JO, 46 JP, 46 JQ, 46 JR, 46 JS, 46 JT, 46 JU, 46 JV, 46 JW, 46 JX, 46 JY, 46 JZ, 46 KA, 46 KB, 46 KC, 46 KD, 46 KE, 46 KF, 46 KG, 46 KH, 46 KI, 46 KJ, 46 KK, 46 KL, 46 KM, 46 KN, 46 KO, 46 KP, 46 KQ, 46 KR, 46 KS, 46 KT, 46 KU, 46 KV, 46 KW, 46 KX, 46 KY, 46 KZ, 46 LA, 46 LB, 46 LC, 46 LD, 46 LE, 46 LF, 46 LG, 46 LH, 46 LI, 46 LJ, 46 LK, 46 LL, 46 LM, 46 LN, 46 LO, 46 LP, 46 LQ, 46 LR, 46 LS, 46 LT, 46 LU, 46 LV, 46 LW, 46 LX, 46 LY, 46 LZ, 46 MA, 46 MB, 46 MC, 46 MD, 46 ME, 46 MF, 46 MG, 46 MH, 46 MI, 46 MJ, 46 MK, 46 ML, 46 MM, 46 MN, 46 MO, 46 MP, 46 MQ, 46 MR, 46 MS, 46 MT, 46 MU, 46 MV, 46 MW, 46 MX, 46 MY, 46 MZ, 46 NA, 46 NB, 46 NC, 46 ND, 46 NE, 46 NF, 46 NG, 46 NH, 46 NI, 46 NJ, 46 NK, 46 NL, 46 NM, 46 NO, 46 NP, 46 NQ, 46 NR, 46 NS, 46 NT, 46 NU, 46 NV, 46 NW, 46 NX, 46 NY, 46 NZ, 46 OA, 46 OB, 46 OC, 46 OD, 46 OE, 46 OF, 46 OG, 46 OH, 46 OI, 46 OJ, 46 OK, 46 OL, 46 OM, 46 ON, 46 OO, 46 OP, 46 OQ, 46 OR, 46 OS, 46 OT, 46 OU, 46 OV, 46 OW, 46 OX, 46 OY, 46 OZ, 46 PA, 46 PB, 46 PC, 46 PD, 46 PE, 46 PF, 46 PG, 46 PH, 46 PI, 46 PJ, 46 PK, 46 PL, 46 PM, 46 PN, 46 PO, 46 PP, 46 PQ, 46 PR, 46 PS, 46 PT, 46 PU, 46 PV, 46 PW, 46 PX, 46 PY, 46 PZ, 46 QA, 46 QB, 46 QC, 46 QD, 46 QE, 46 QF, 46 QG, 46 QH, 46 QI, 46 QJ, 46 QK, 46 QL, 46 QM, 46 QN, 46 QO, 46 QP, 46 QQ, 46 QR, 46 QS, 46 QT, 46 QU, 46 QV, 46 QW, 46 QX, 46 QY, 46 QZ, 46 RA, 46 RB, 46 RC, 46 RD, 46 RE, 46 RF, 46 RG, 46 RH, 46 RI, 46 RJ, 46 RK, 46 RL, 46 RM, 46 RN, 46 RO, 46 RP, 46 RQ, 46 RR, 46 RS, 46 RT, 46 RU, 46 RV, 46 RW, 46 RX, 46 RY, 46 RZ, 46 SA, 46 SB, 46 SC, 46 SD, 46 SE, 46 SF, 46 SG, 46 SH, 46 SI, 46 SJ, 46 SK, 46 SL, 46 SM, 46 SN, 46 SO, 46 SP, 46 SQ, 46 SR, 46 SS, 46 ST, 46 SU, 46 SV, 46 SW, 46 SX, 46 SY, 46 SZ, 46 TA, 46 TB, 46 TC, 46 TD, 46 TE, 46 TF, 46 TG, 46 TH, 46 TI, 46 TJ, 46 TK, 46 TL, 46 TM, 46 TN, 46 TO, 46 TP, 46 TQ, 46 TR, 46 TS, 46 TT, 46 TU, 46 TV, 46 TW, 46 TX, 46 TY, 46 TZ, 46 UA, 46 UB, 46 UC, 46 UD, 46 UE, 46 UF, 46 UG, 46 UH, 46 UI, 46 UJ, 46 UK, 46 UL, 46 UM, 46 UN, 46 UO, 46 UP, 46 UQ, 46 UR, 46 US, 46 UT, 46 UU, 46 UV, 46 UW, 46 UX, 46 UY, 46 UZ, 46 VA, 46 VB, 46 VC, 46 VD, 46 VE, 46 VF, 46 VG, 46 VH, 46 VI, 46 VJ, 46 VK, 46 VL, 46 VM, 46 VN, 46 VO, 46 VP, 46 VQ, 46 VR, 46 VS, 46 VT, 46 VU, 46 VV, 46 VW, 46 VX, 46 VY, 46 VZ, 46 WA, 46 WB, 46 WC, 46 WD, 46 WE, 46 WF, 46 WG, 46 WH, 46 WI, 46 WJ, 46 WK, 46 WL, 46 WM, 46 WN, 46 WO, 46 WP, 46 WQ, 46 WR, 46 WS, 46 WT, 46 WU, 46 WV, 46 WW, 46 WX, 46 WY, 46 WZ, 46 XA, 46 XB, 46 XC, 46 XD, 46 XE, 46 XF, 46 XG, 46 XH, 46 XI, 46 XJ, 46 XK, 46 XL, 46 XM, 46 XN, 46 XO, 46 XP, 46 XQ, 46 XR, 46 XS, 46 XT, 46 XU, 46 XV, 46 XW, 46 XX, 46 XY, 46 XZ, 46 YA, 46 YB, 46 YC, 46 YD, 46 YE, 46 YF, 46 YG, 46 YH, 46 YI, 46 YJ, 46 YK, 46 YL, 46 YM, 46 YN, 46 YO, 46 YP, 46 YQ, 46 YR, 46 YS, 46 YT, 46 YU, 46 YV, 46 YW, 46 YX, 46 YY, 46 YZ, 46 ZA, 46 ZB, 46 ZC, 46 ZD, 46 ZE, 46 ZF, 46 ZG, 46 ZH, 46 ZI, 46 ZJ, 46 ZK, 46 ZL, 46 ZM, 46 ZN, 46 ZO, 46 ZP, 46 ZQ, 46 ZR, 46 ZS, 46 ZT, 46 ZU, 46 ZV, 46 ZW, 46 ZX, 46 ZY, 46 ZZ

