

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text
PageRank
Other tweaks

Video Search and Recommendation

Summary

Web Search

COMP90049 Knowledge Technologies

Lea Frermann and Justin Zobel and Karin Verspoor, CIS

Semester 2, 2019



THE UNIVERSITY OF

MELBOURNE

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text
PageRank
Other tweaks

Video Search and
Recommendation

Summary

So far,

- indexing
- boolean and ranked retrieval
- each document was considered independently

But,

- **Link structure** of the web is highly informative
consider: personal blogs, wedding planning sites, ...
vs.: news sites, wikipedia articles, ...
- # incoming links predict page's **importance** (to general public)
- # incoming links predict page's **authority**
- Anchor text
- PageRank

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text

PageRank

Other tweaks

Video Search and
Recommendation

Summary

For more information on unit tests, please go to `the unit test page`.

Anchor text, is crucial for web search. For example,

- There are many thousands of ‘Library’ references in the unimelb web site.
- The Library home page does not mention the word often (hardly any plain text at all).
- Most of the within-Unimelb ‘library’ links point to the Library home page.
- Most of the links to the home page contain the word ‘library’.

Anchor text is treated as a form of **zone**.

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text

PageRank

Other tweaks

Video Search and Recommendation

Summary

For more information on unit tests, please go to `the unit test page`.

Why useful?

- topic indicator
- very concise (easy to index, typically unambiguous)
- we usually get many of those per (important) page

Link = vote of importance (independent of anchor text content).
PageRank uses this fact.

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text

PageRank

Other tweaks

Video Search and Recommendation

Summary

Let me introduce **Bob**. Bob is bored.

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text

PageRank

Other tweaks

Video Search and Recommendation

Summary

Let me introduce **Bob**. Bob is bored.

So bored that he spends his days **clicking through webpages**.

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text
PageRank
Other tweaks

Video Search and Recommendation

Summary

Let me introduce **Bob**. Bob is bored.

So bored that he spends his days **clicking through webpages**.

Bob's keyboard has a **teleport** button which takes him to a random website

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text
PageRank
Other tweaks

Video Search and Recommendation

Summary

Let me introduce **Bob**. Bob is bored.

So bored that he spends his days **clicking through webpages**.

Bob's keyboard has a **teleport** button which takes him to a random website

At each webpage, Bob decides to

EITHER click a random link on that webpage

OR click his teleport button

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text
PageRank
Other tweaks

Video Search and Recommendation

Summary

Let me introduce **Bob**. Bob is bored.

So bored that he spends his days **clicking through webpages**.

Bob's keyboard has a **teleport** button which takes him to a random website

At each webpage, Bob decides to

EITHER click a random link on that webpage

OR click his teleport button

If he decides to click a link, he chooses it completely at random.

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text
PageRank
Other tweaks

Video Search and Recommendation

Summary

Let me introduce **Bob**. Bob is bored.

So bored that he spends his days **clicking through webpages**.

Bob's keyboard has a **teleport** button which takes him to a random website

At each webpage, Bob decides to

EITHER click a random link on that webpage

OR click his teleport button

If he decides to click a link, he chooses it completely at random.

Bob is so bored that he keeps doing this forever

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text

PageRank

Other tweaks

Video Search and Recommendation

Summary

Let me introduce **Bob**. Bob is bored.

So bored that he spends his days **clicking through webpages**.

Bob's keyboard has a **teleport** button which takes him to a random website

At each webpage, Bob decides to

EITHER click a random link on that webpage

OR click his teleport button

If he decides to click a link, he chooses it completely at random.

Bob is so bored that he keeps doing this forever

Bob performs a random walk over the internet.

- Bob can't get stuck (because of the catapult)
- Bob will eventually reach every webpage there is
- Bob will spend most time on popular webpages (e.g., `abc.com.au`)

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text
PageRank
Other tweaks

Video Search and Recommendation

Summary

Let me introduce **Bob**. Bob is bored.

So bored that he spends his days **clicking through webpages**.

