How to build a search engine for the Web?

Louis Jachiet

What is a search engine?

What is a search engine ?

What is a search engine ?

Do you know examples of search engines?

Several flavors of search engines

- For the Web

  e.g. Bing, Duckduckgo, Google, Yandex

- For a Website

  e.g. on Telecom website or on Youtube

- For a collection of documents

  e.g. Google Scholar

- For books

  e.g. a library

Commonalities to search engines

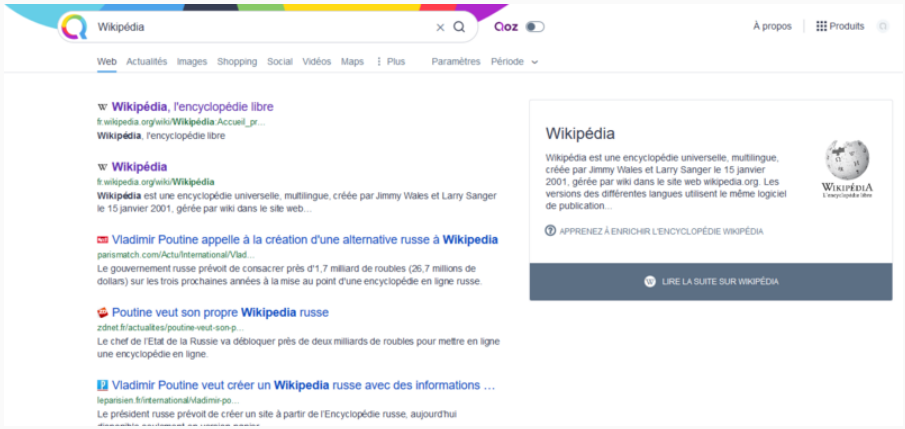
- A set of "entities" (often called documents) to be searched

  Webpages, Videos, Books, Scientific Articles, etc.

- A query language

  Contains X and Y, Natural language, Author/Title/Type/ISBN, etc.

- The output is a list of selected documents with a ranking



Reminder on HTML

- Hypertext documents link to other documents

Reminder on HTML

- Hypertext documents link to other documents
- together these links forms a directed graph: the Web!

Reminder on HTML

- Hypertext documents link to other documents
- together these links forms a directed graph: the Web!
- Web browsing also uses other forms of links (forms, JS, etc.)

Specificity of a Web search engine

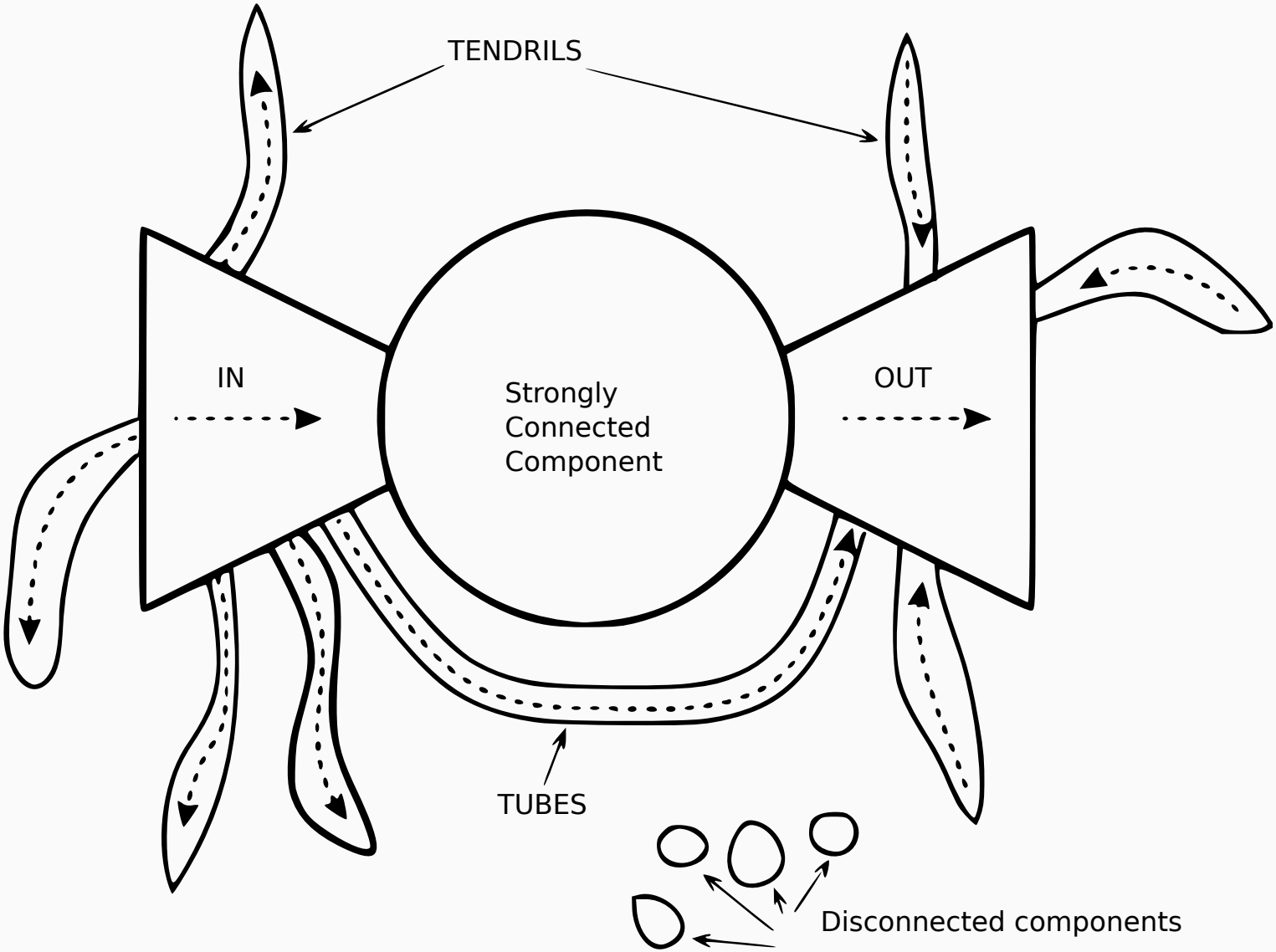
- The documents are mostly hypertext and form a graph
- Search engines are not supposed to be normative on documents
- This graph can be manipulated by anyone
- The Web is too large (6×10^{10} webpages) and too rapidly evolving to be manually curated
- Query languages should stay simple