Aidan Hogan's Homepage
[aidanhogan.com](#)
 Aidan Hogan is the Director, Homepage: Aidan.Hogan@deri.ie
 Director, Scalable and Distributed Methods for Entity Matching, Classification, Journal: 2008-2010, conference - workshop

thumbs up

Two Sides to Ranking: Relevance

Google

obama

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?name=foods&id=9](#)
 Broccoli can provide you with some special cholesterol-lowering benefits if you will cook ...
 by steaming. The fiber-related components in broccoli do a better job ...

News for broccoli

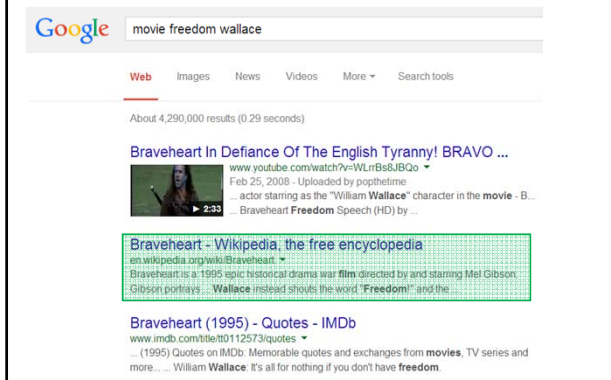
Mistakes We All Make With Spaghetti, Steak And ...
 Huffington Post - 2 days ago
 But in her new book *Brassicas: Conquering the World's Healthiest Vegetables*, she says ...
 plunking **broccoli**, cauliflower or Brussels sprouts into ...

