

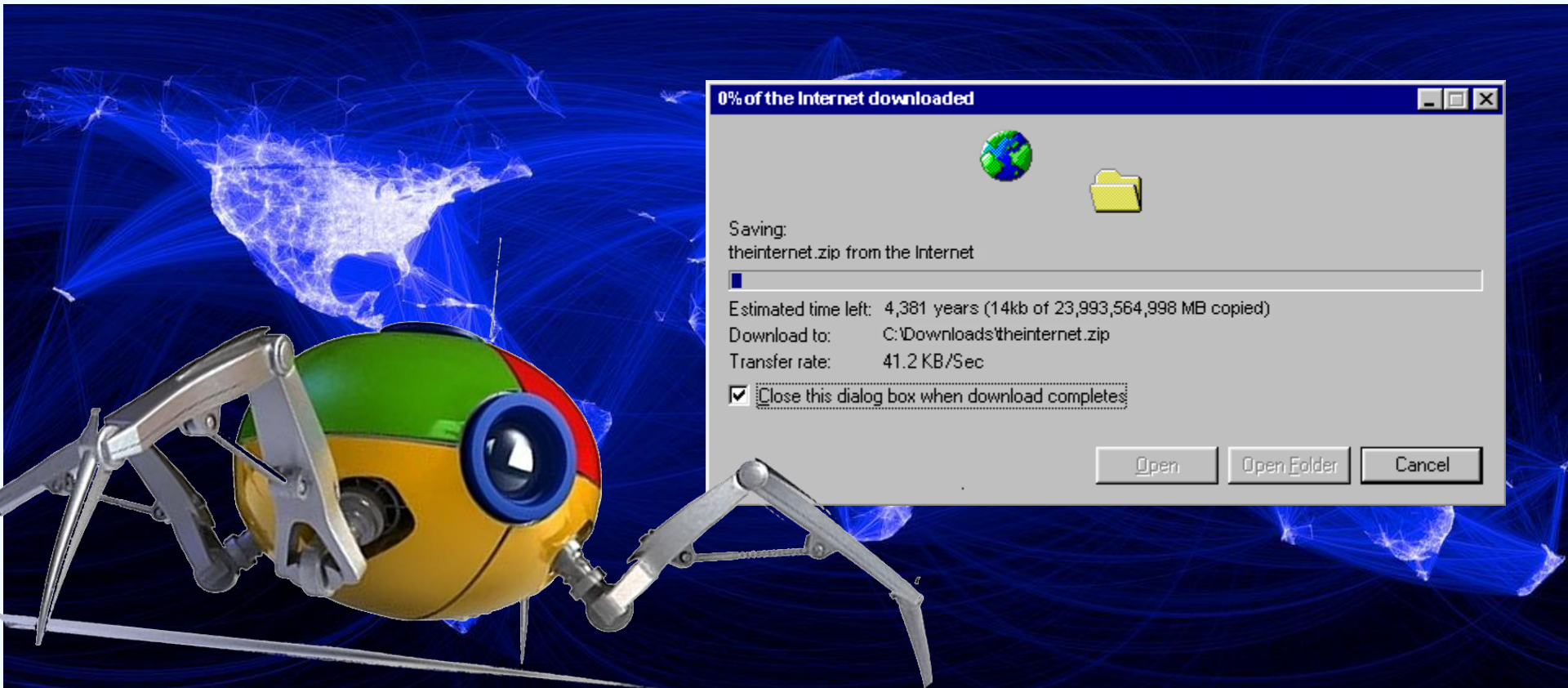
CC5212-1

PROCESAMIENTO MASIVO DE DATOS
OTOÑO 2017

Lecture 7: Information Retrieval II

Aidan Hogan
aidhog@gmail.com

How does Google know about the Web?



Inverted Index: Example



1

Fruitvale Station

From Wikipedia, the free encyclopedia

Fruitvale Station is a 2013 American [drama film](#) written and directed by [Ryan Coogler](#).

Inverted index:

Term List	Posting List
a	(1,2,...)
american	(1,5,...)
and	(1,2,...)
by	(1,2,...)
directed	(1,2,...)
drama	(1,16,...)
...	...

Apache Lucene



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

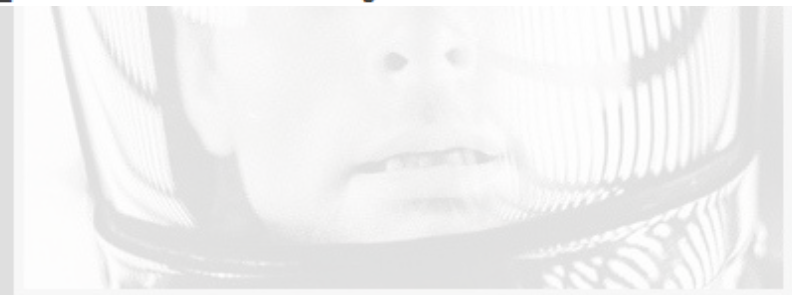


My God. It's full of win.

Apache Lucene



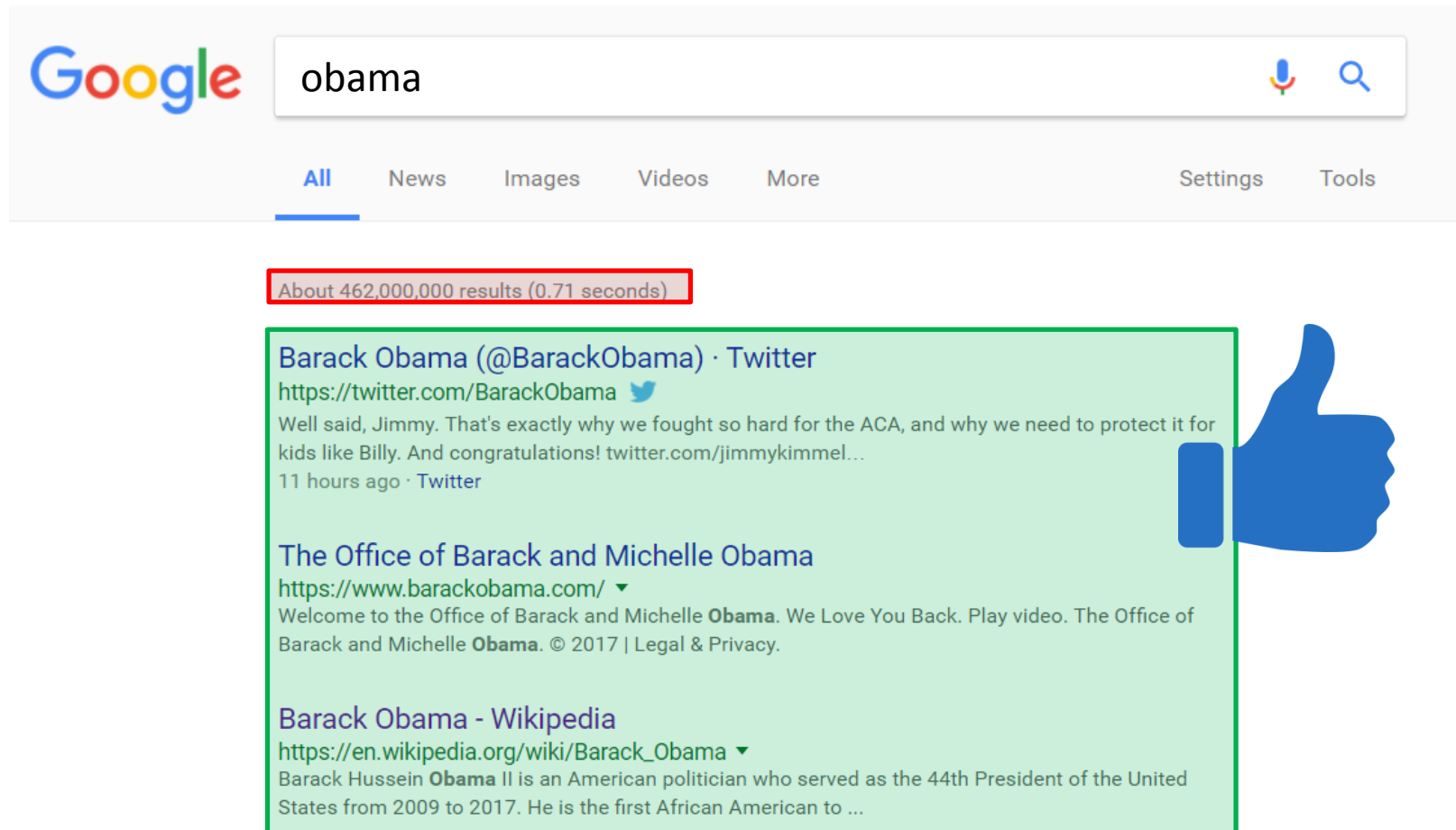
```
Tasks Console
SearchWikiIndex [Java Application] C:\Program Files\Java\jre1.8.0_65\bin\javaw.exe (03-05-2017 12:45:22 a. m.)
Opening directory at lucene
Enter a keyword search phrase:
obama
Running query: obama
Parsed query: TITLE:obam^5.0 ABSTRACT:obam
Matching documents: 255
Showing top 10 results
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/Transición_presidencial_de_Barack_Obama Transición presidencial de Barack Obama
9 http://es.wikipedia.org/wiki/Por_qué_Obama_ganará_en_2008_y_en_2012 Por qué Obama ganará en 2008 y en 2012
10 http://es.wikipedia.org/wiki/Ricardo_Mangue_Obama_Nfubea Ricardo Mangue Obama Nfubea
```



My God. It's full of win.

INFORMATION RETRIEVAL: RANKING

How Does Google Get Such Good Results?




Google

obama

All News Images Videos More Settings Tools

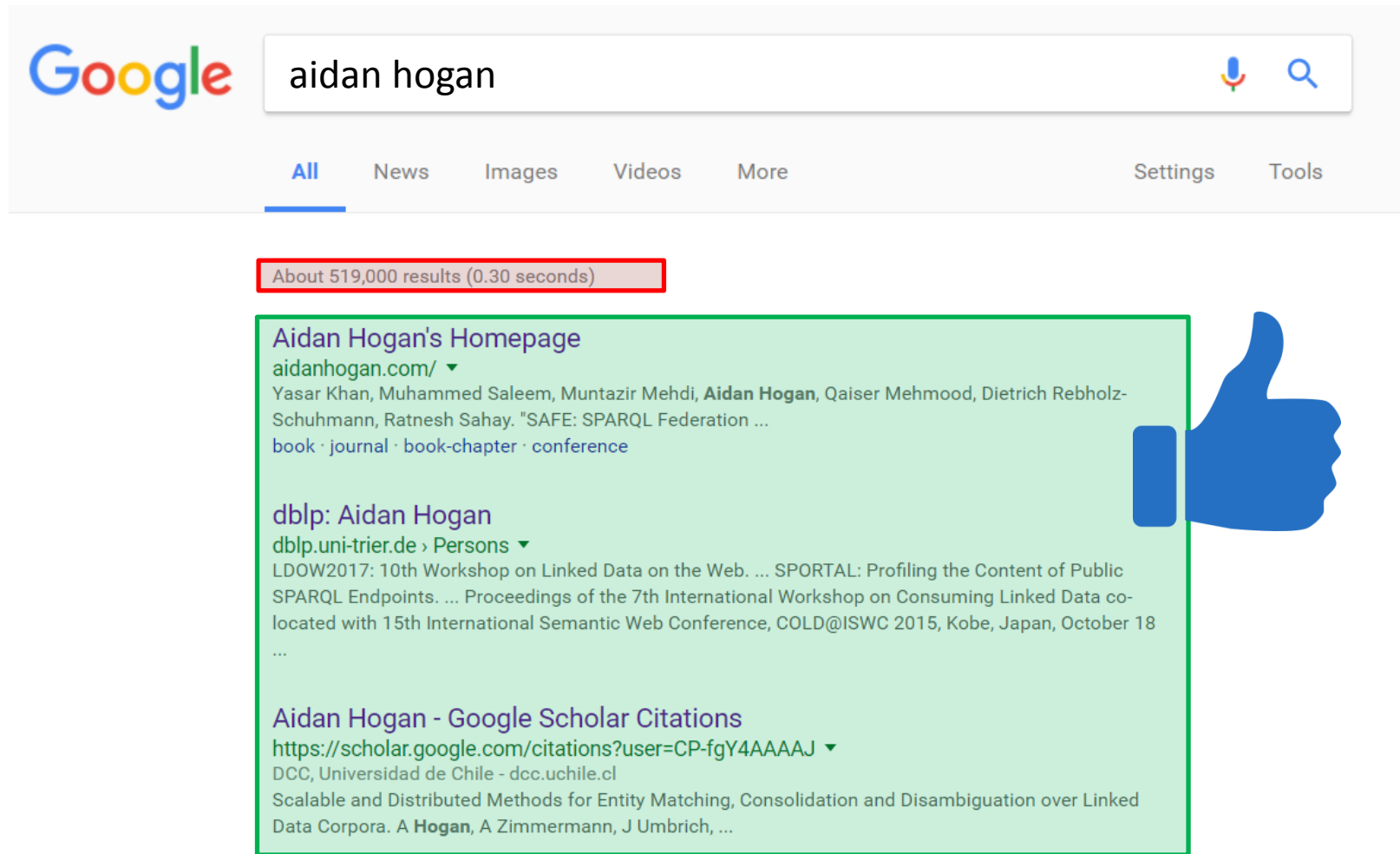
About 462,000,000 results (0.71 seconds)

Barack Obama (@BarackObama) · Twitter
<https://twitter.com/BarackObama> 
Well said, Jimmy. That's exactly why we fought so hard for the ACA, and why we need to protect it for kids like Billy. And congratulations! [twitter.com/jimmykimmel...](https://twitter.com/jimmykimmel)
11 hours ago · [Twitter](#)

The Office of Barack and Michelle Obama
<https://www.barackobama.com/> ▼
Welcome to the Office of Barack and Michelle **Obama**. We Love You Back. Play video. The Office of Barack and Michelle **Obama**. © 2017 | [Legal & Privacy](#).