Bob's keyboard has a **teleport** button which takes him to a random website

At each webpage, Bob decides to

EITHER click a random link on that webpage

OR click his teleport button

If he decides to click a link, he chooses it completely at random.

Bob is so bored that he keeps doing this forever

At some point, you walk past and glance at Bob's screen.
What is the probability that Bob is looking at `abc.com.au`?

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text

PageRank

Other tweaks

Video Search and Recommendation

Summary

Let me introduce **Bob**. Bob is bored.

So bored that he spends his days **clicking through webpages**.

Bob's keyboard has a **teleport** button which takes him to a random website

At each webpage, Bob decides to

EITHER click a random link on that webpage

OR click his teleport button

If he decides to click a link, he chooses it completely at random.

Bob is so bored that he keeps doing this forever

At some point, you walk past and glance at Bob's screen.
What is the probability that Bob is looking at `abc.com.au`?
`unimelb.edu.au`?

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text

PageRank

Other tweaks

Video Search and Recommendation

Summary

Let me introduce **Bob**. Bob is bored.

So bored that he spends his days **clicking through webpages**.

Bob's keyboard has a **teleport** button which takes him to a random website

At each webpage, Bob decides to

EITHER click a random link on that webpage

OR click his teleport button

If he decides to click a link, he chooses it completely at random.

Bob is so bored that he keeps doing this forever

At some point, you walk past and glance at Bob's screen.
What is the probability that Bob is looking at `abc.com.au`?
`unimelb.edu.au`? `houseofcardsespresso.com`?

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text
PageRank
Other tweaks

Video Search and Recommendation

Summary

Let me introduce **Bob**. Bob is bored.

So bored that he spends his days **clicking through webpages**.

Bob's keyboard has a **teleport** button which takes him to a random website

At each webpage, Bob decides to

EITHER click a random link on that webpage

OR click his teleport button

If he decides to click a link, he chooses it completely at random.

Bob is so bored that he keeps doing this forever

At some point, you walk past and observe Bob's current webpage.

What is the probability that you observe
unimelb.edu.au? houseof

← The webpage's **PageRank score**!

Calculating PageRank Scores

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text

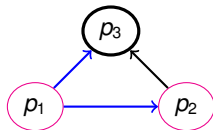
PageRank

Other tweaks

Video Search and
Recommendation

Summary

- $PR(v)$ pagerank score of page v
- I_v set of webpages pointing to v
(e.g., $I_{p_3} = \{p_1, p_2\}$; set of pages pointing to p_3)
- O_v number of outgoing links of page v
(e.g., $O_{p_1} = 2$ outgoing links from p_1)



Calculating PageRank Scores

Web Search

COMP90049
Knowledge
Technologies

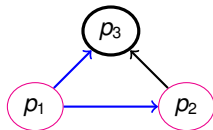
Link Analysis

Anchor Text
PageRank
Other tweaks

Video Search and Recommendation

Summary

- $PR(v)$ pagerank score of page v
- I_v set of webpages pointing to v
(e.g., $I_{p_3} = \{p_1, p_2\}$; set of pages pointing to p_3)
- O_v number of outgoing links of page v
(e.g., $O_{p_1} = 2$ outgoing links from p_1)



$$PR(v) = \sum_{u \in I_v} \frac{PR(u)}{O_u} \quad (\text{no catapult})$$

Calculating PageRank Scores

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text

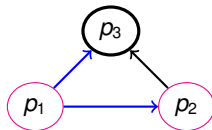
PageRank

Other tweaks

Video Search and
Recommendation

Summary

- $PR(v)$ pagerank score of page v
- I_v set of webpages pointing to v
(e.g., $I_{p_3} = \{p_1, p_2\}$; set of pages pointing to p_3)
- O_v number of outgoing links of page v
(e.g., $O_{p_1} = 2$ outgoing links from p_1)
- N all webpages in our collection (here: $N = 3$)
- λ probability to catapult



$$PR(v) = \frac{\lambda}{N} + (1 - \lambda) \times \sum_{u \in I_v} \frac{PR(u)}{O_u} \quad (\text{with catapult})$$