thumbs down

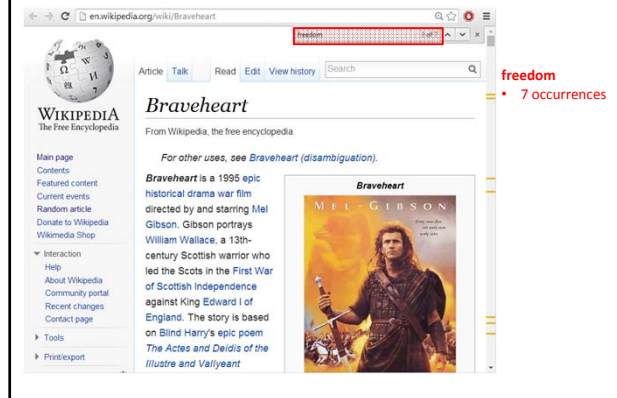
Two Sides to Ranking: Importance

RANKING:
RELEVANCE

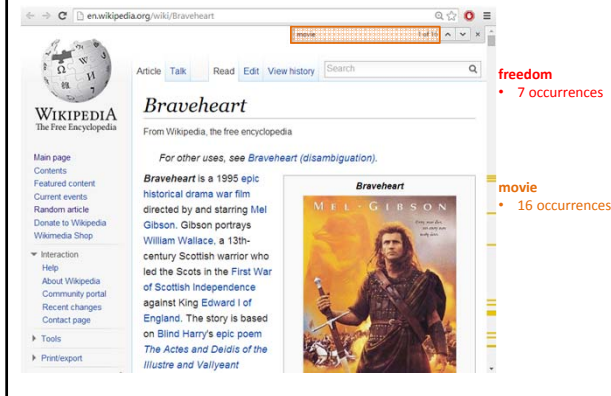
Example Query



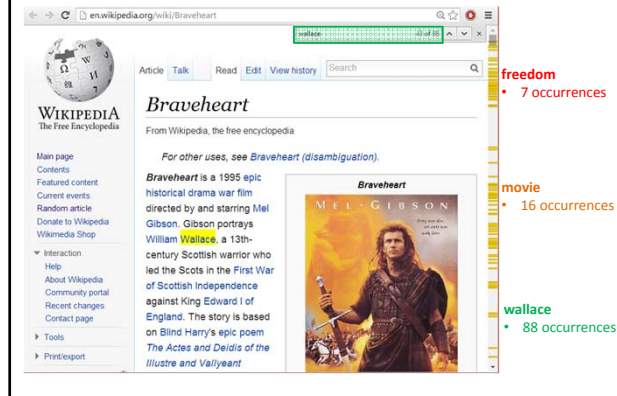
Matches in a Document



Matches in a Document



Matches in a Document



Usefulness of Words

Google search results for the query "movie freedom wallace". The results show the frequency of each word in the query:

- 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
 - $tf(t, d)$... 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
 $tf(t, d) = \frac{\text{count}(t, d)}{\sum_{t' \in d} \text{count}(t', d)}$
 $tf(t, d) = \frac{\text{count}(t, d)}{\max\{\text{count}(t', d) | t' \in d\}}$
 - A combination / something else ☺

Relevance Measure: TF-IDF

- **IDF: Inverse Document Frequency**
 - Measures how rare/common a term is across **all** documents
 - $idf(t, D)$...
 - Logarithmically scaled document occurrences
 $idf(t, D) = \log\left(\frac{|D|}{|\{d' \in D : t \in d'\}| + 1}\right)$

Relevance Measure: TF-IDF

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

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

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

Relevance Measure: TF-IDF

Diagram illustrating the calculation of TF-IDF for the query "movie freedom wallace".