Barack Obama - Wikipedia
https://en.wikipedia.org/wiki/Barack_Obama ▼
Barack Hussein **Obama** II is an American politician who served as the 44th President of the United States from 2009 to 2017. He is the first African American to ...

How Does Google Get Such Good Results?



The image shows a Google search interface. The search bar contains the text "aidan hogan". Below the search bar, the "All" tab is selected. The search results show "About 519,000 results (0.30 seconds)". The first result is "Aidan Hogan's Homepage" with the URL "aidanhogan.com/". The second result is "dblp: Aidan Hogan" with the URL "dblp.uni-trier.de > Persons". The third result is "Aidan Hogan - Google Scholar Citations" with the URL "https://scholar.google.com/citations?user=CP-fgY4AAAAJ". A large blue thumbs up icon is positioned to the right of the search results.

Google

aidan hogan

All News Images Videos More Settings Tools

About 519,000 results (0.30 seconds)

Aidan Hogan's Homepage
aidanhogan.com/ ▼
Yasar Khan, Muhammed Saleem, Muntazir Mehdi, **Aidan Hogan**, Qaiser Mehmood, Dietrich Rebholz-Schuhmann, Ratnesh Sahay. "SAFE: SPARQL Federation ...
book · journal · book-chapter · conference

dblp: Aidan Hogan
dblp.uni-trier.de > Persons ▼
LDOW2017: 10th Workshop on Linked Data on the Web. ... SPOTAL: Profiling the Content of Public SPARQL Endpoints. ... Proceedings of the 7th International Workshop on Consuming Linked Data co-located with 15th International Semantic Web Conference, COLD@ISWC 2015, Kobe, Japan, October 18 ...

Aidan Hogan - Google Scholar Citations
https://scholar.google.com/citations?user=CP-fgY4AAAAJ ▼
DCC, Universidad de Chile - dcc.uchile.cl
Scalable and Distributed Methods for Entity Matching, Consolidation and Disambiguation over Linked Data Corpora. A **Hogan**, A Zimmermann, J Umbrich, ...

How does Google Get Such Good Results?

Google

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

Web Videos News Images Shopping More Search tools

About 64,700,000 results (0.91 seconds)

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

Google

da da da dum symphony

Web Videos News Shopping Images More Search tools

About 107,000 results (0.36 seconds)



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




sometimes when i'm

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

Google Search I'm Feeling Lucky

Two 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?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 sautés plunking **broccoli**, cauliflower or brussels sprouts into ...

Barack Obama

Broccoli

=

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 Park Area, the committee organized a site visit to the C

RANKING:
RELEVANCE

Example Query

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



movie freedom wallace

Web

Images

News

Videos

More ▼

Search tools

About 4,290,000 results (0.29 seconds)

Braveheart In Defiance Of The English Tyranny! BRAVO ...



www.youtube.com/watch?v=WLrrBs8JBQo ▼

Feb 25, 2008 - Uploaded by popthetime

... actor starring as the "William **Wallace**" character in the **movie** - B...

... Braveheart **Freedom** Speech (HD) by ...

Braveheart - Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/Braveheart ▼

Braveheart is a 1995 epic historical drama war **film** directed by and starring Mel Gibson.

Gibson portrays ... **Wallace** instead shouts the word "**Freedom!**" and the ...

Braveheart (1995) - Quotes - IMDb

www.imdb.com/title/tt0112573/quotes ▼

... (1995) Quotes on IMDb: Memorable quotes and exchanges from **movies**, TV series and more... ... William **Wallace**: It's all for nothing if you don't have **freedom**.

Matches in a Document

W Braveheart - Wikipedia

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

Not logged in | **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]

freedom
• 7 occurrences

1 Plot
2 Cast
3 Production

Upload file

Braveheart
MEL GIBSON
Every man dies,
not every man
really lives.
BRAVEHEART

Matches in a Document

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

W Braveheart - Wikipedia

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

Not logged in

wallace 44 de 88

Article Talk Read Edit View history Search Wikipedia

Braveheart

From Wikipedia, the free encyclopedia

For other uses, see *Braveheart* (disambiguation).

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

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

Contents [hide]