Calculating PageRank Scores

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text

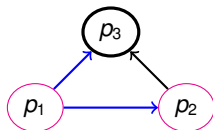
PageRank

Other tweaks

Video Search and
Recommendation

Summary

- $PR(v)$ pagerank score of page v
- I_v set of webpages pointing to v
(e.g., $I_{p_3} = \{p_1, p_2\}$; set of pages pointing to p_3)
- O_v number of outgoing links of page v
(e.g., $O_{p_1} = 2$ outgoing links from p_1)
- N all webpages in our collection (here: $N = 3$)
- λ probability to catapult



$$PR(v) = \frac{\lambda}{N} + (1 - \lambda) \times \sum_{u \in I_v} \frac{PR(u)}{O_u} \quad (\text{with catapult})$$

but we don't know $PR(u)$

Calculating PageRank Scores

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text

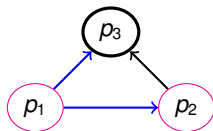
PageRank

Other tweaks

Video Search and
Recommendation

Summary

- $PR(v)$ pagerank score of page v
- I_v set of webpages pointing to v
(e.g., $I_{p_3} = \{p_1, p_2\}$; set of pages pointing to p_3)
- O_v number of outgoing links of page v
(e.g., $O_{p_1} = 2$ outgoing links from p_1)
- N all webpages in our collection (here: $N = 3$)
- λ probability to catapult



$$PR(v) = \frac{\lambda}{N} + (1 - \lambda) \times \sum_{u \in I_v} \frac{PR(u)}{O_u} \quad (\text{with catapult})$$

but we don't know $PR(u)$

Estimate them iteratively:

- each web page has a fixed number of credits
- redistributes credits to pages it links to
- receives credits from pages that link to it
- repeat until stable distribution is reached

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text

PageRank

Other tweaks

Video Search and Recommendation

Summary

Other Tweaks & Heuristics

- Note which pages people visit: count click-throughs.
- Manually alter the behavior of common queries.
- Cache the answers to common queries.
- Index selected phrases.
- Have separate servers for crawling and index construction.
- Accept feeds from dynamic data providers (booksellers, newspapers, ...)
- Integrate diverse data resources, such as maps and videos.

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text
PageRank
Other tweaks

Video Search and Recommendation

Summary

Video Snippets in Google Results

- Demo
- How do they do this? Secret. Hints in a research paper (below).

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text
PageRank
Other tweaks

Video Search and Recommendation

Summary

Video Snippets in Google Results

- Demo
- How do they do this? Secret. Hints in a research paper (below).

Major challenges in **parsing** the crawled information?

- how to parse a video?
 - text → words
 - video → frames? snippets (how long, how many)?
- how to canonicalize a video?
- meta-data (title, author, descriptions, ...)
- how to deal with comments?
- all of the above?
- none of the above?

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text
PageRank
Other tweaks

Video Search and Recommendation

Summary

Video Snippets in Google Results

- Demo
- How do they do this? Secret. Hints in a research paper (below).

Major challenges in **indexing**?

- we can index text by terms ...
what are the 'terms' for videos?
- index as a whole? index snippets (which snippets)?
- index text and video in a single structure? how to link them?
- all of the above?
- none of the above?

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text
PageRank
Other tweaks

Video Search and Recommendation

Summary

Video Snippets in Google Results

- Demo
- How do they do this? Secret. Hints in a research paper (below).

Major challenges in **querying**?

- query is in text form, result is in video (image) form!
how to relate the two?
- boolean querying for videos?
- tfidf for videos? (what are 'terms' for videos, again?)
- exact query-video matching?
approximate query-video matching?
term-based querying?
- all of the above?
- none of the above?

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text
PageRank
Other tweaks

Video Search and Recommendation

Summary

Video Snippets in Google Results

- Demo
- How do they do this? Secret. Hints in a research paper (below).

Maybe, **machine learning** is the key!