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 Calculation

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

t	$tf(t, d)$
movie	16
freedom	7
wallace	43

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

Relevance Measure: **TF-IDF****Term Frequency**

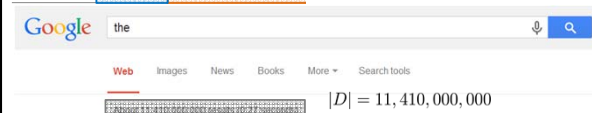
$$tf(t, d) = \text{count}(t, d)$$

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	

Relevance Measure: **TF-IDF****Term Frequency**

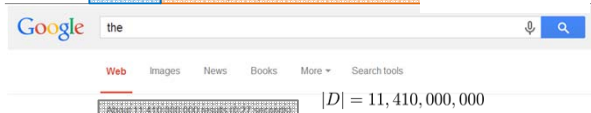
$$tf(t, d) = \text{count}(t, d)$$

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	13.66
freedom	7	198,000,000	57.62
wallace	43	49,200,000	231.91

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

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

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

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

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

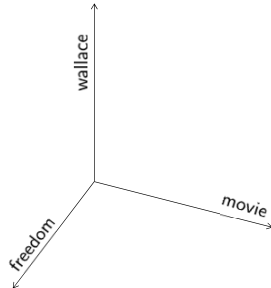
$$\text{score}(q, d) = \sum_{t \in q} tf-idf(t, d)$$

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

Vector Space Model (a mention)

t	$tf(t, d)$
movie	16
freedom	7
wallace	43

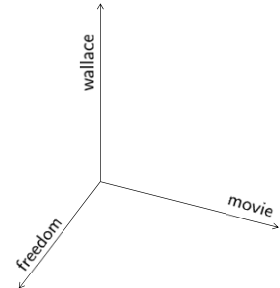
$$l = \sqrt{\sum_{t \in q} tf(t, d)^2}$$



Vector Space Model (a mention)

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

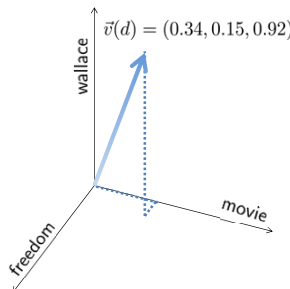
$$l = \sqrt{\sum_{t \in q} tf(t, d)^2}$$



Vector Space Model (a mention)

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

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



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

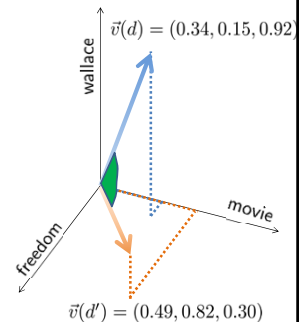
$$\text{sim}(d, d') \approx 0.57$$

• 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

Field-Based Boosting

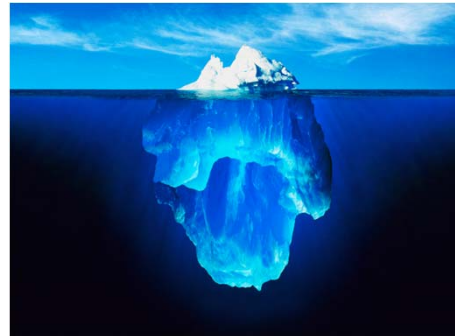
- Not all text is equal: titles, headers, etc.