Comparison of search engine types

Name	Shape	Size	Metadata information	Evolution	Personalized
Web	Graph	10^{10}	Low	Fast	Somewhat
Website	Items	10^4	Low	Slow	No
Youtube	Videos	5×10^9	High	No	Yes
G. Scholar	Graph	75×10^6	High	No	No
Books	Items	10^8	High	No	Maybe

Retrieve the Web

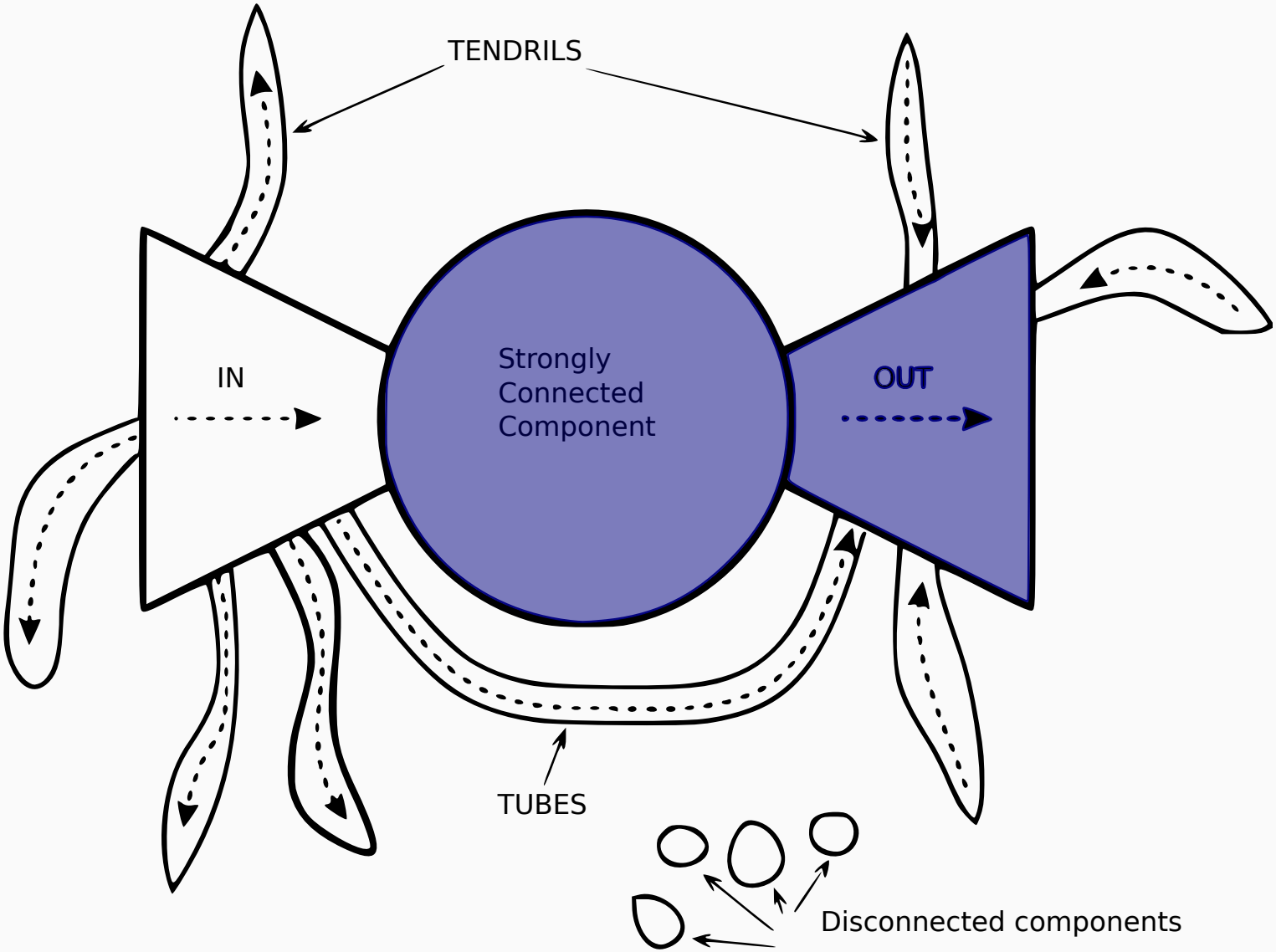