- **learn to** align words or queries with videos
- **learn to** segment videos into meaningful snippets
- ...
- 2nd part of Knowledge Technologies (starting tomorrow)

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text
PageRank
Other tweaks

Video Search and Recommendation

Summary

Video Snippets in Google Results

- Demo
- How do they do this? Secret. Hints in a research paper (below).

In your groups:

- Discuss Google video snippets as a knowledge technology
- Play with the Google feature
- In which contexts is it useful?
- What information is required? Where could it come from?
- How can the information be indexed and retrieved?
- For hints, check out this Google research paper:
Malmaud et al., ‘‘What’s Cookin’? Interpreting Cooking
Videos using Text, Speech and Vision.’’ NAACL 2015.
- **Answer the quiz at** <https://pollev.com/kt19>

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text
PageRank
Other tweaks

Video Search and Recommendation

Summary

- **Search** involves crawling, parsing, indexing, querying (and more!)
- **Crawling** is in principle straightforward queuing, but practical issues make it more complex
- **Parsing** involves discarding metadata and hidden information; tokenization; canonicalisation; zoning; and stemming.
- **Inverted indices** describe text collections as lists of the pages with each word, rather than the list of words on each page.
- Inverted indices can be used for **Boolean and ranked querying**.
- On the web, **link and anchor information** can be the dominant evidence of relevance.
- Search goes **beyond word-based matching**: images, videos, phrases and sentences, ads, user behavior...

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text
PageRank
Other tweaks

Video Search and Recommendation

Summary

Malmaud et al., (2015) “What’s Cookin’? Interpreting Cooking Videos using Text, Speech and Vision.” Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

W. Bruce Croft and Donald Metzler and Trevor Strohman (2015), “Search Engines: Information Retrieval in Practice”. Online version. Pearson Education, Inc. Chapter 4.

Pagerank algorithm

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text
PageRank
Other tweaks

Video Search and Recommendation

Summary

Input: D = document set

Output: Π_T = set of pagerank scores for each document $d_i \in D$

```

1: for all  $d_i \in D$  do                                ▷ Initialise the starting probabilities
2:    $\pi(d_{(i,0)}) \leftarrow \frac{1}{N}$                         ▷  $N$  is the total number of documents
3: end for
4: for  $t = 1..T$  do                                    ▷ Repeat over  $T$  iterations
5:   for all  $d_i \in D$  do                                ▷ Initialise the document probabilities
6:      $\pi(d_{(i,t)}) \leftarrow 0$ 
7:   end for
8:   for all  $d_i \in D$  do
9:     if  $\exists d_j : d_i \mapsto d_j$  then
10:      for all  $d_j \in D$  do                                ▷ EITHER teleport randomly
11:         $\pi(d_{(j,t)}) \leftarrow \pi(d_{(j,t)}) + \lambda \times \pi(d_{(i,t-1)}) \times \frac{1}{N}$ 
12:      end for
13:      for all  $d_j$  where  $d_i \mapsto d_j$  do
14:         $\pi(d_{(j,t)}) \leftarrow \pi(d_{(j,t)}) + (1 - \lambda) \times \pi(d_{(i,t-1)}) \times \frac{1}{m}$   ▷ OR follow an outlink (one of  $m$ )
15:      end for
16:    else
17:      for all  $d_j \in D$  do                                ▷ teleport to a random document
18:         $\pi(d_{(j,t)}) \leftarrow \pi(d_{(j,t)}) + \pi(d_{(i,t-1)}) \times \frac{1}{N}$ 
19:      end for
20:    end if
21:  end for
22: end for
23: end for

```

Web Search

COMP90049
Knowledge
Technologies

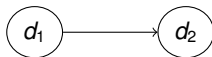
Link Analysis

Anchor Text
PageRank
Other tweaks

Video Search and Recommendation

Summary

A (tiny) web



$$PR(v) = \frac{\lambda}{N} + (1 - \lambda) \times \sum_{u \in I_v} \frac{PR(u)}{O_u}$$