1 Plot

2 Cast

3 Production

4 Reception

5 Trivia

6 External links

7 References

8 Notes

9 See also

10 Further reading

11 Citations

12 External links

13 References

14 Notes

15 See also

16 Further reading

17 Citations

18 External links

19 References

20 Notes

21 See also

22 Further reading

23 Citations

24 External links

25 References

26 Notes

27 See also

28 Further reading

29 Citations

30 External links

31 References

32 Notes

33 See also

34 Further reading

35 Citations

36 External links

37 References

38 Notes

39 See also

40 Further reading

41 Citations

42 External links

43 References

44 Notes

45 See also

46 Further reading

47 Citations

48 External links

49 References

50 Notes

51 See also

52 Further reading

53 Citations

54 External links

55 References

56 Notes

57 See also

58 Further reading

59 Citations

60 External links

61 References

62 Notes

63 See also

64 Further reading

65 Citations

66 External links

67 References

68 Notes

69 See also

70 Further reading

71 Citations

72 External links

73 References

74 Notes

75 See also

76 Further reading

77 Citations

78 External links

79 References

80 Notes

81 See also

82 Further reading

83 Citations

84 External links

85 References

86 Notes

87 See also

88 Further reading

89 Citations

90 External links

91 References

92 Notes

93 See also

94 Further reading

95 Citations

96 External links

97 References

98 Notes

99 See also

100 Further reading

101 Citations

102 External links

103 References

104 Notes

105 See also

106 Further reading

107 Citations

108 External links

109 References

110 Notes

111 See also

112 Further reading

113 Citations

114 External links

115 References

116 Notes

117 See also

118 Further reading

119 Citations

120 External links

121 References

122 Notes

123 See also

124 Further reading

125 Citations

126 External links

127 References

128 Notes

129 See also

130 Further reading

131 Citations

132 External links

133 References

134 Notes

135 See also

136 Further reading

137 Citations

138 External links

139 References

140 Notes

141 See also

142 Further reading

143 Citations

144 External links

145 References

146 Notes

147 See also

148 Further reading

149 Citations

150 External links

151 References

152 Notes

153 See also

154 Further reading

155 Citations

156 External links

157 References

158 Notes

159 See also

160 Further reading

161 Citations

162 External links

163 References

164 Notes

165 See also

166 Further reading

167 Citations

168 External links

169 References

170 Notes

171 See also

172 Further reading

173 Citations

174 External links

175 References

176 Notes

177 See also

178 Further reading

179 Citations

180 External links

181 References

182 Notes

183 See also

184 Further reading

185 Citations

186 External links

187 References

188 Notes

189 See also

190 Further reading

191 Citations

192 External links

193 References

194 Notes

195 See also

196 Further reading

197 Citations

198 External links

199 References

200 Notes

201 See also

202 Further reading

203 Citations

204 External links

205 References

206 Notes

207 See also

208 Further reading

209 Citations

210 External links

211 References

212 Notes

213 See also

214 Further reading

215 Citations

216 External links

217 References

218 Notes

219 See also

220 Further reading

221 Citations

222 External links

223 References

224 Notes

225 See also

226 Further reading

227 Citations

228 External links

229 References

230 Notes

231 See also

232 Further reading

233 Citations

234 External links

235 References

236 Notes

237 See also

238 Further reading

239 Citations

240 External links

241 References

242 Notes

243 See also

244 Further reading

245 Citations

246 External links

247 References

248 Notes

249 See also

250 Further reading

251 Citations

252 External links

253 References

254 Notes

255 See also

256 Further reading

257 Citations

258 External links

259 References

260 Notes

261 See also

262 Further reading

263 Citations

264 External links

265 References

266 Notes

267 See also

268 Further reading

269 Citations

270 External links

271 References

272 Notes

273 See also

274 Further reading

275 Citations

276 External links

277 References

278 Notes

279 See also

280 Further reading

281 Citations

282 External links

283 References

284 Notes

285 See also

286 Further reading

287 Citations

288 External links

289 References

290 Notes

291 See also

292 Further reading

293 Citations

294 External links

295 References

296 Notes

297 See also

298 Further reading

299 Citations

300 External links

301 References

302 Notes

303 See also

304 Further reading

305 Citations

306 External links

307 References

308 Notes

309 See also

310 Further reading

311 Citations

312 External links

313 References

314 Notes

315 See also

316 Further reading

317 Citations

318 External links

319 References

320 Notes

321 See also

322 Further reading

323 Citations

324 External links

325 References

326 Notes

327 See also

328 Further reading

329 Citations

330 External links

331 References

332 Notes

333 See also

334 Further reading

335 Citations

336 External links

337 References

338 Notes

339 See also

340 Further reading

341 Citations

342 External links

343 References

344 Notes

345 See also

346 Further reading

347 Citations

348 External links

349 References

350 Notes

351 See also

352 Further reading

353 Citations

354 External links

355 References

356 Notes

357 See also

358 Further reading

359 Citations

360 External links

361 References

362 Notes

363 See also

364 Further reading

365 Citations

366 External links

367 References

368 Notes

369 See also

370 Further reading

371 Citations

372 External links

373 References

374 Notes

375 See also

376 Further reading

377 Citations

378 External links

379 References

380 Notes

381 See also

382 Further reading

383 Citations

384 External links

385 References

386 Notes

387 See also

388 Further reading

389 Citations

390 External links

391 References

392 Notes

393 See also

394 Further reading

395 Citations

396 External links

397 References

398 Notes

399 See also

400 Further reading

401 Citations

402 External links

403 References

404 Notes

405 See also

406 Further reading

407 Citations

408 External links

409 References

410 Notes

411 See also

412 Further reading

413 Citations

414 External links

415 References

416 Notes

417 See also

418 Further reading

419 Citations

420 External links

421 References

422 Notes

423 See also

424 Further reading

425 Citations

426 External links

427 References

428 Notes

429 See also

430 Further reading

431 Citations

432 External links

433 References

434 Notes

435 See also

436 Further reading

437 Citations

438 External links

439 References

440 Notes

441 See also

442 Further reading

443 Citations

444 External links

445 References

446 Notes

447 See also

448 Further reading

449 Citations

450 External links

451 References

452 Notes

453 See also

454 Further reading

455 Citations

456 External links

457 References

458 Notes

459 See also

460 Further reading

461 Citations

462 External links

463 References

464 Notes

465 See also

466 Further reading

467 Citations

468 External links

469 References

470 Notes

471 See also

472 Further reading

473 Citations

474 External links

475 References

476 Notes

477 See also

478 Further reading

479 Citations

480 External links

481 References

482 Notes

483 See also

484 Further reading

485 Citations

486 External links

487 References

488 Notes

489 See also

490 Further reading

491 Citations

492 External links

493 References

494 Notes

495 See also

496 Further reading

497 Citations

498 External links

499 References

500 Notes

501 See also

502 Further reading

503 Citations

504 External links

505 References

506 Notes

507 See also

508 Further reading

509 Citations

510 External links

511 References

512 Notes

513 See also

514 Further reading

515 Citations

516 External links

517 References

518 Notes

519 See also

520 Further reading

521 Citations

522 External links

523 References

524 Notes

525 See also

526 Further reading

527 Citations

528 External links

529 References

530 Notes

531 See also

532 Further reading

533 Citations

534 External links

535 References

536 Notes

537 See also

538 Further reading

539 Citations

540 External links

541 References

542 Notes

543 See also

544 Further reading

545 Citations

546 External links

547 References

548 Notes

549 See also

550 Further reading

551 Citations

552 External links

553 References

554 Notes

555 See also

556 Further reading

557 Citations

558 External links

559 References

560 Notes

561 See also

562 Further reading

563 Citations

564 External links

565 References

566 Notes

567 See also

568 Further reading

569 Citations

570 External links

571 References

572 Notes

573 See also

574 Further reading

575 Citations

576 External links

577 References

578 Notes

579 See also

580 Further reading

581 Citations

582 External links

583 References

584 Notes

585 See also

586 Further reading

587 Citations

588 External links

589 References

590 Notes

591 See also

592 Further reading

593 Citations

594 External links

595 References

596 Notes

597 See also

598 Further reading

599 Citations

600 External links

601 References

602 Notes

603 See also

604 Further reading

605 Citations

606 External links

607 References

608 Notes

609 See also

610 Further reading

611 Citations

612 External links

613 References

614 Notes

615 See also

616 Further reading

617 Citations

618 External links

619 References

620 Notes

621 See also

622 Further reading

623 Citations

624 External links

625 References

626 Notes

627 See also

628 Further reading

629 Citations

630 External links

631 References

632 Notes

633 See also

634 Further reading

635 Citations

636 External links

637 References

638 Notes

639 See also

640 Further reading

641 Citations

642 External links

643 References

644 Notes

645 See also

646 Further reading

647 Citations

648 External links

649 References

650 Notes

651 See also

652 Further reading

653 Citations

654 External links

655 References

656 Notes

657 See also

658 Further reading

659 Citations

660 External links

661 References

662 Notes

663 See also

664 Further reading

665 Citations

666 External links

667 References

668 Notes

669 See also

670 Further reading

671 Citations

672 External links

673 References

674 Notes

675 See also

676 Further reading

677 Citations

678 External links

679 References

680 Notes

681 See also

682 Further reading

683 Citations

684 External links

685 References

686 Notes

687 See also

688 Further reading

689 Citations

690 External links

691 References

692 Notes

693 See also

694 Further reading

695 Citations

696 External links

697 References

698 Notes

699 See also

700 Further reading

701 Citations

702 External links

703 References

704 Notes

705 See also

706 Further reading

707 Citations

708 External links

709 References

710 Notes

711 See also

712 Further reading

713 Citations

714 External links

715 References

716 Notes

717 See also

718 Further reading

719 Citations

720 External links

721 References

722 Notes

723 See also

724 Further reading

725 Citations

726 External links

727 References

728 Notes

729 See also

730 Further reading

731 Citations

732 External links

733 References

734 Notes

735 See also

736 Further reading

737 Citations

738 External links

739 References

740 Notes

741 See also

742 Further reading

743 Citations

744 External links

745 References

746 Notes

747 See also

748 Further reading

749 Citations

750 External links

751 References

752 Notes

753 See also

754 Further reading

755 Citations

756 External links

757 References

758 Notes

759 See also

760 Further reading

761 Citations

762 External links

763 References

764 Notes

765 See also

766 Further reading

767 Citations

768 External links

769 References

770 Notes

771 See also

772 Further reading

773 Citations

774 External links

775 References

776 Notes

777 See also

778 Further reading

779 Citations

780 External links

781 References

782 Notes

783 See also

784 Further reading

785 Citations

786 External links

787 References

788 Notes

789 See also

790 Further reading

791 Citations

792 External links

793 References

794 Notes

795 See also

796 Further reading

797 Citations

798 External links

799 References

800 Notes

801 See also

802 Further reading

803 Citations

804 External links

805 References

806 Notes

807 See also

808 Further reading

809 Citations

810 External links

811 References

812 Notes

813 See also

814 Further reading

815 Citations

816 External links

817 References

818 Notes

819 See also

820 Further reading

821 Citations

822 External links

823 References

824 Notes

825 See also

826 Further reading

827 Citations

828 External links

829 References

830 Notes

831 See also

832 Further reading

833 Citations

834 External links

835 References

836 Notes

837 See also

838 Further reading

839 Citations

840 External links

841 References

842 Notes

843 See also

844 Further reading

845 Citations

846 External links

847 References

848 Notes

849 See also

850 Further reading

851 Citations

852 External links

853 References

854 Notes

855 See also

856 Further reading

857 Citations

858 External links

859 References

860 Notes

861 See also

862 Further reading

863 Citations

864 External links

865 References

866 Notes

867 See also

868 Further reading

869 Citations

870 External links

871 References

872 Notes

873 See also

874 Further reading

875 Citations

876 External links

877 References

878 Notes

879 See also

880 Further reading

881 Citations

882 External links

883 References

884 Notes

885 See also

886 Further reading

887 Citations

888 External links

889 References

890 Notes

891 See also

892 Further reading

893 Citations

894 External links

895 References

896 Notes

897 See also

898 Further reading

899 Citations

900 External links

901 References

902 Notes

903 See also

904 Further reading

905 Citations

906 External links

907 References

908 Notes

909 See also

910 Further reading

911 Citations

912 External links

913 References

914 Notes

915 See also

916 Further reading

917 Citations

918 External links

919 References

920 Notes

921 See also

922 Further reading

923 Citations

924 External links

925 References

926 Notes

927 See also

928 Further reading

929 Citations

930 External links

931 References

932 Notes

933 See also

934 Further reading

935 Citations

936 External links

937 References

938 Notes

939 See also

940 Further reading

941 Citations

942 External links

943 References

944 Notes

945 See also

946 Further reading

947 Citations

948 External links

949 References

950 Notes

951 See also

952 Further reading

953 Citations

954 External links

955 References

956 Notes

957 See also

958 Further reading

959 Citations

960 External links

961 References

962 Notes

963 See also

964 Further reading

965 Citations

966 External links

967 References

968 Notes

969 See also

970 Further reading

971 Citations

972 External links

973 References

974 Notes

975 See also

976 Further reading

977 Citations

978 External links

979 References

980 Notes

981 See also

982 Further reading

983 Citations

984 External links

985 References

986 Notes

987 See also

988 Further reading

989 Citations

990 External links

991 References

992 Notes

993 See also

994 Further reading

995 Citations

996 External links

997 References

998 Notes

999 See also

1000 Further reading

1001 Citations

1002 External links

1003 References

1004 Notes

1005 See also

1006 Further reading

1007 Citations

1008 External links

1009 References

1010 Notes

1011 See also

1012 Further reading

1013 Citations

1014 External links

1015 References

1016 Notes

1017 See also

1018 Further reading

1019 Citations

1020 External links

1021 References

1022 Notes

1023 See also

1024 Further reading

1025 Citations

1026 External links

1027 References

1028 Notes

1029 See also

1030 Further reading

1031 Citations

1032 External links

1033 References

1034 Notes

1035 See also

1036 Further reading

1037 Citations

1038 External links

1039 References

1040 Notes

1041 See also

1042 Further reading

1043 Citations

1044 External links

1045 References

1046 Notes

1047 See also

1048 Further reading

1049 Citations

1050 External links

1051 References

1052 Notes

1053 See also

1054 Further reading

1055 Citations

1056 External links

1057 References

1058 Notes

1059 See also

1060 Further reading

1061 Citations

1062 External links

1063 References

1064 Notes

1065 See also

1066 Further reading

1067 Citations

1068 External links

1069 References

1070 Notes

1071 See also

1072 Further reading

1073 Citations

1074 External links

1075 References

1076 Notes

1077 See also

1078 Further reading

1079 Citations

1080 External links

1081 References

1082 Notes

1083 See also

1084 Further reading

1085 Citations

1086 External links

1087 References

1088 Notes

1089 See also

1090 Further reading

1091 Citations

1092 External links

1093 References

1094 Notes

1095 See also

1096 Further reading

1097 Citations

1098 External links

1099 References

1100 Notes

1101 See also

1102 Further reading

1103 Citations

1104 External links

1105 References

1106 Notes

1107 See also

1108 Further reading

1109 Citations

1110 External links

1111 References

1112 Notes

1113 See also

1114 Further reading

1115 Citations

1116 External links

1117 References

1118 Notes

1119 See also

1120 Further reading

1121 Citations

1122 External links

1123 References

1124 Notes

1125 See also

1126 Further reading

1127 Citations

1128 External links

1129 References

1130 Notes

1131 See also

1132 Further reading

1133 Citations

1134 External links

1135 References

1136 Notes

1137 See also

1138 Further reading

1139 Citations

1140 External links

1141 References

1142 Notes

1143 See also

1144 Further reading

1145 Citations

1146 External links

1147 References

1148 Notes

1149 See also

1150 Further reading

1151 Citations

1152 External links

1153 References

1154 Notes

1155 See also

1156 Further reading

1157 Citations

1158 External links

1159 References

1160 Notes

1161 See also

1162 Further reading

1163 Citations

1164 External links

1165 References

1166 Notes

1167 See also

1168 Further reading

1169 Citations

1170 External links

1171 References

1172 Notes

1173 See also

1174 Further reading

1175 Citations

1176 External links

1177 References

1178 Notes

1179 See also

1180 Further reading

1181 Citations

1182 External links

1183 References

1184 Notes

1185 See also

1186 Further reading

1187 Citations

1188 External links

1189 References

1190 Notes

1191 See also

1192 Further reading

1193 Citations

1194 External links

1195 References

1196 Notes

1197 See also

1198 Further reading

1199 Citations

1200 External links

1201 References

1202 Notes

1203 See also

1204 Further reading

1205 Citations

1206 External links

1207 References

1208 Notes

1209 See also

1210 Further reading

1211 Citations

1212 External links

1213 References

1214 Notes

1215 See also

1216 Further reading

1217 Citations

1218 External links

1219 References

1220 Notes

1221 See also

1222 Further reading

1223 Citations

1224 External links

1225 References

1226 Notes

1227 See also

1228 Further reading

1229 Citations

1230 External links

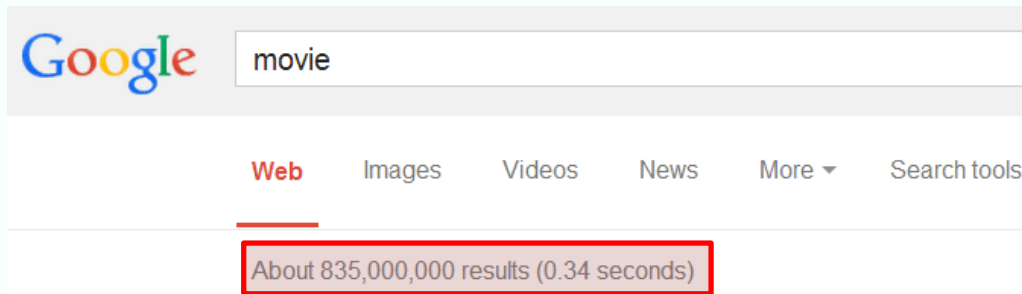
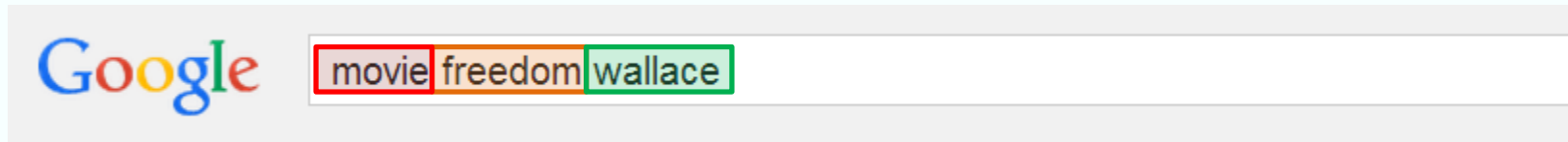
1231 References

1232 Notes

1233 See also

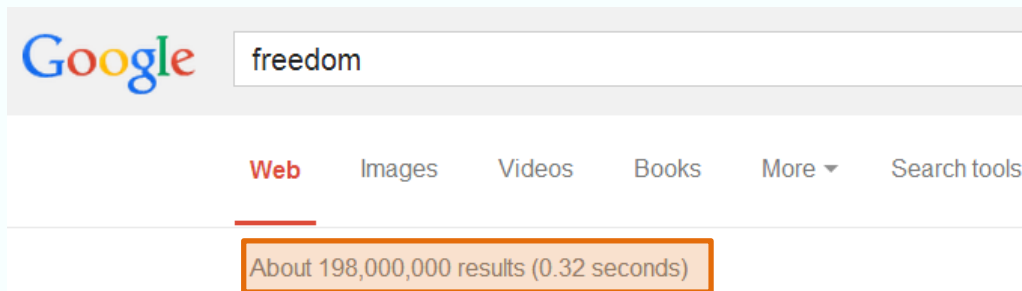
12

Usefulness of Words



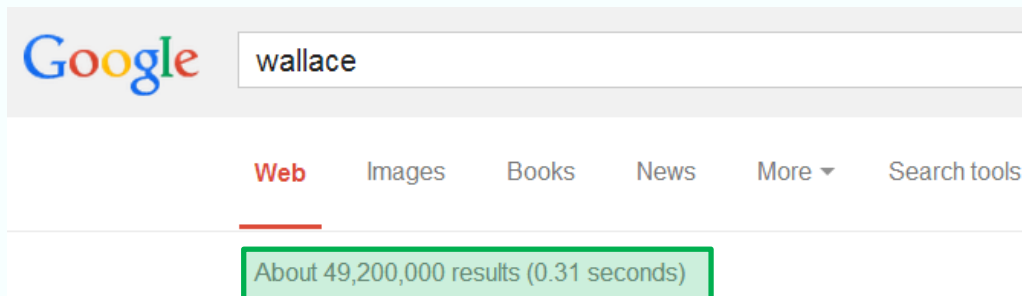
movie

- 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

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

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

- 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

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

- Note: The more rare, the larger the value

Relevance Measure: TF-IDF

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

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

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

(There are other possibilities)

Relevance Measure: TF-IDF



Term Frequency

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

Inverse Document Frequency

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

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

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

Relevance Measure: TF-IDF



Term Frequency

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

Inverse Document Frequency

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

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

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

Relevance Measure: TF-IDF



Term Frequency

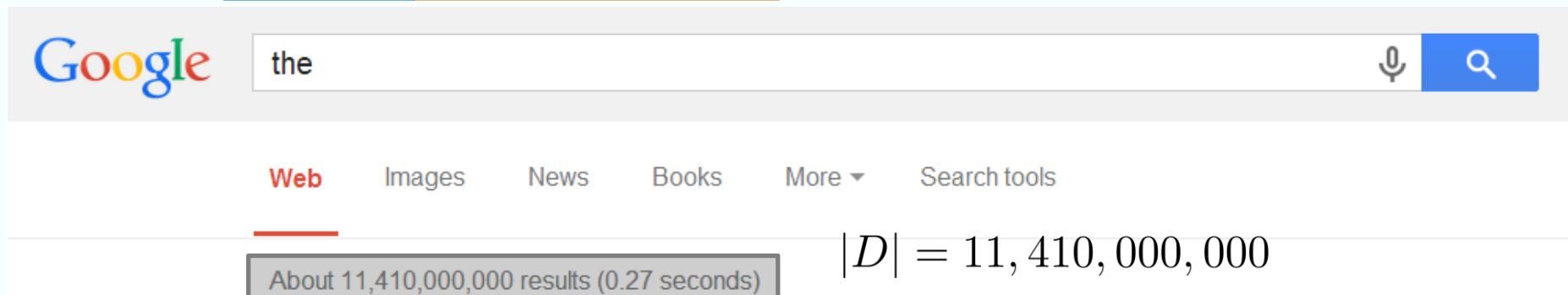
$$\text{tf}(t, d) = \text{count}(t, d)$$