Bowtie Structure of the Web



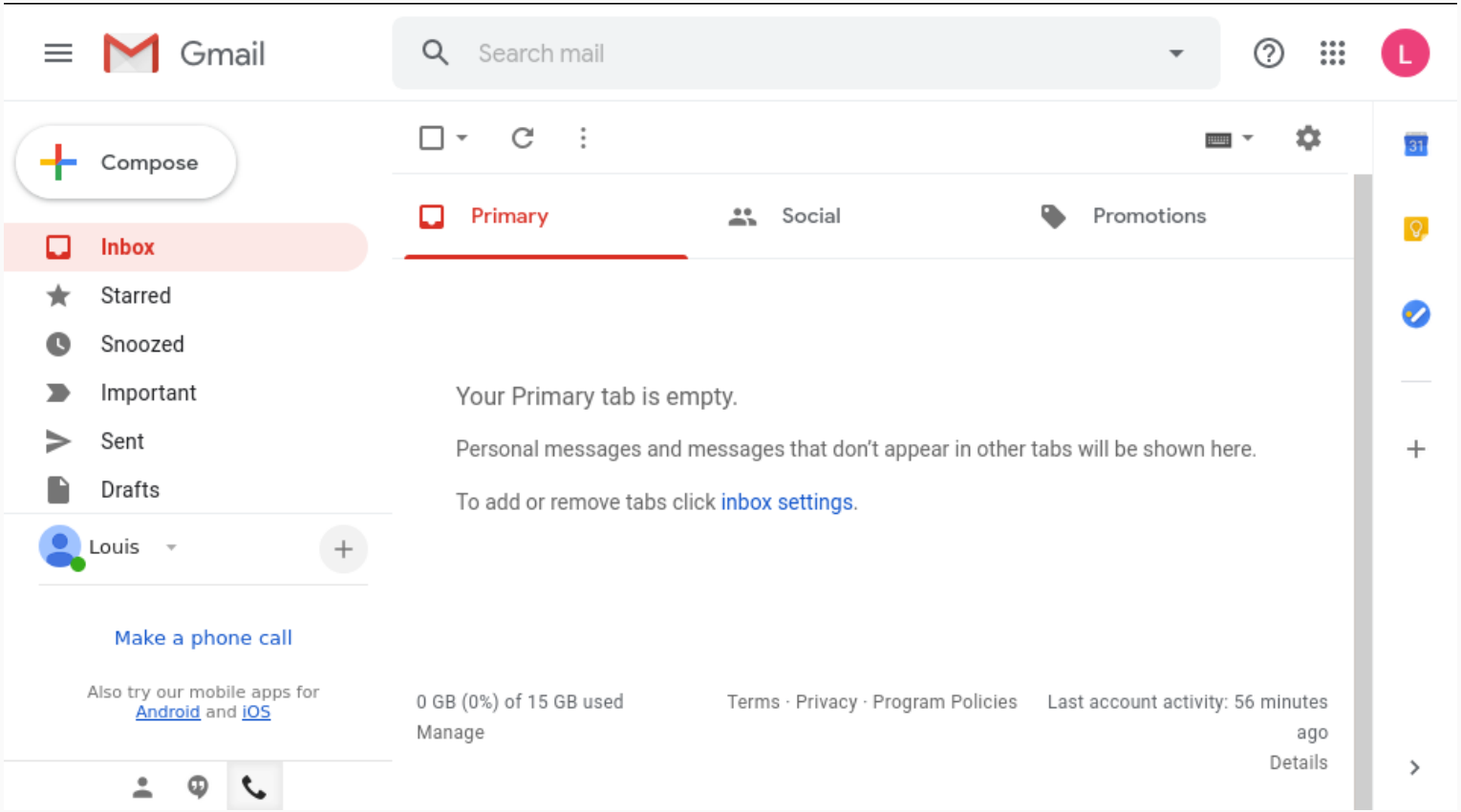
Exploring the Web

```
todo = [start]
webpages = set(todo)
while len(todo):
    page = todo.pop()
    for nxt in neighbor(page):
        if nxt not in webpages:
            todo.append(nxt)
            webpages.insert(nxt)
```