$\lambda = 0.5$		
t	$\pi(d_{(1,t)})$	$\pi(d_{(2,t)})$
0	0.5	0.5

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text
PageRank
Other tweaks

Video Search and Recommendation

Summary

A (tiny) web



$$PR(v) = \frac{\lambda}{N} + (1 - \lambda) \times \sum_{u \in I_v} \frac{PR(u)}{O_u}$$

$$\lambda = 0.5$$

t	$\pi(d_{(1,t)})$	$\pi(d_{(2,t)})$
0	0.5	0.5
1	$0.5 \times 0.2 \times 0.5 + 0.5 \times 0.5 = 0.3$	$0.5 \times 0.2 \times 0.5 + 0.5 \times 0.8 + 0.5 \times 0.5 = 0.7$

Web Search

COMP90049
Knowledge
Technologies

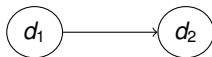
Link Analysis

Anchor Text
PageRank
Other tweaks

Video Search and Recommendation

Summary

A (tiny) web



$$PR(v) = \frac{\lambda}{N} + (1 - \lambda) \times \sum_{u \in I_v} \frac{PR(u)}{O_u}$$

$$\lambda = 0.5$$

t	$\pi(d_{(1,t)})$	$\pi(d_{(2,t)})$
0	0.5	0.5
1	$0.5 \times 0.2 \times 0.5 + 0.5 \times 0.5 = 0.3$	$0.5 \times 0.2 \times 0.5 + 0.5 \times 0.8 + 0.5 \times 0.5 = 0.7$
2	$0.3 \times 0.2 \times 0.5 + 0.7 \times 0.5 = 0.38$	$0.3 \times 0.2 \times 0.5 + 0.3 \times 0.8 + 0.7 \times 0.5 = 0.62$

Web Search

COMP90049
Knowledge
Technologies

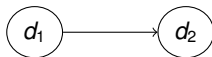
Link Analysis

Anchor Text
PageRank
Other tweaks

Video Search and Recommendation

Summary

A (tiny) web



$$PR(v) = \frac{\lambda}{N} + (1 - \lambda) \times \sum_{u \in I_v} \frac{PR(u)}{O_u}$$

$$\lambda = 0.5$$

t	$\pi(d_{(1,t)})$	$\pi(d_{(2,t)})$
0	0.5	0.5
1	$0.5 \times 0.2 \times 0.5 + 0.5 \times 0.5 = 0.3$	$0.5 \times 0.2 \times 0.5 + 0.5 \times 0.8 + 0.5 \times 0.5 = 0.7$
2	$0.3 \times 0.2 \times 0.5 + 0.7 \times 0.5 = 0.38$	$0.3 \times 0.2 \times 0.5 + 0.3 \times 0.8 + 0.7 \times 0.5 = 0.62$
3	$0.38 \times 0.2 \times 0.5 + 0.62 \times 0.5 = 0.348$	$0.38 \times 0.2 \times 0.5 + 0.38 \times 0.8 + 0.62 \times 0.5 = 0.652$

Web Search

COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text
PageRank
Other tweaks

Video Search and Recommendation

Summary

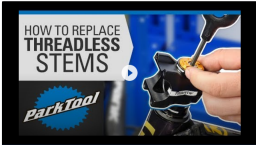
https://www.google.com/search?client=ubuntu&hl=hs&channel=fs&ei=7u8Kxv5Ldey9QOK4i48Q6q=properly+mount+stem&q=properly+mount+stem&gs_l=psy-ab..3.33160.2382.4712..5071...0.4.0.185.293j

Melbourne Forecast COMP90049 Calendar Email Files - OneDrive Thems polleverywhere yoga Get Ready to Work - H... electronics vkods Tim's Calendar Amazon Research Aw... MLS Tax Clinic - Client ...