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	



Relevance Measure: TF-IDF



Term Frequency

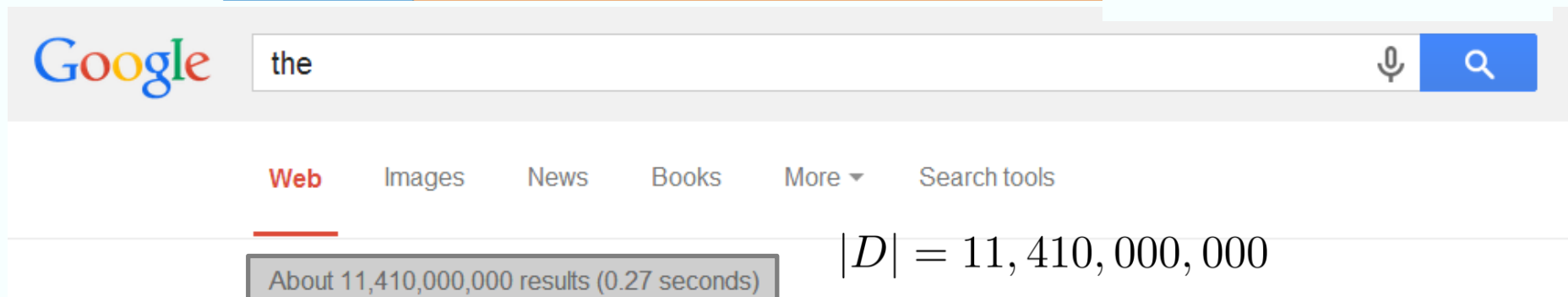
$$\text{tf}(t, d) = \text{count}(t, d)$$

Inverse Document Frequency

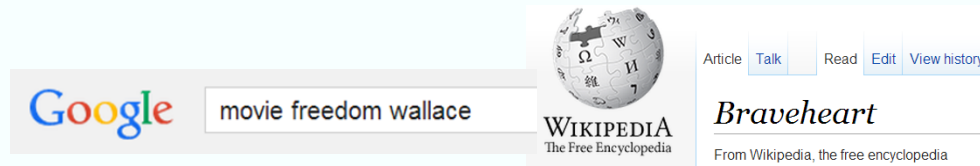
$$\text{idf}(t, D) = \log_2 \left(\frac{|D|}{|\{d \in D : t \in d\}| + 1} \right)$$

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

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



Relevance Measure: TF-IDF



Term Frequency

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

Inverse Document Frequency

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

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

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

Relevance Measure: TF-IDF



Term Frequency

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

Inverse Document Frequency

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

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

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

Relevance Measure: TF-IDF



Term Frequency

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

Inverse Document Frequency

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

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

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

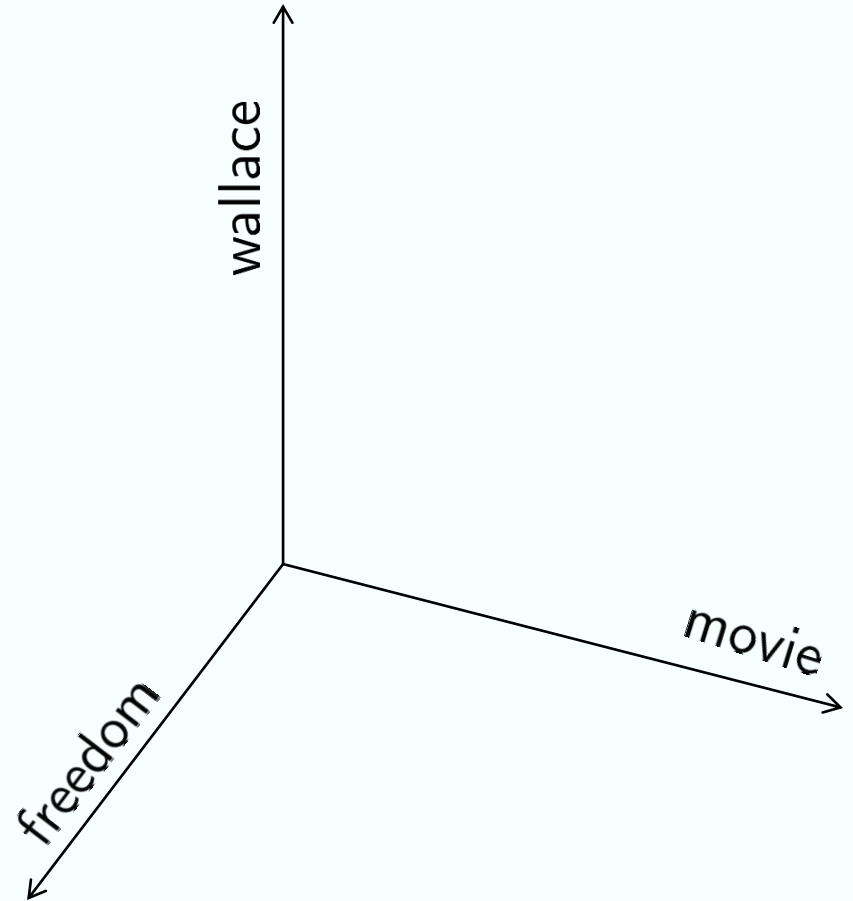
$$\text{score}(q, d) = \sum_{t \in q} \text{tf-idf}(t, d)$$

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

Vector Space Model (a mention)

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

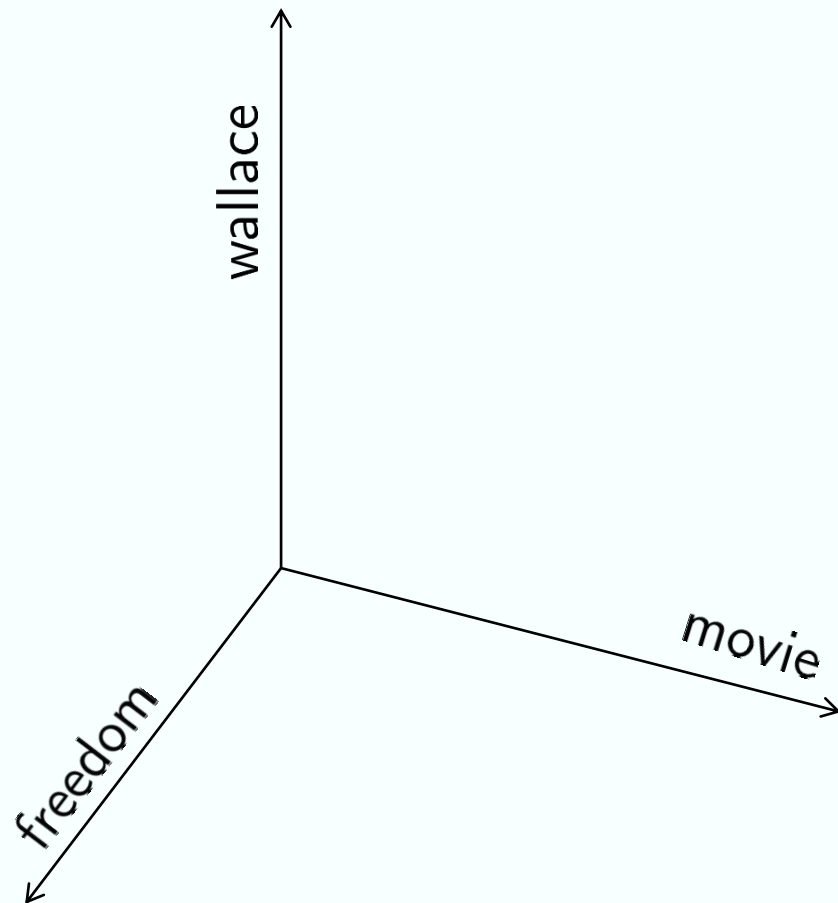
$$l = \sqrt{\sum_{t \in q} \text{tf}(t, d)^2}$$



Vector Space Model (a mention)

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

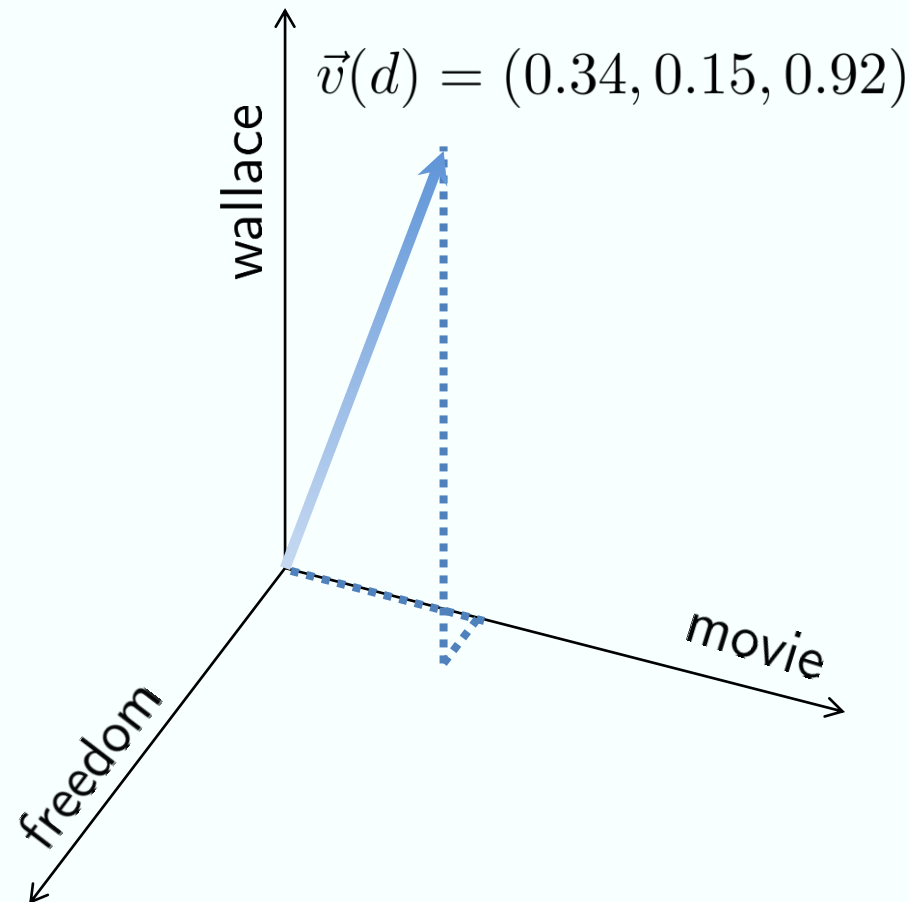
$$l = \sqrt{\sum_{t \in q} \text{tf}(t, d)^2}$$



Vector Space Model (a mention)

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

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



Dividing by l normalises length of vector to 1

Vector Space Model (a mention)

- Cosine Similarity

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

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

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

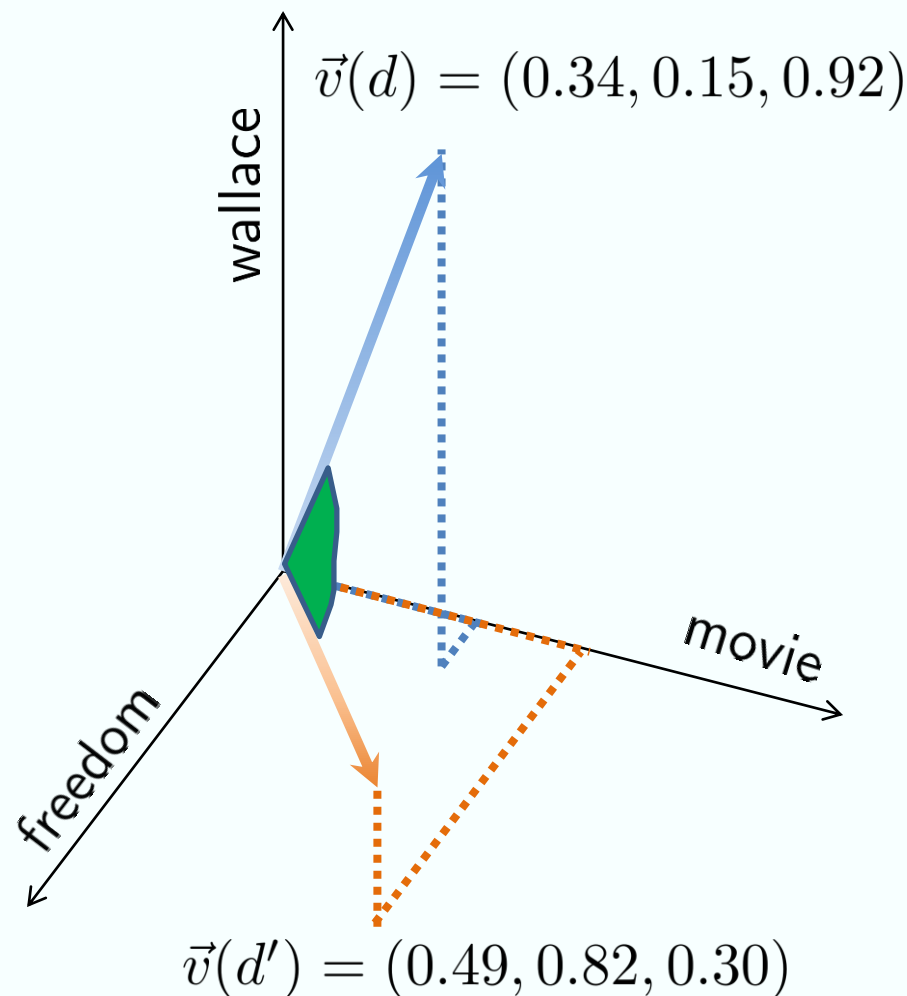
Σ

- Note:

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

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

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



Relevance Measure: TF-IDF

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

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

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

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

Two Sides to Ranking: Relevance



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) - 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 Eggs
Huffington Post - 2 days ago
But in her new book Brassicas: The World's Healthiest Vegetables, she says plunking **broccoli**, cauliflower or Brussels sprouts into ...

Field-Based Boosting

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

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



The screenshot shows the Wikipedia page for Barack Obama. The title "Barack Obama - Wikipedia, the free encyclopedia" is highlighted in the HTML code above. On the page, the title "Barack Obama" is highlighted with an orange box. The page includes a sidebar with navigation links, a search bar, and a main content area with a biography of Barack Obama. The biography text is partially visible, starting with "Barack Hussein Obama II" and mentioning his birth date and role as the 44th President of the United States.