Explorable part of the Web



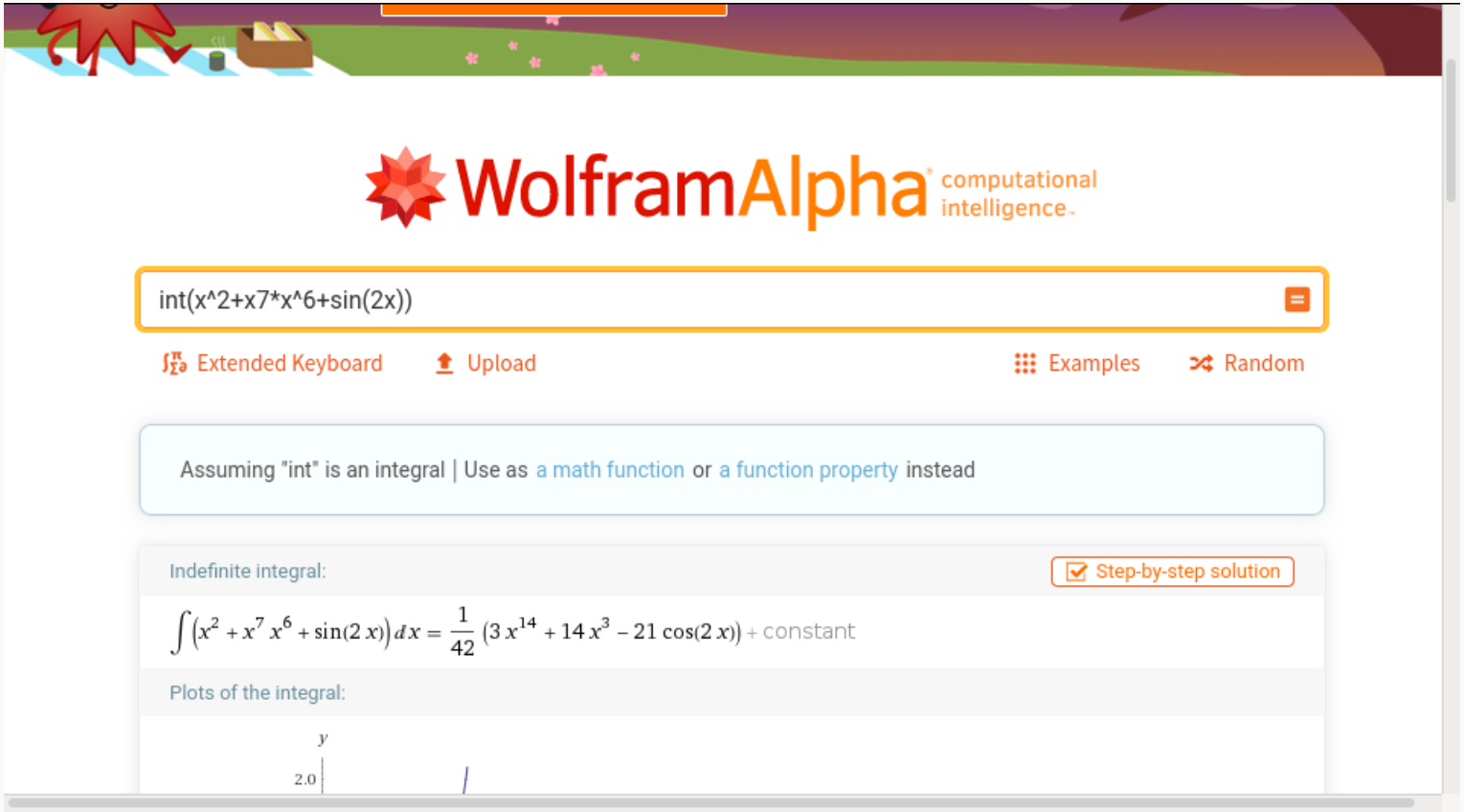
The Deep Web, the unexplorable part of the Web



The Deep Web, the unexplorable part of the Web

The Deep Web, the unexplorable part of the Web

The Deep Web, the unexplorable part of the Web



The Deep Web, the unexplorable part of the Web

Legal & Technical Aspects

Technical aspects

- Risk of DDOS-ing a website

  Too frequent access or resource heavy webpages

- Going through computer generated webpages

  July 3452 Calendar...

- Bots browsing an app might generate a lot of noise

Legal aspects

- Everything on the internet is at risk of being copied

  Beware of what you put on the internet!

- A link on Google $\neq$ the webpage being stored by the NSA...

  Right to forget

Robots Exclusion Protocol (`robots.txt`)

```
Crawl-delay: 4

User-agent: *
Allow: *

User-agent: *
Disallow: /tmp/

User-agent: wget
Allow: /tmp/

User-agent: curl
Allow: /tmp/

User-agent: ia_archiver
Disallow: /
```

Other Robots Exclusion Methods

In the HTML

```
<meta name="robots" content="noindex">
<!-- prevent this page to be indexed -->
```

```
<meta name="robots" content="nofollow">
<!-- do not follow links in the webpage -->
```

In the HTTP conversation

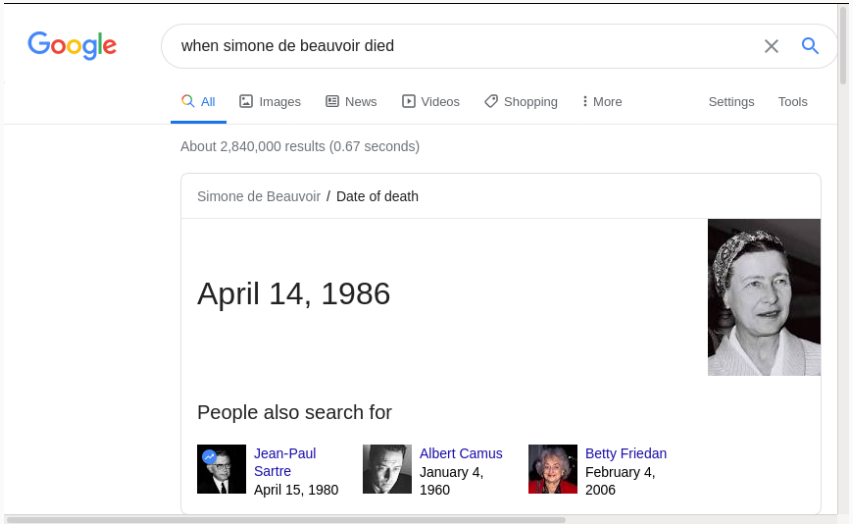
```
HTTP/1.1 200 OK
(...)
X-Robots-Tag: noindex
```

Beyond simple crawling

Extracting semantic information from crawling

- from online phonebooks
- from shopping websites
- from restaurants
- ...