Google properly mount stem

AI Images Shopping Videos Maps More Settings Tools

About 15,100,000 results (0.58 seconds)



Suggested clip 02 seconds

How to Replace a Bicycle Stem - Threadless - YouTube
<https://www.youtube.com/watch>

About Featured Snippets Feedback

People also ask

- What is an ahead stem? ▾
- What is a threadless stem? ▾
- How do you install a quill stem? ▾
- How are quill stems measured? ▾

Feedback

Web Search

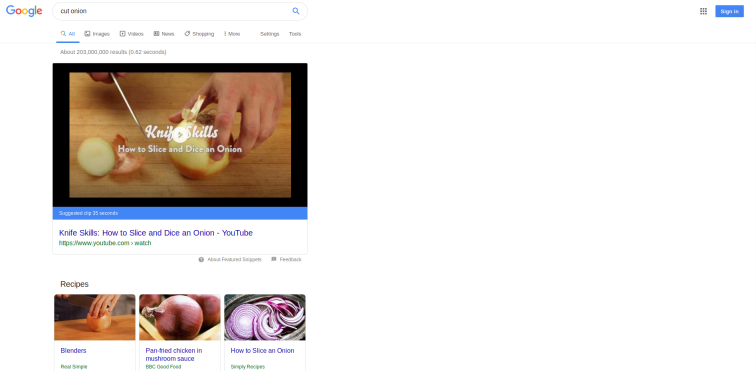
COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text
PageRank
Other tweaks

Video Search and Recommendation

Summary



The screenshot shows a Google search for "cut onion". The search bar at the top contains the text "cut onion" and a magnifying glass icon. Below the search bar, there are tabs for "All", "Images", "Videos", "News", "Shopping", "More", "Settings", and "Tools". The "All" tab is selected. Below the tabs, it says "About 103,000,000 results (0.82 seconds)". The main result is a video thumbnail from YouTube titled "Knife Skills: How to Slice and Dice an Onion". The thumbnail shows a person's hands slicing an onion on a wooden cutting board. Below the thumbnail, there is a blue bar that says "Suggested clip: 31 seconds". Below that, the text reads "Knife Skills: How to Slice and Dice an Onion - YouTube" and "https://www.youtube.com/watch". At the bottom of the video result, there are links for "About Featured Snippets" and "Feedback". Below the video result, there is a section titled "Recipes" with three small images. The first image shows a hand holding a knife over a cutting board, with the caption "Blenders" and "Real Simple". The second image shows a whole onion, with the caption "Pan-fried chicken in mushroom sauce" and "BBC Good Food". The third image shows sliced onions, with the caption "How to Slice an Onion" and "Simply Recipes".

Web Search

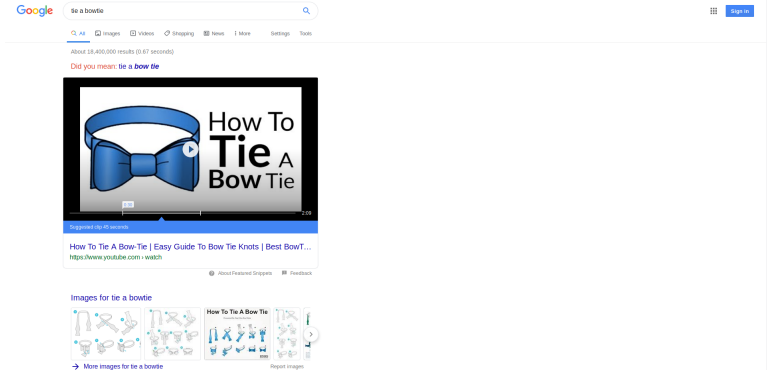
COMP90049
Knowledge
Technologies

Link Analysis

Anchor Text
PageRank
Other tweaks

Video Search and Recommendation

Summary



Google tie a bowtie

About 18,400,000 results (0.67 seconds)

Did you mean: tie a **bow tie**

How To Tie A Bow Tie

Suggested clip: 45 seconds

How To Tie A Bow-Tie | Easy Guide To Bow Tie Knots | Best BowT...
<https://www.youtube.com/watch>

Images for tie a bowtie

More images for tie a bowtie