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

Search

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 (ⁱ/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

Barack Obama



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)



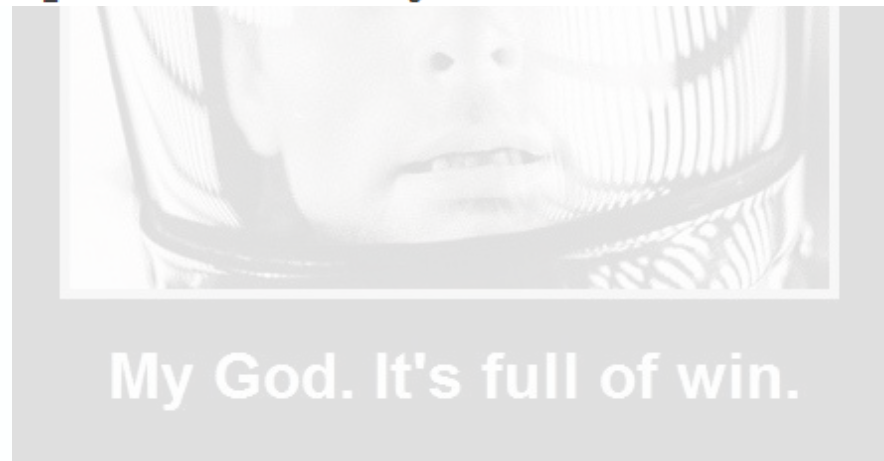
Beethoven - Symphony No. 5 in C Minor (1) - YouTube

www.youtube.com/watch?v=W2qW6fOtAMY

Lucene uses relevance scoring

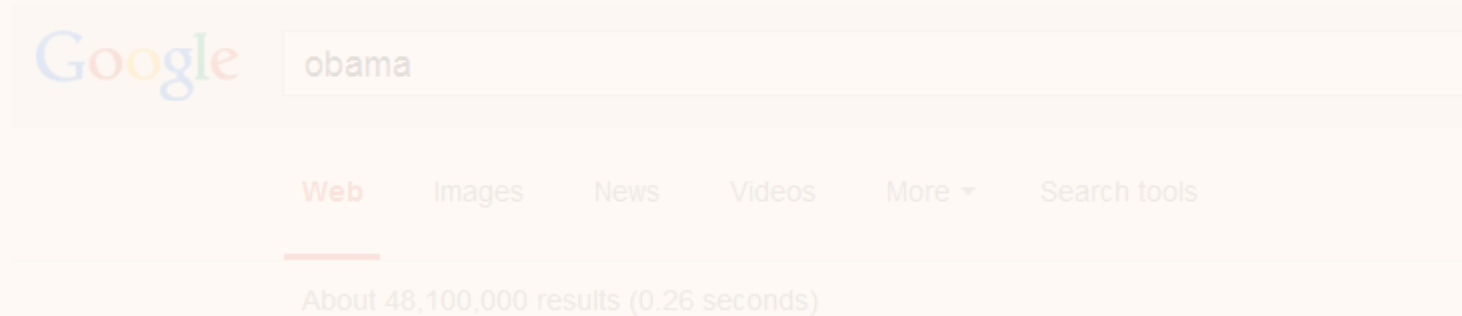


```
Tasks Console
SearchWikiIndex [Java Application] C:\Program Files\Java\jre1.8.0_65\bin\javaw.exe (03-05-2017 12:45:22 a. m.)
Opening directory at lucene
Enter a keyword search phrase:
obama
Running query: obama
Parsed query: TITLE:obam^5.0 ABSTRACT:obam
Matching documents: 255
Showing top 10 results
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/Transición_presidencial_de_Barack_Obama Transición presidencial de Barack Obama
9 http://es.wikipedia.org/wiki/Por_qué_Obama_ganará_en_2008_y_en_2012 Por qué Obama ganará en 2008 y en 2012
10 http://es.wikipedia.org/wiki/Ricardo_Mangue_Obama_Nfubea Ricardo Mangue Obama Nfubea
```



RANKING:
IMPORTANCE

Two Sides to Ranking: Importance

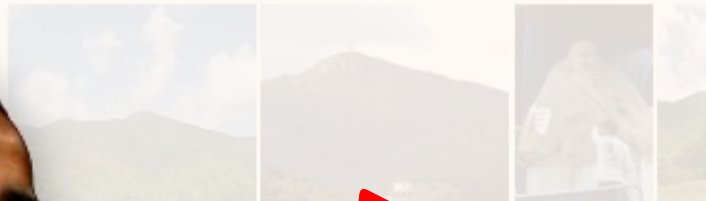


How could we determine that Barack Obama is more important than Mount Obama as a search result on the Web?



Images for mount obama

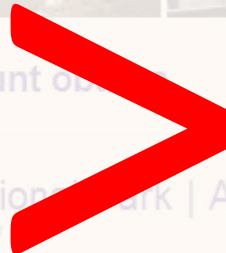
Report images



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 O



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](#) [Deltasone](#) [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](#) [Himcoid](#) [Himcolin](#) [Himcospaz](#) [Himplasia](#) [Hoodia](#) [Hytrin](#)
[Hyzaar](#) [Imdur](#) [Imitrex](#) [Inderal](#) [Ismo](#) [Isoptin](#) [Isordil](#) [Kamagra](#) [Karela](#) [Keftab](#) [Koflet](#) [Kytril](#) [Lamictal](#)
[Lamisil](#) [Lanoxin](#) [Lariam](#) [Lasix](#) [Lasuna](#) [Leukeran](#) [Levaquin](#) [Levlen](#) [Levothroid](#) [Lincocin](#) [Lioresal](#)
[Lisinopril](#) [Liv.52](#) [Lopid](#) [Lopressor](#) [Loprox](#) [Lotensin](#) [Lotrisone](#) [Loxitane](#) [Lozol](#) [Lukol](#) [Lynoral](#)
[Maxaquin](#) [Menosan](#) [Mentat](#) [Mentax](#) [Mevacor](#) [Mexitil](#) [Miacalcin](#) [Micardis](#) [Mobic](#) [Monoket](#) [Motrin](#)
[Myambutol](#) [Mycellex-G](#) [Mysoline](#) [Naprosyn](#) [Neurontin](#) [Nicotinell](#) [Nimotop](#) [Nirdosh](#) [Nizoral](#)
[Nolvadex](#) [Nonoxinol](#) [Noroxin](#) [Omnicef](#) [Ophthalmicare](#) [Oxytrol](#) [Pamelor](#) [Parlodel](#) [Paxil](#) [Penisole](#)
[Phentermine](#) [Pilex](#) [Plan B](#) [Plavix](#) [Plendil](#) [Pletal](#) [Prandin](#) [Pravachol](#) [Prednisone](#) [Pregamin](#) [Prevacid](#)
[Prilosec](#) [Prinivil](#) [Procardia](#) [Prograf](#) [Prometrium](#) [Propecia](#) [Proscar](#) [Protonix](#) [Proventil](#) [Prozac](#) [Purim](#)
[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](#) [Stretch](#) [Nil](#) [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 Lamp](#) [Coral Calcium](#)
[Cialis Jelly](#) [Cialis Soft Tabs](#) [Calcium Carbonate](#) [Bust Enhancer](#) [Beconase AQ](#) [Anatrim 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](#)



black magic specialist baba ji call now +919610897260



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



java

edit | close | undelete | more ▼

Link Importance

So which should count for more?

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

Or a link from <http://freev1agra.com/shop.html>?



Link Importance

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



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

PageRank



PageRank

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

PageRank is Recursive

- Not just a count of inlinks
 - A link from a more important page is more important
 - A link from a page with lots of inlinks is less important
 - ∴ A page with lots of inlinks and few outgoing links is more important



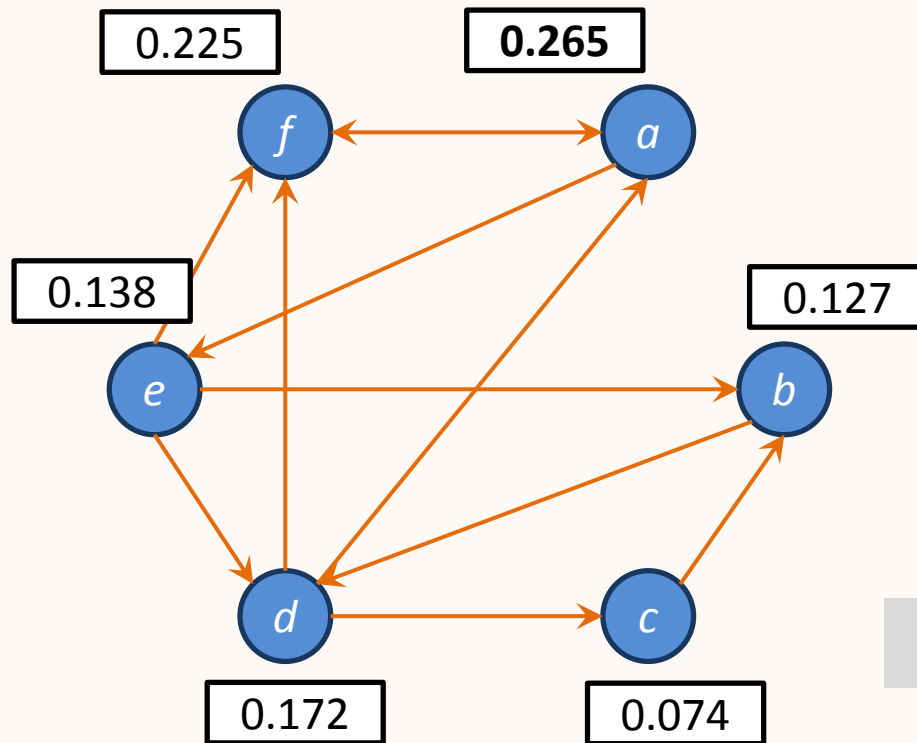
PageRank Model

- The Web: a directed graph

$$G = [V, E]$$

Vertices
(pages)

Edges
(links)



Which vertex is most important?



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

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

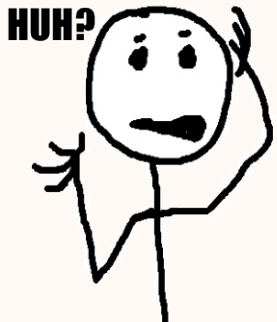
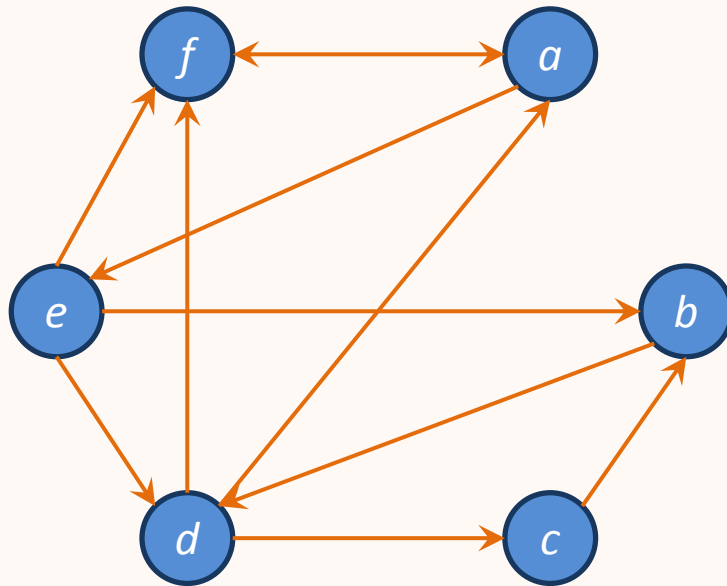
PageRank Model

- The Web: a directed graph

$$G = [V, E]$$

Vertices
(pages)

Edges
(links)



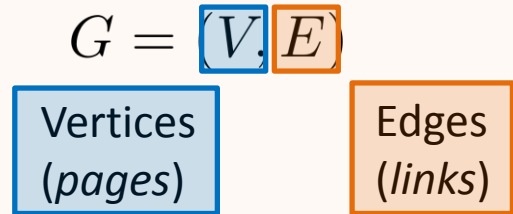
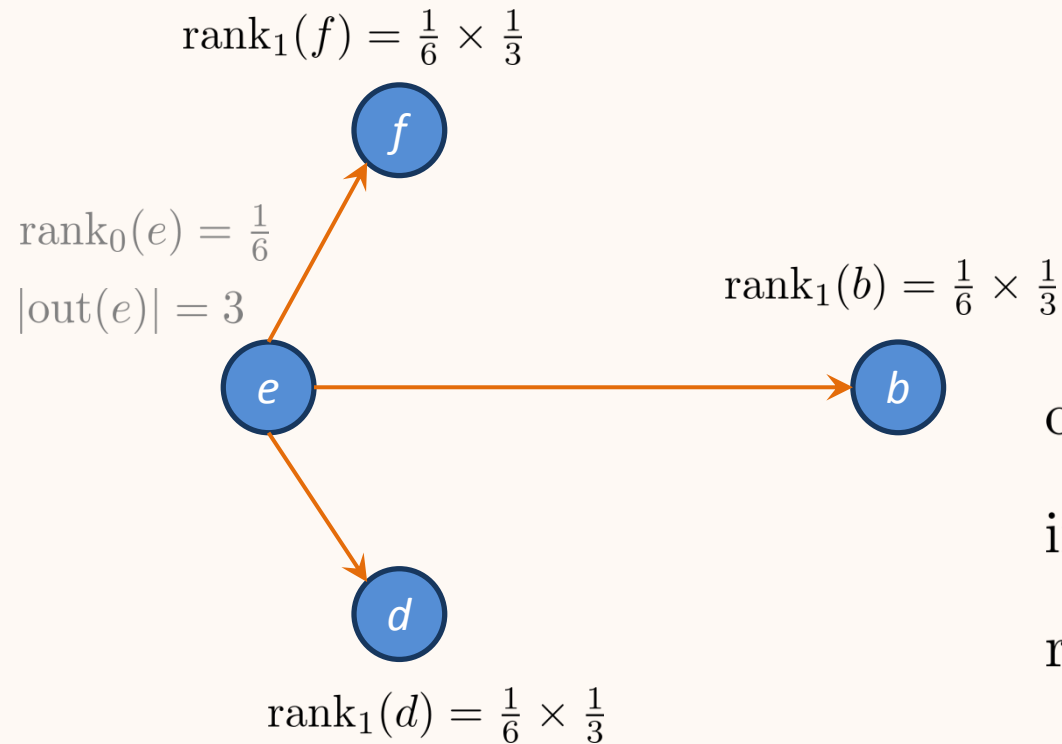
$$\text{out}(v) \doteq \{v' \in V \mid (v, v') \in E\}$$