Using it for semantic answering



Information Retrieval

We have collected billions of webpages...

... how to get the relevant ones?

Query language

Which query language for a Web search engine?

- Simple

but with advanced features

- very Selective

The Web has billion of webpages

Easy solution: bag of words

- We compute solutions for each word
- Partial solutions are merged
- We remove stop words

This allows for an English-like querying

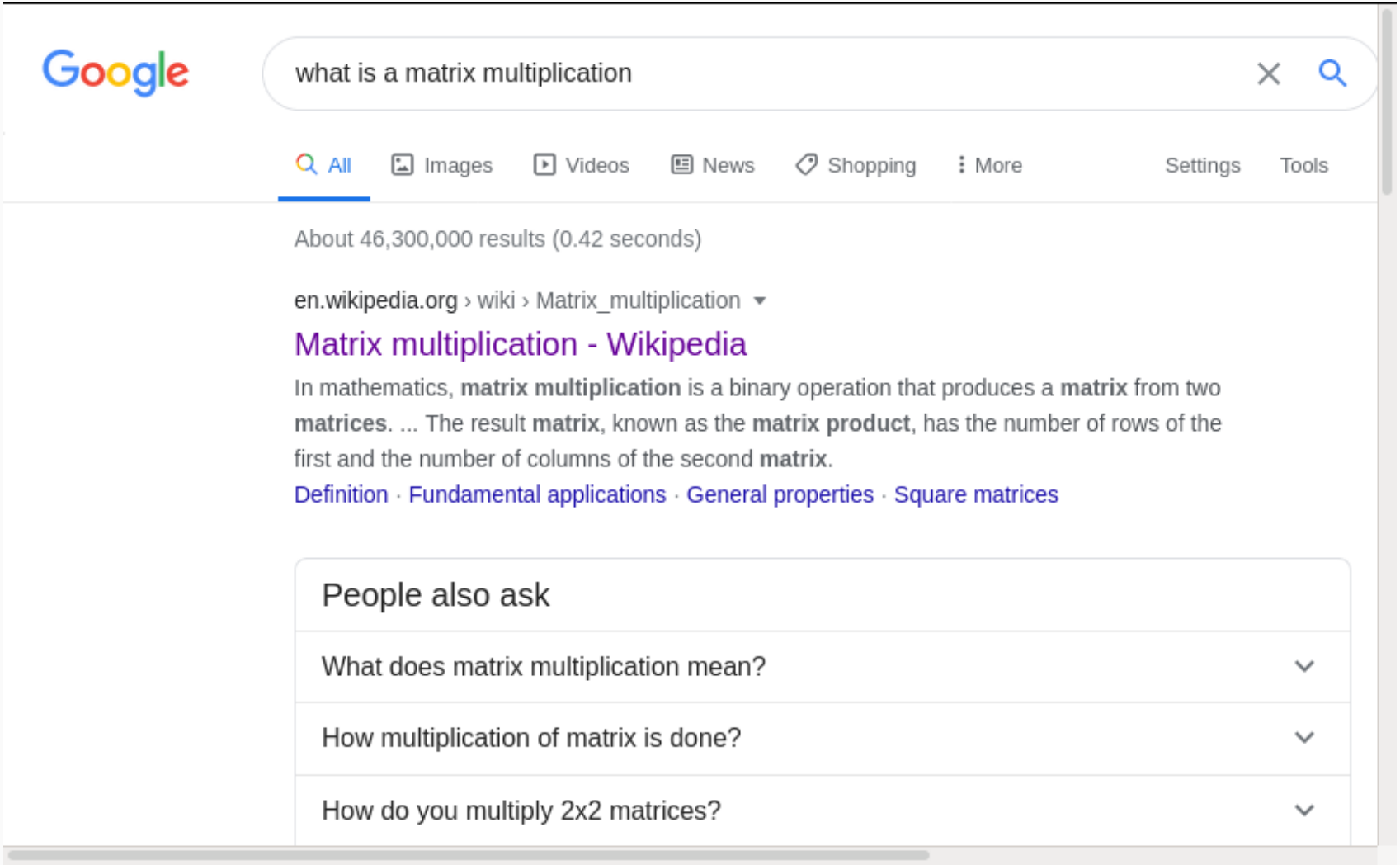
Example

What is a matrix multiplication?

Example

What is a matrix multiplication?

What is a matrix multiplication?