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/wiki/Mount_Obama" style="color: red; text-decoration: underline;">http://en.wikipedia.org/wiki/Mount_Obama
</p>
</body>
</html>
```

Information Retrieval & Relevance



Apache to the rescue again!

Lucene: An Inverted Index Engine

- Open Source Java Project
- Will play with it in the labs



RANKING: IMPORTANCE

Two Sides to Ranking: Importance



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



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



Link Importance

Which is more "important": a link from Barack Obama's Wikipedia page or a link from buyv1agra.com?



PageRank



PageRank

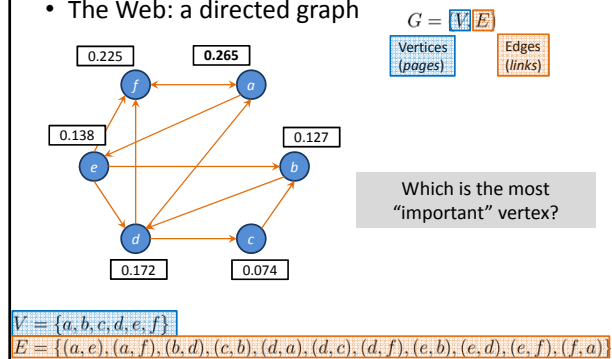
- Not just a count of inlinks
 - A link from a more important page is more important
 - A link from a page with fewer links is more important
- ∴ A page with lots of inlinks from important pages (which have few outlinks) is more important

PageRank is Recursive



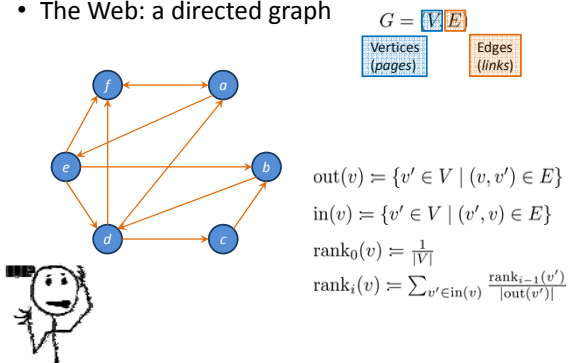
PageRank Model

- The Web: a directed graph

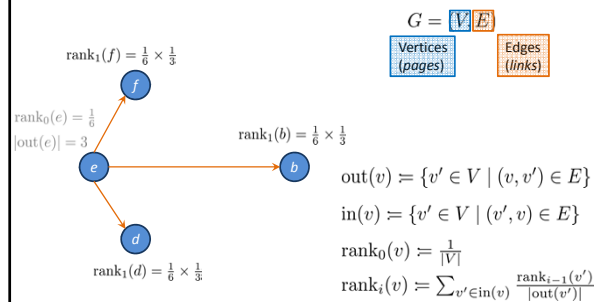


PageRank Model

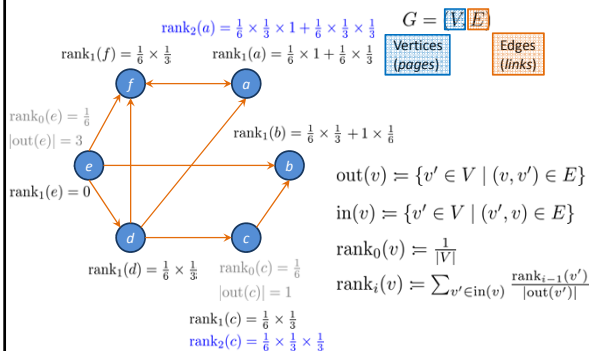
- The Web: a directed graph



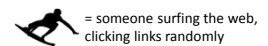
PageRank Model



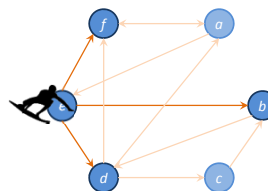
PageRank Model




PageRank: Random Surfer Model

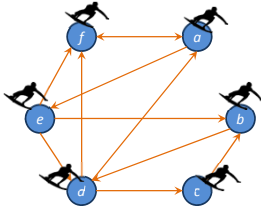


- What is the probability of being at page x after n hops?




PageRank: Random Surfer Model

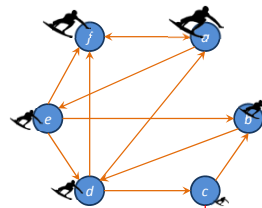
 = someone surfing the web, clicking links randomly



- 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

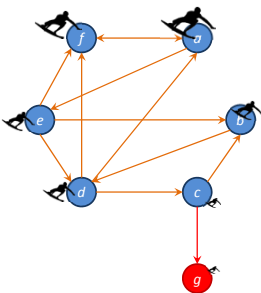


- What is the probability of being at page x after n hops?
- *Initial state*: surfer equally likely to start at any node
- PageRank applied iteratively for each hop: score indicates probability of being at that page after than many hops

What would happen with g over time?


PageRank: Random Surfer Model

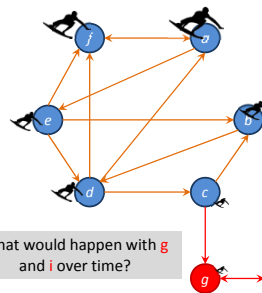
 = someone surfing the web, clicking links randomly



- 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

PageRank: Random Surfer Model


 = someone surfing the web, clicking links randomly

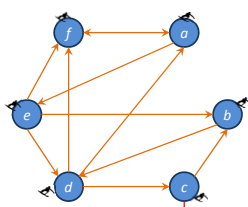