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

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

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

PageRank Model



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

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

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

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

PageRank Model

$$\text{rank}_2(a) = \frac{1}{6} \times \frac{1}{3} \times 1 + \frac{1}{6} \times \frac{1}{3} \times \frac{1}{3}$$

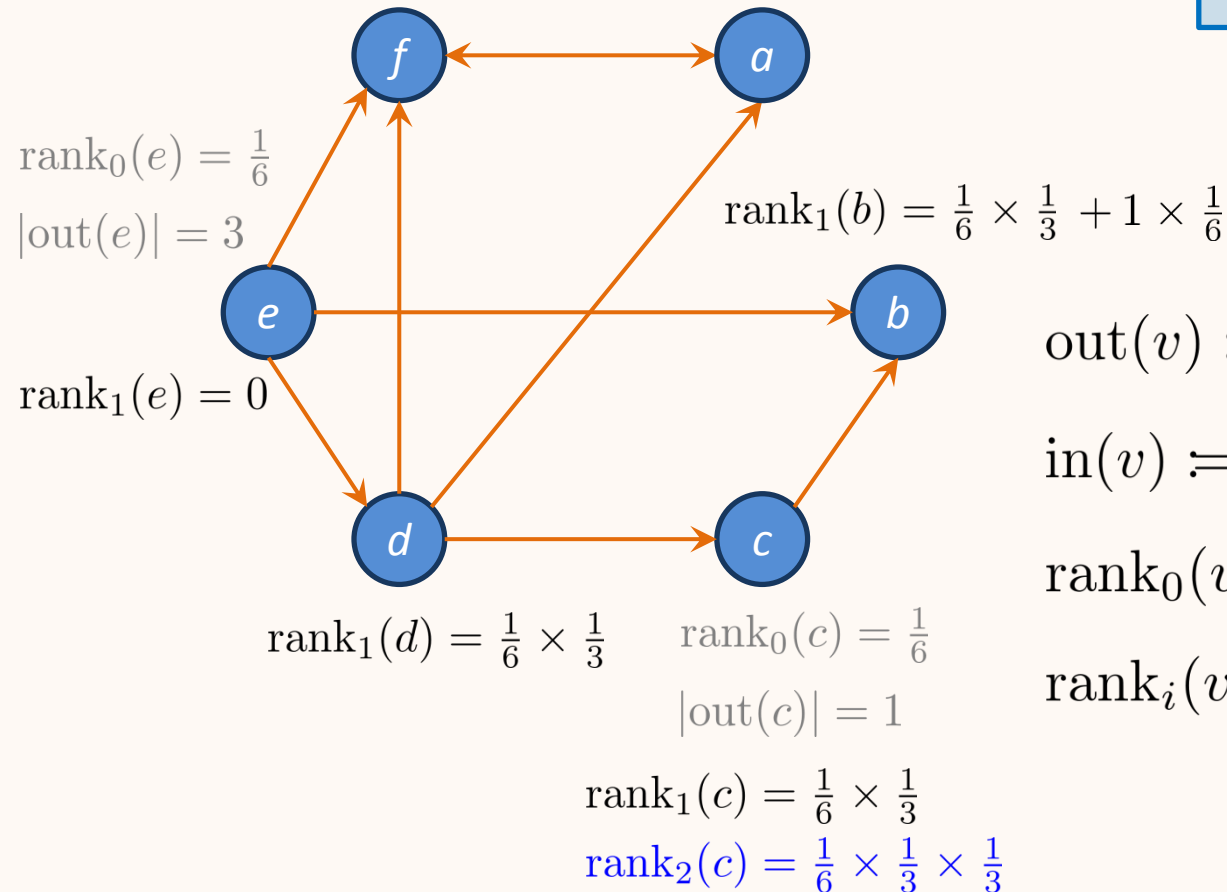
$$G = [V, E]$$

Vertices
(pages)

Edges
(links)

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

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



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

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

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

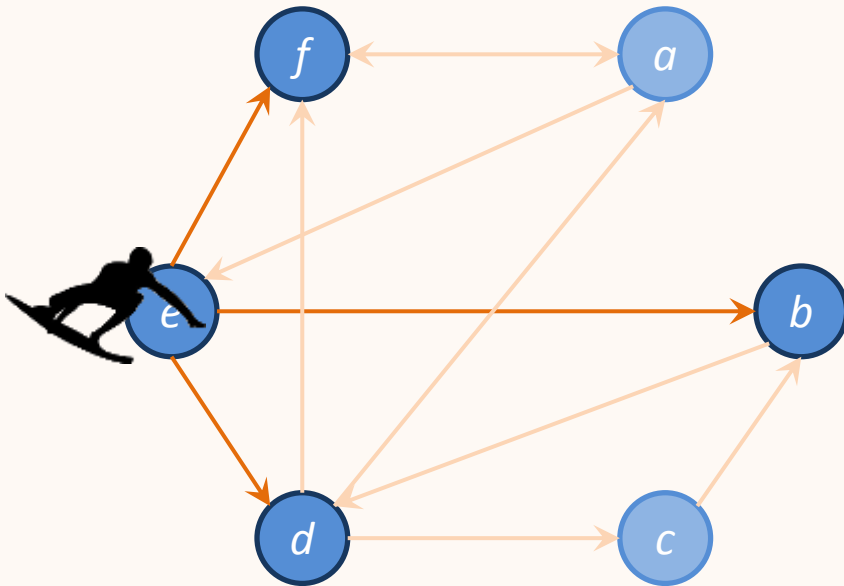
$$\text{rank}_i(v) \doteq \sum_{v' \in \text{in}(v)} \frac{\text{rank}_{i-1}(v')}{|\text{out}(v')|}$$

PageRank: Random Surfer Model



= someone surfing the web,
clicking links randomly

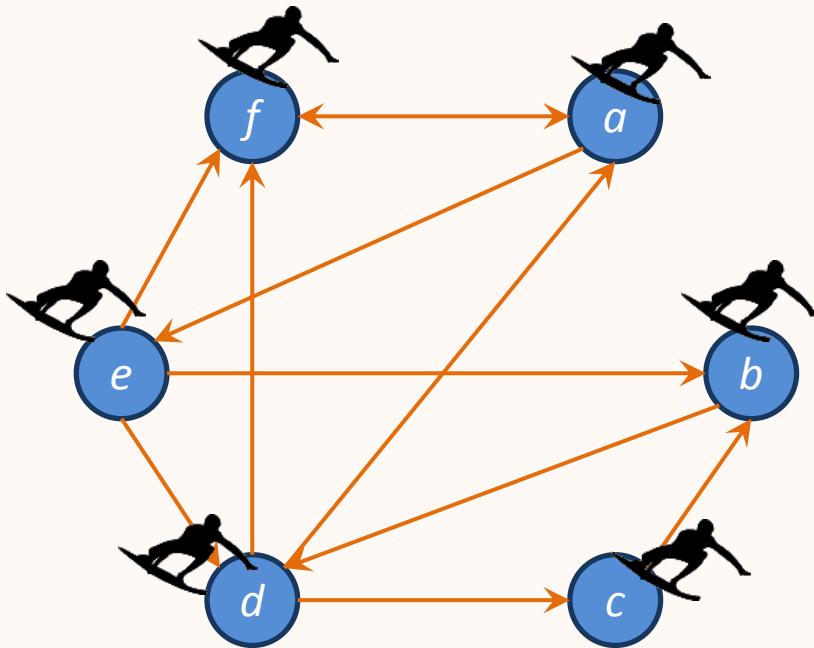
- What is the probability of being at page x after n hops?



PageRank: Random Surfer Model



= someone surfing the web,
clicking links randomly

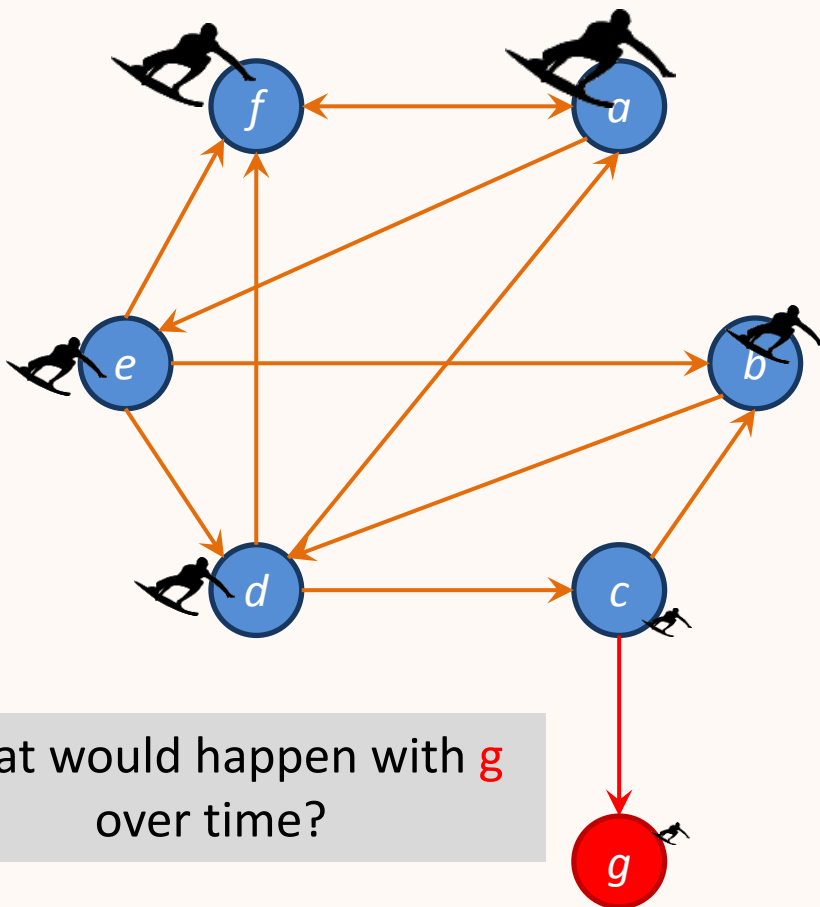


- What is the probability of being at page x after n hops?
- *Initial state*: surfer equally likely to start at any node

PageRank: Random Surfer Model



= someone surfing the web,
clicking links randomly



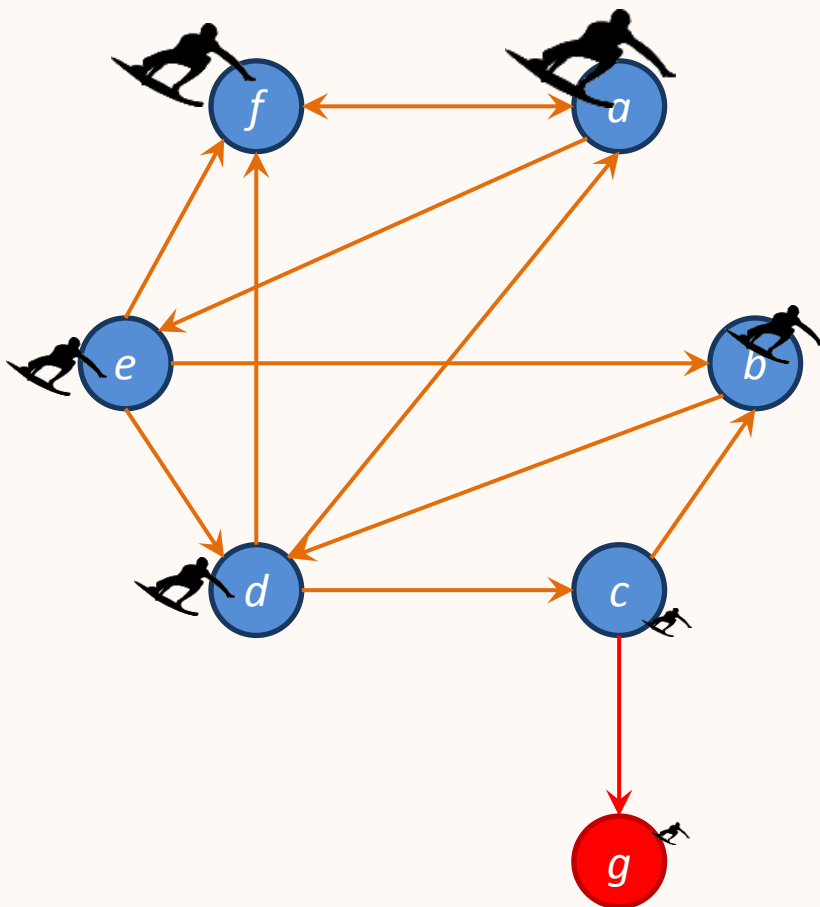
What would happen with **g**
over time?