Example

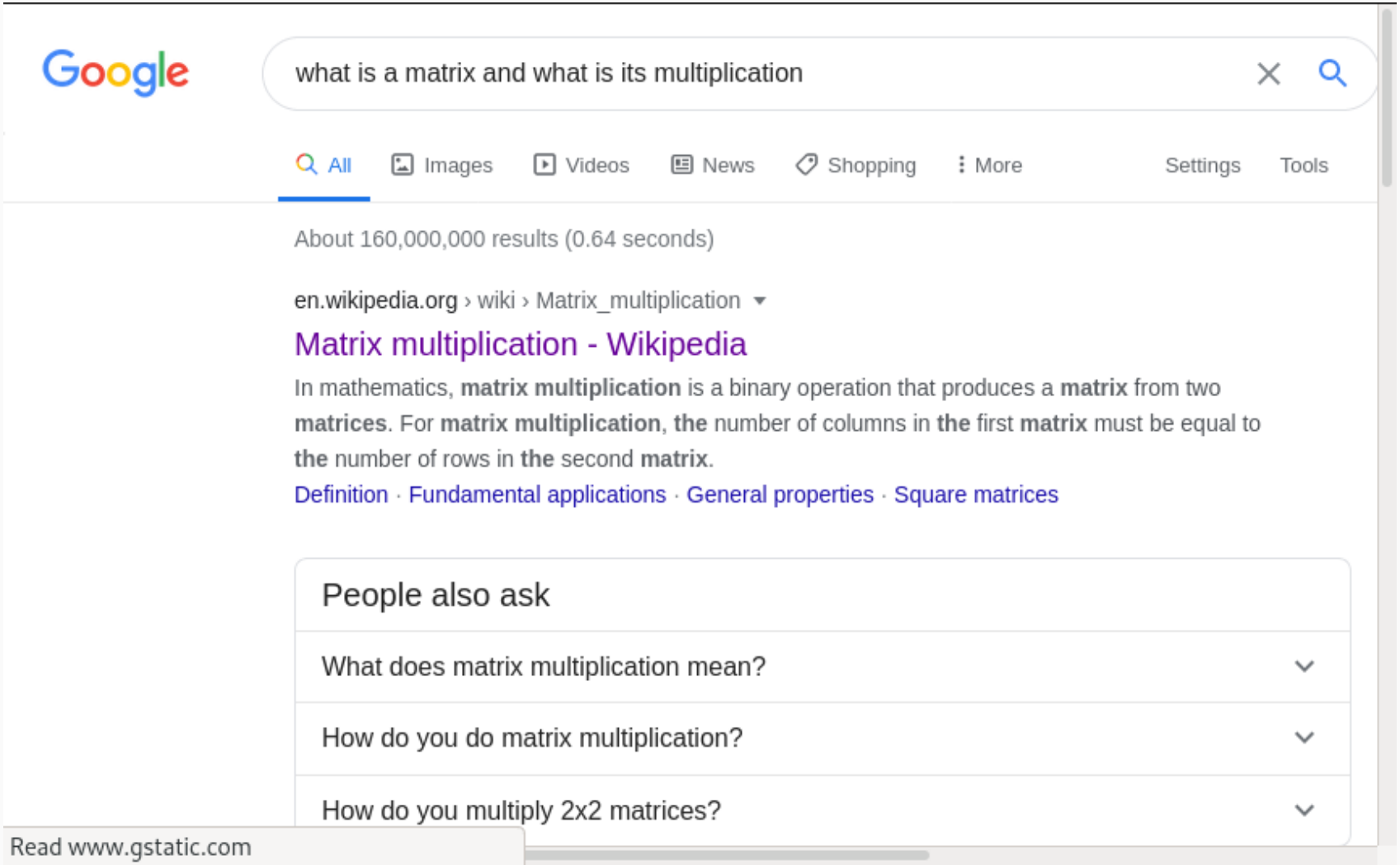
What is a matrix and what is its multiplication?

Example

What is a matrix and what is its multiplication?

Example

What is a matrix and what is its multiplication?



Example

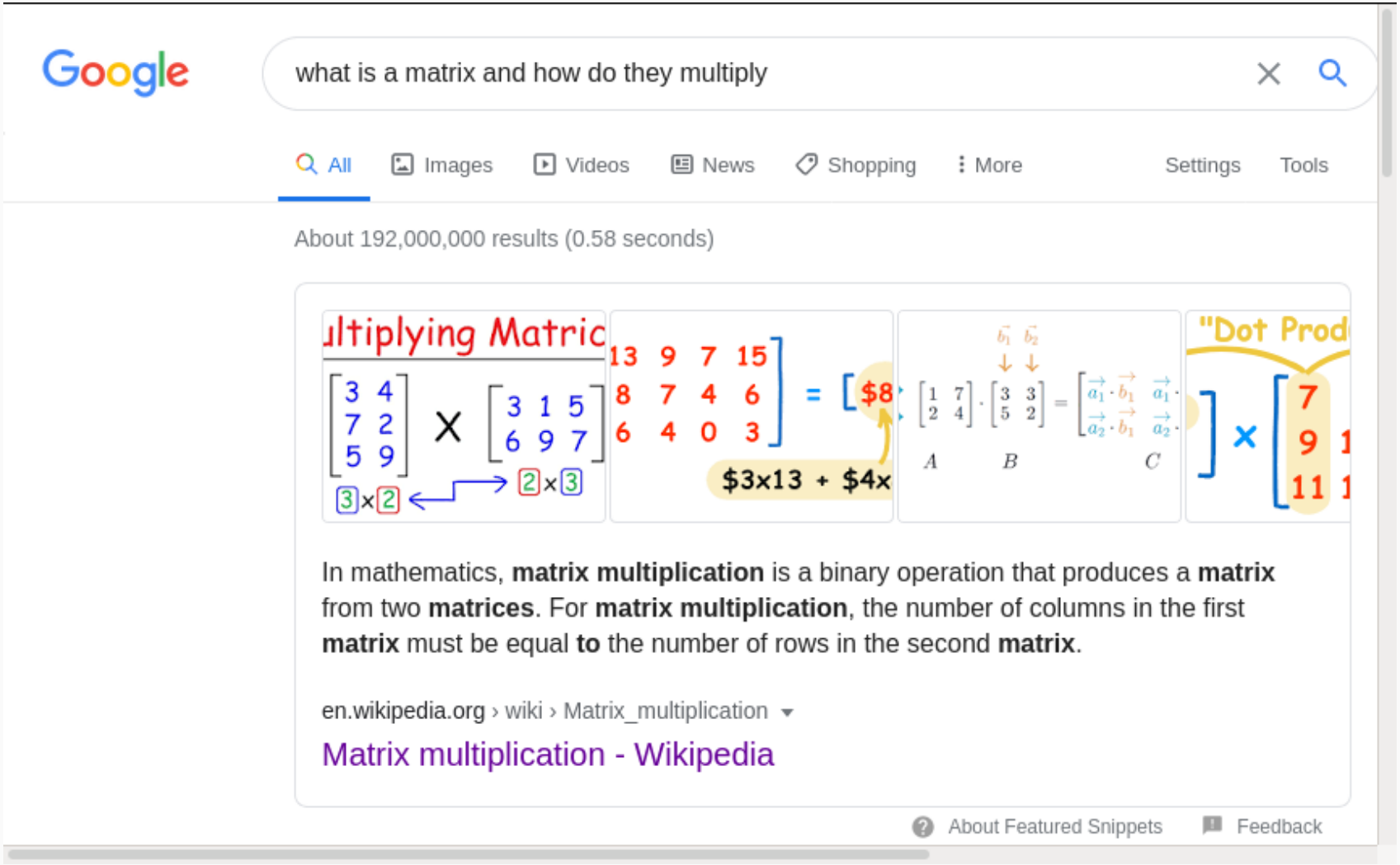
What is a matrix and how do they multiply?