- What is the probability of being at page x after n hops?
- *Initial state*: surfer equally likely to start at any node
- PageRank applied iteratively for each hop: score indicates probability of being at that page after than many hops
- If the surfer reaches a page without links, the surfer randomly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

What would happen with g and i over time?

PageRank: Random Surfer Model

 = someone surfing the web, clicking links randomly




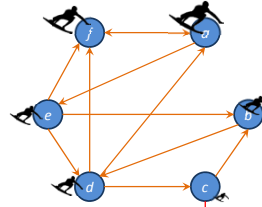
- What is the probability of being at page x after n hops?
- *Initial state*: surfer equally likely to start at any node
- PageRank applied iteratively for each hop: score indicates probability of being at that page after than many hops
- If the surfer reaches a page without links, the surfer randomly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

What would happen with g and i over time?



PageRank: Random Surfer Model

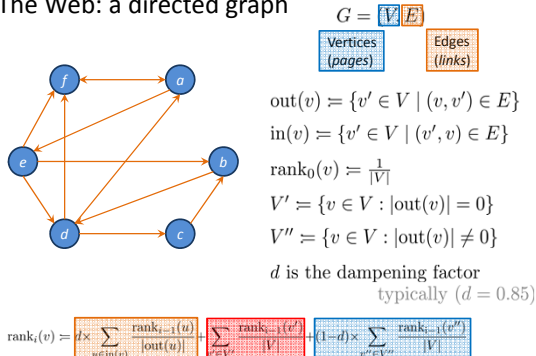
 = someone surfing the web, clicking links randomly



- 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



PageRank: Benefits

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

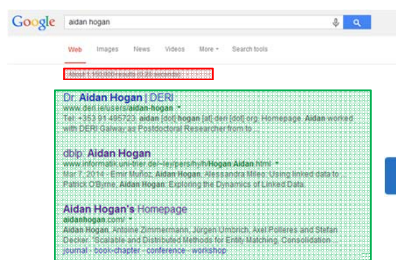


Two Sides to Ranking: Importance



INFORMATION RETRIEVAL: RECAP

How Does Google Get Such Good Results?



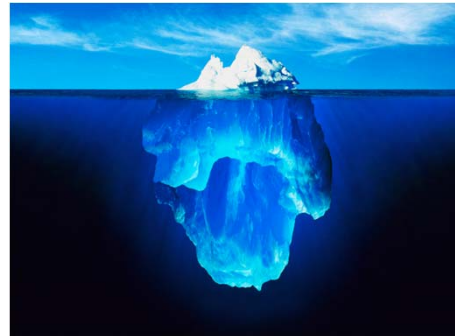
Ranking in Information Retrieval

- 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?



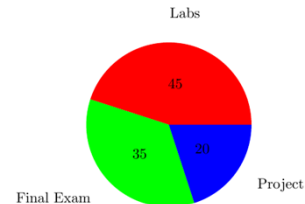
Information Retrieval & Relevance



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

datos.gob.cl

¿qué estás buscando?

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 (3)
- > 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)

CATEGORÍAS

Gobierno (600)	Salud (245)
Comunicación (1277)	General (169)
Geografía (154)	Sociedad (148)
Finanzas (132)	Educación (117)
Planificación (116)	
Negocios (109)	Industria (89)
Seguridad (80)	Empleo (79)
Turismo (76)	
Comunicaciones (73)	

Next Week (May 4th, 6th)

- No official classes or labs next week



- but ...
- **Good opportunity to meet with your lab partner to explore project ideas!**
- **Deadline for finding a topic: May 13th**

Questions