- 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

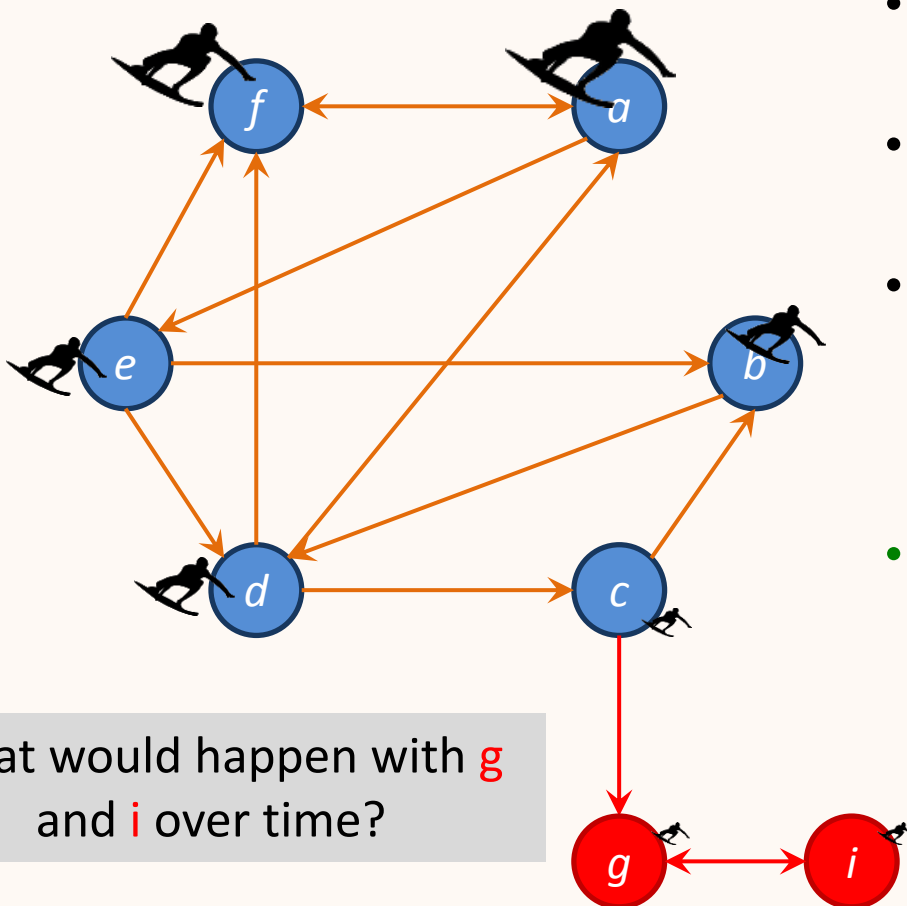


- What is the probability of being at page x after n hops?
- *Initial state*: surfer equally likely to start at any node
- PageRank applied iteratively for each hop: score indicates probability of being at that page after than many hops
- If the surfer reaches a page without links, the surfer randomly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

PageRank: Random Surfer Model



= someone surfing the web,
clicking links randomly



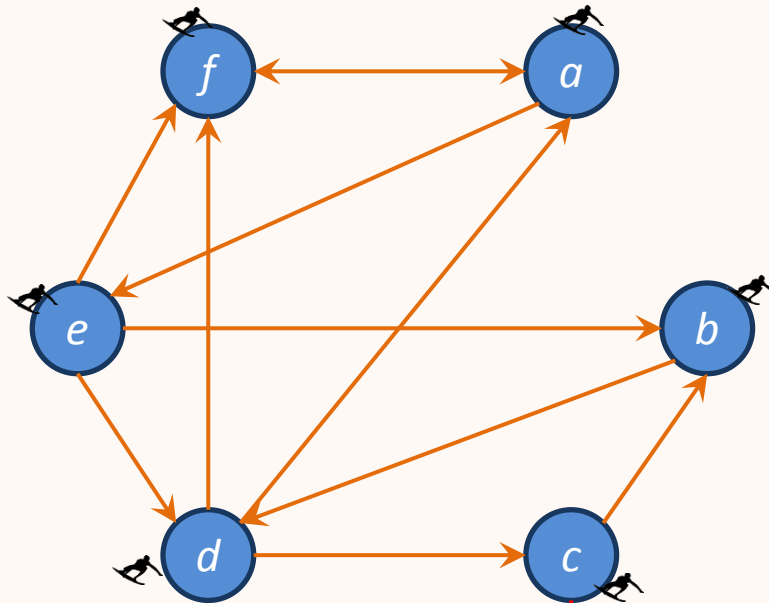
What would happen with **g**
and **i** over time?

- 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



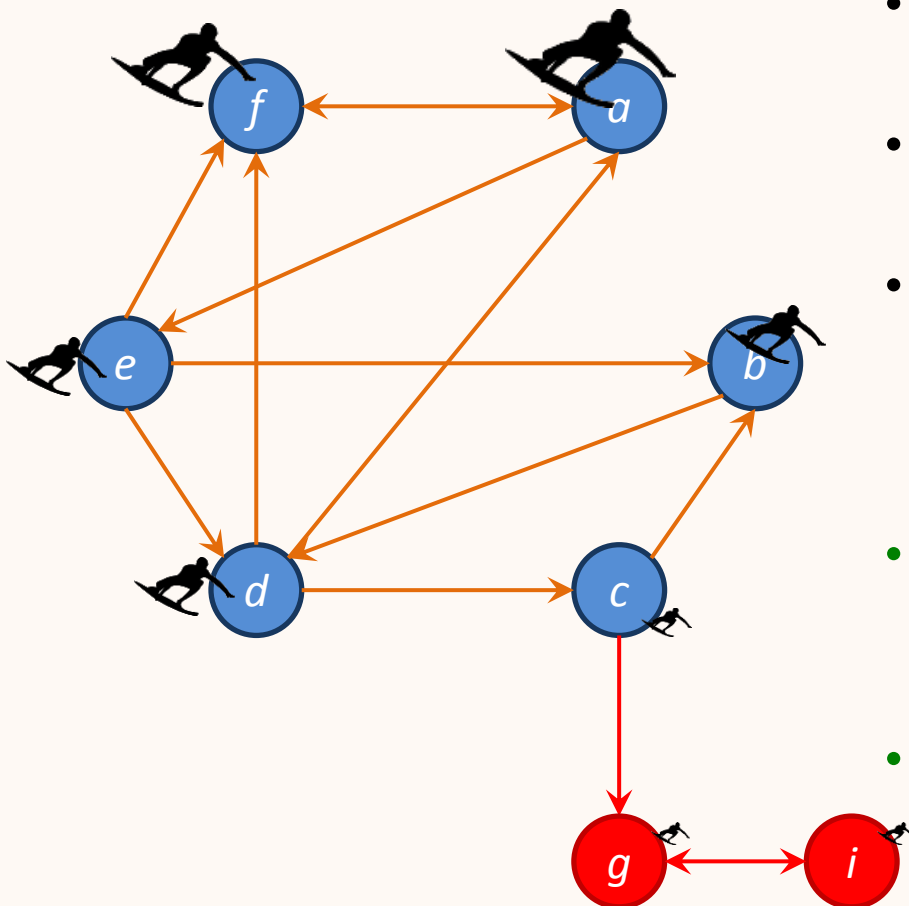
What would happen with **g**
and **i** over time?

- 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



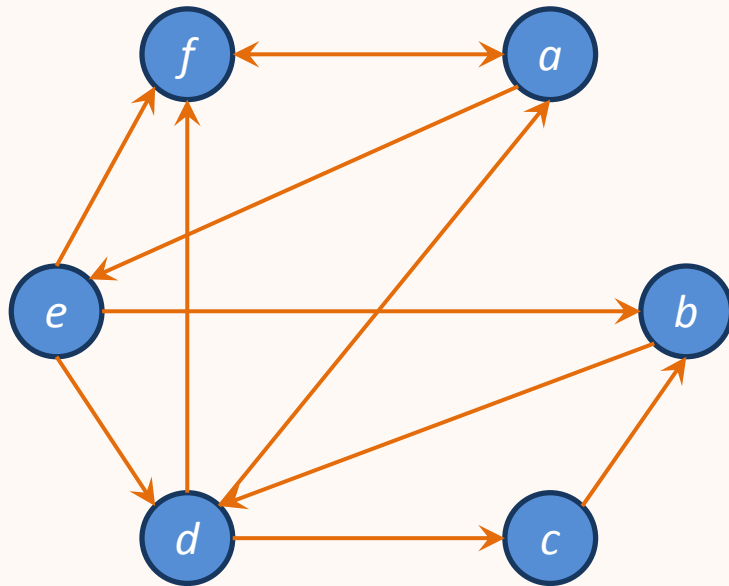
= someone surfing the web,
clicking links randomly



- 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
- 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)

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

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

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

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

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

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

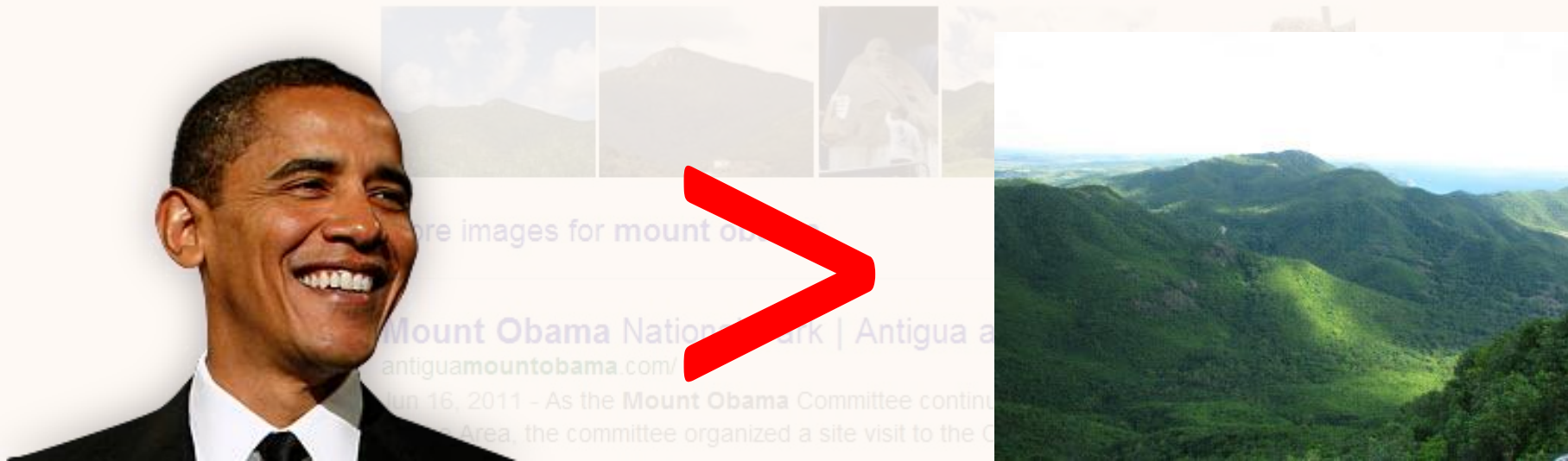
$$\text{rank}_i(v) \coloneqq d \times \sum_{u \in \text{in}(v)} \frac{\text{rank}_{i-1}(u)}{|\text{out}(u)|} + \sum_{v' \in V'} \frac{\text{rank}_{i-1}(v')}{|V|} + (1-d) \times \sum_{v'' \in V''} \frac{\text{rank}_{i-1}(v'')}{|V|}$$

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

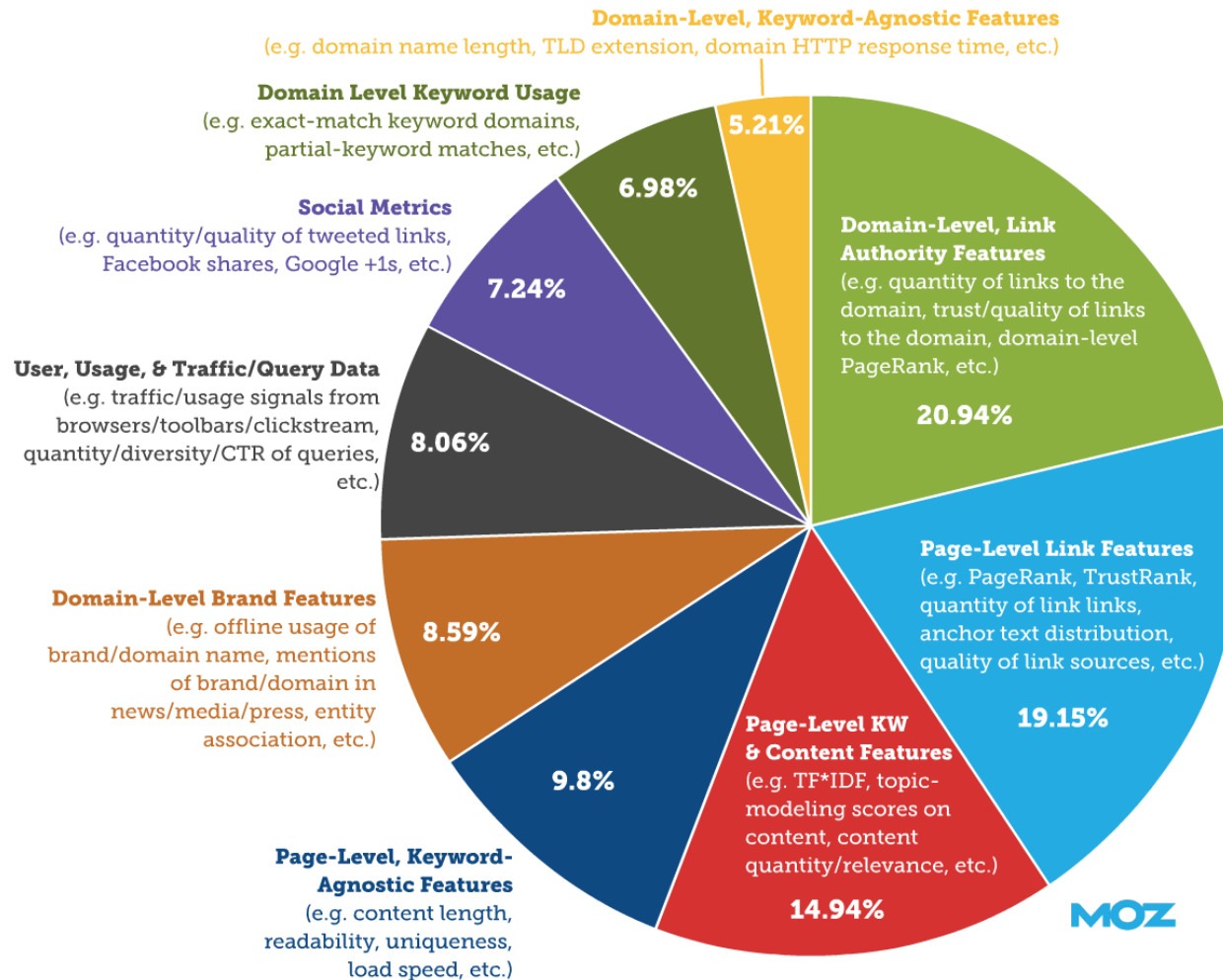


GOOGLE: A GUESS

How Modern Google ranks results (maybe)

Weighting of Thematic Clusters of Ranking Factors in Google

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

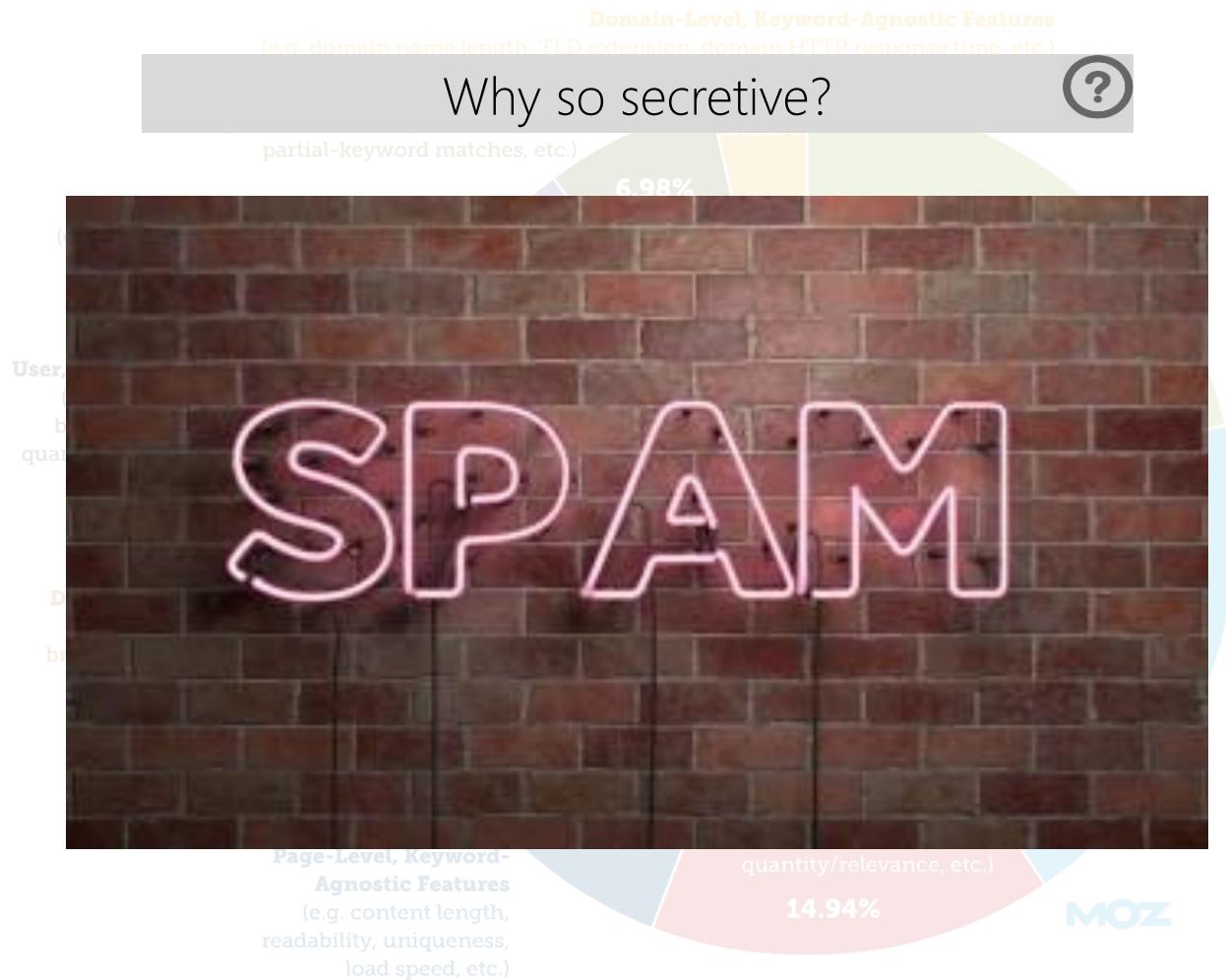


According to survey of SEO experts, not people in Google

How Modern Google ranks results (maybe)

Weighting of Thematic Clusters of Ranking Factors in Google

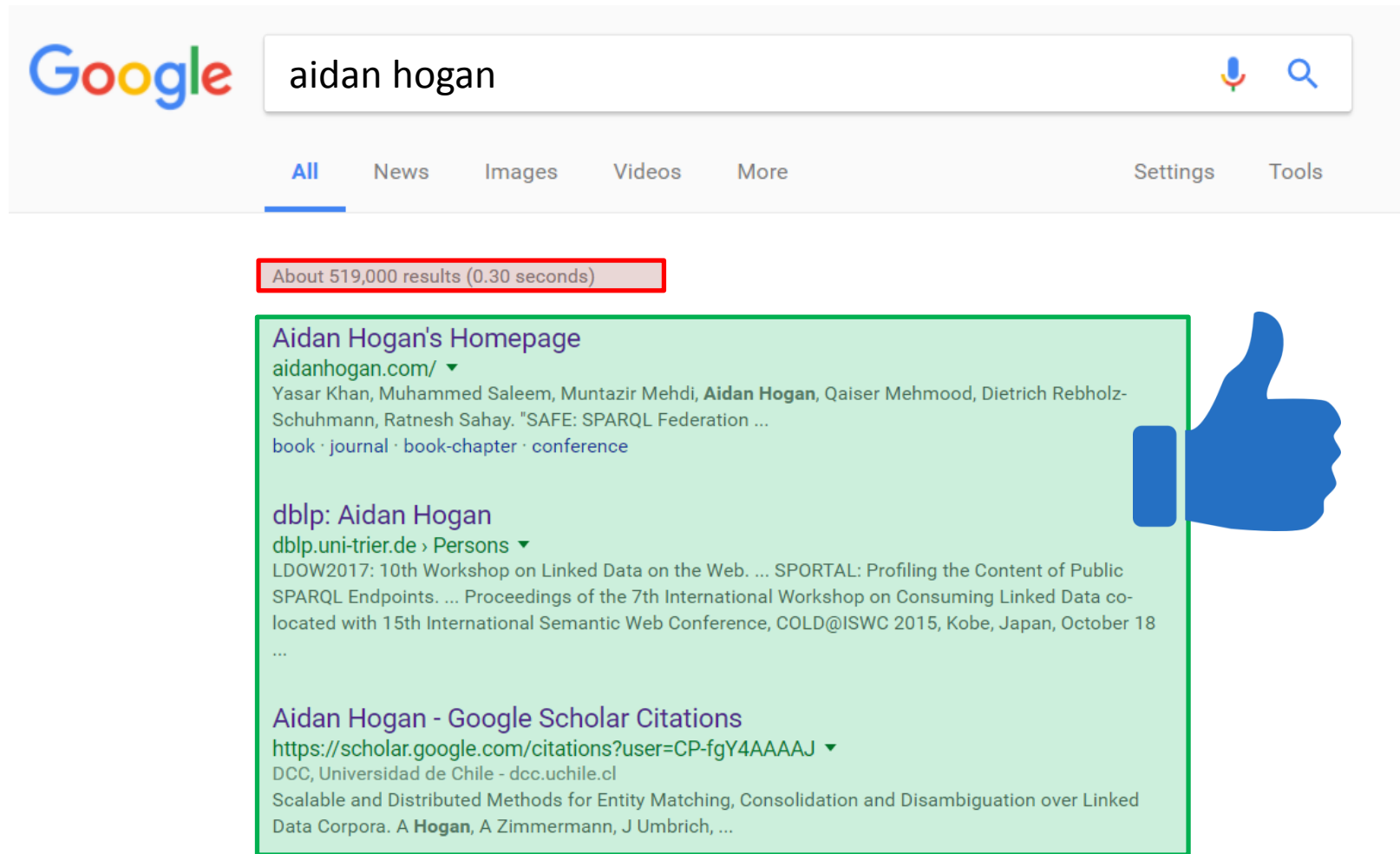
(based on survey responses by 128 SEO professionals in June 2013)



According to survey of SEO experts, not people in Google

INFORMATION RETRIEVAL: RECAP

How Does Google Get Such Good Results?



The image shows a Google search interface. The search bar contains the text "aidan hogan". Below the search bar, the "All" tab is selected. The search results show "About 519,000 results (0.30 seconds)". The first result is "Aidan Hogan's Homepage" with the URL "aidanhogan.com/". The second result is "dblp: Aidan Hogan" with the URL "dblp.uni-trier.de > Persons". The third result is "Aidan Hogan - Google Scholar Citations" with the URL "https://scholar.google.com/citations?user=CP-fgY4AAAAJ". A large blue thumbs up icon is positioned to the right of the search results.

Google

aidan hogan

All News Images Videos More Settings Tools

About 519,000 results (0.30 seconds)

Aidan Hogan's Homepage
aidanhogan.com/ ▼
Yasar Khan, Muhammed Saleem, Muntazir Mehdi, **Aidan Hogan**, Qaiser Mehmood, Dietrich Rebholz-Schuhmann, Ratnesh Sahay. "SAFE: SPARQL Federation ...
book · journal · book-chapter · conference

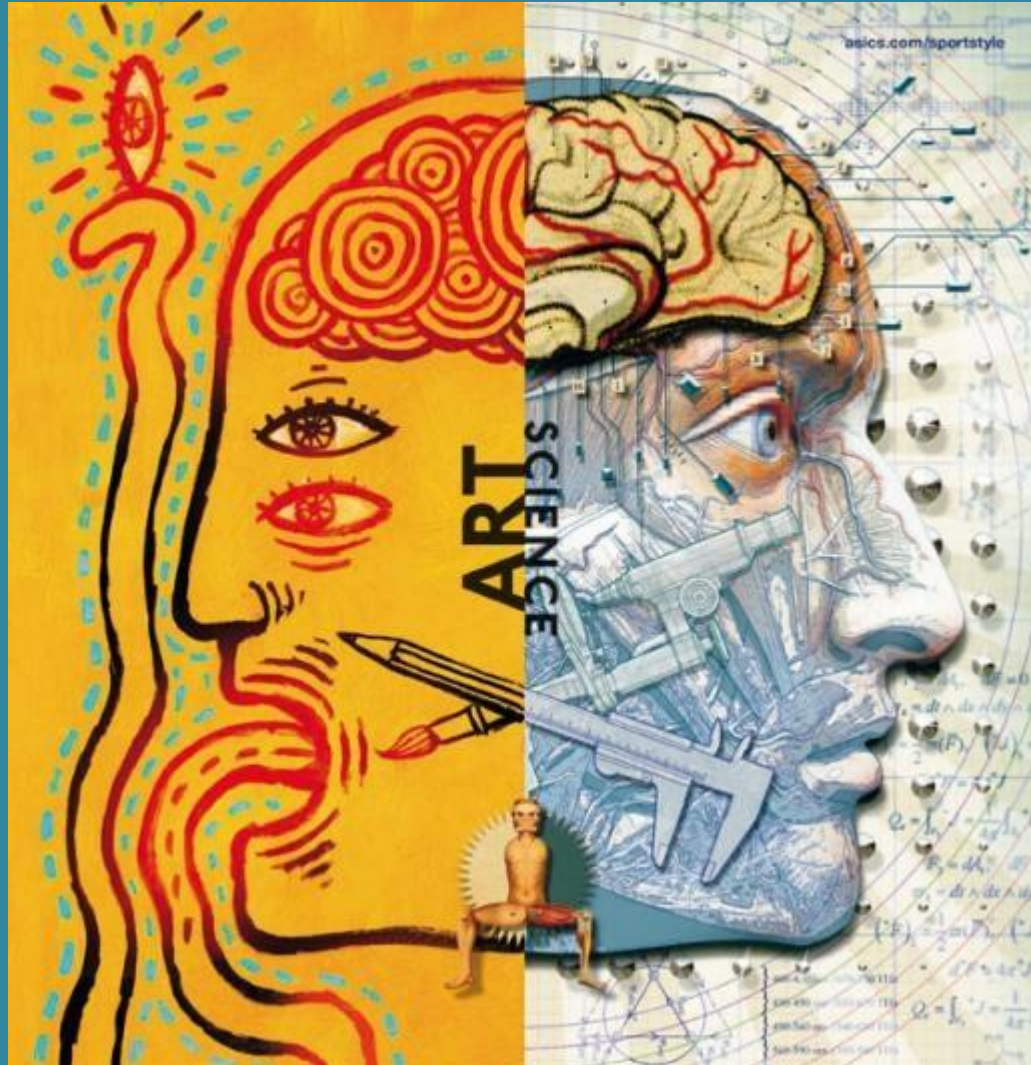
dblp: Aidan Hogan
dblp.uni-trier.de > Persons ▼
LDOW2017: 10th Workshop on Linked Data on the Web. ... SPOTAL: Profiling the Content of Public SPARQL Endpoints. ... Proceedings of the 7th International Workshop on Consuming Linked Data co-located with 15th International Semantic Web Conference, COLD@ISWC 2015, Kobe, Japan, October 18 ...

Aidan Hogan - Google Scholar Citations
https://scholar.google.com/citations?user=CP-fgY4AAAAJ ▼
DCC, Universidad de Chile - dcc.uchile.cl
Scalable and Distributed Methods for Entity Matching, Consolidation and Disambiguation over Linked Data Corpora. A **Hogan**, A Zimmermann, J Umbrich, ...

Ranking in Information Retrieval

- **Relevance**: Is the document relevant for the query?
 - Term Frequency * Inverse Document Frequency
 - Cosine similarity
- **Importance**: Is the document a popular one?
 - Links analysis
 - PageRank

Ranking: Science or Art?





Questions?