Example

What is a matrix and how do they multiply?

Example

What is a matrix and how do they multiply?



Some Terminology for the Bag-of-word Model

Term-document Matrix

A Matrix $\mathcal{M}$ where $\mathcal{M}_{w,d}1$ corresponds to the frequency at which w appears in document d.

Inverted Index

A list $(f_1, d_1), \ldots$ for each word w stating that w appears with frequency f_1 in d_1.

Full Inverted Index

A list $(f_1, o_1), \ldots$ for each word w stating that w appears at positions o_1 in d_1.

Multiple Models for bag of words

The Boolean Model

- Each word of the query appears or not in the document
- A bag query can be ranked by the number of matches

The Vector Space Model

- A document d is associated with a vector $\vec{v}(d)$
- $\vec{v}(d)_i$ corresponds to the weight of term i in d
- Ranking can be done using classical metrics (such as norm-2 or norm-1)

Multiple Models for bag of words

The Probabilistic Model

- For each word w and each document d we compute $P(d, w)$, the probability that d talks about w
- With independence assumption the probability of a document d being a match is $\prod_w P(d, w)$

Latent semantic analysis

- A low-rank decomposition of the term-document-matrix
- Distance between document d and bag b is dot-product.

Term Frequency - Inverse Document Frequency (tf-idf)

Given a set of documents $\mathcal{D}$, tf-idf$_{\mathcal{D}}(w, d)$ computes for the importance of w within document d with respect to documents in $\mathcal{D}$.

Formally:

$$\text{tf-idf}_{\mathcal{D}}(w, d) = tf(w, d) \times idf(w, \mathcal{D})$$

... now *tf* comes in several flavors and so does *idf*!

Term Frequency

With $f_{w,d}$ the number of occurrences of w in d:

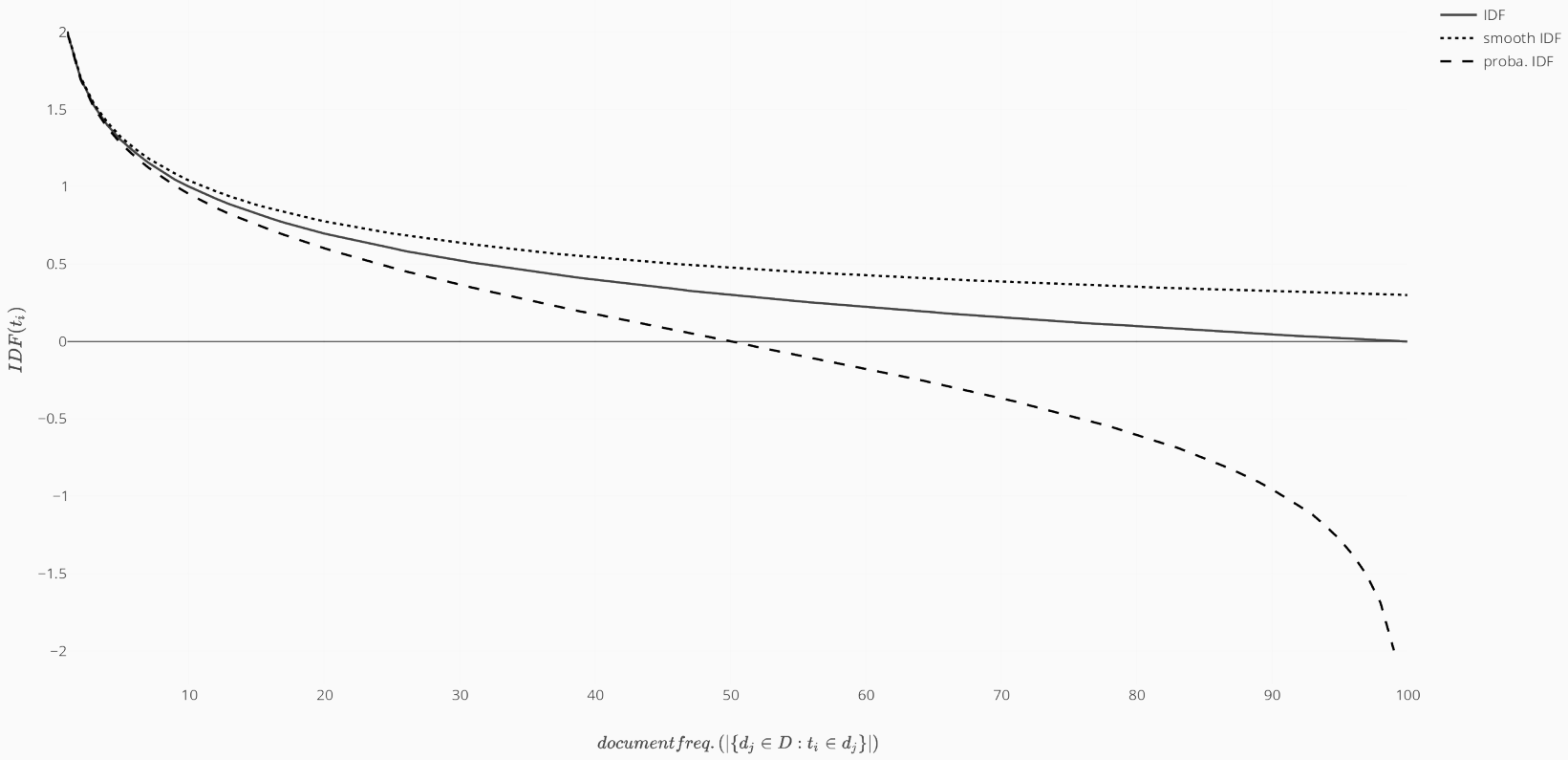
Weighting scheme	tf weight
binary	0 or 1
count	$f_{w,d}$
term frequency	$\frac{f_{w,d}}{\sum_{w'} f_{w',d}}$
log normalization	$log(1 + f_{w,d})$
double normalization	$\alpha + (1 - \alpha) \times \frac{f_{w,d}}{\sum_{w'} f_{w',d}}$

Inverse Document Frequency

With n_w the number of documents with word w:

Weighting scheme	idf weight
unary	1
inverse document frequency	$log\left(\frac{N}{n_w}\right)$
smoothed idf	$log\left(\frac{N}{1+n_w}\right)+1$
max-normalized	$log\left(\frac{\max\ n_{w'}}{1+n_w}\right)$
probabilistic idf	$log\left(\frac{N-n_w}{n_w}\right)$

Inverse Document Frequency



Offline computations for tf-idf

Term-document Matrix

`flatMap` $d_i \rightarrow (i, d_i.split())$

Inverted Index

`groupByKey`

Inverse Document Frequency

`countByKey`

Storage

Store the offline computation into an efficient key-value store. The vocabulary can often be stored in main memory (Heaps' Law: $O(n^{\beta})$ with $\beta \sim 0.5$).

Online computations for tf-idf

Given $w_1 \dots w_k$:

- look up $idf(w_1), \dots, idf(w_k)$
- compute $tf(w_i, d)$ for all d and all i using inverted index
- output the list of documents ordered by $\sum_i tf(w_i, d) \times idf(w_i)$ (norm-1)

Other models than tf-idf

Historically important: Okapi BM-25

$$\text{score}(d, W) = \sum_{i=1}^{n} \text{IDF}(w_i) \cdot \frac{f(w_i, d) \cdot (k_1 + 1)}{f(w_i, d) + k_1 \cdot \left(1 - b + b \cdot \frac{len(d)}{\text{avglen}}\right)}$$

with $1.2 \leq k_1 \leq 2$ and $b \sim 0.75$.

Persin's Algorithm

Ranking-in-the-vector-model(query terms t)

(1) Create P as C-candidate similarities initialized to $(P_d, P_w) = (0, 0)$
(2) Sort the query terms t by decreasing weight
(3) $c \leftarrow 1$.
(4) **for** each sorted term t in the query **do** {
(5) Compute the value of the threshold t_{add}.
(6) Retrieve the inverted list for t, L_t.
(7) **for** each document d in L_t **do** {
(8) **if** $w_{d,t} < t_{add}$ **then break**
(9) $psim \leftarrow w_{d,t} \times w_{q,t}/W_d$.
(10) **if** $d \in P_d(i)$ **then** $P_w(i) \leftarrow P_w(i) + psim$
(11) **elif** $psim > min_j(P_w(j))$ **then** $n \leftarrow min_j(P_w(j))$
(12) **elif** $c \leq C$ **then** {
(13) $n \leftarrow c$
(14) $c \leftarrow c + 1$
}
(15) **if** $n \leq C$ **then** $P(n) \leftarrow (d, psim)$
} }
(16) **return** the top-k documents according to P_w

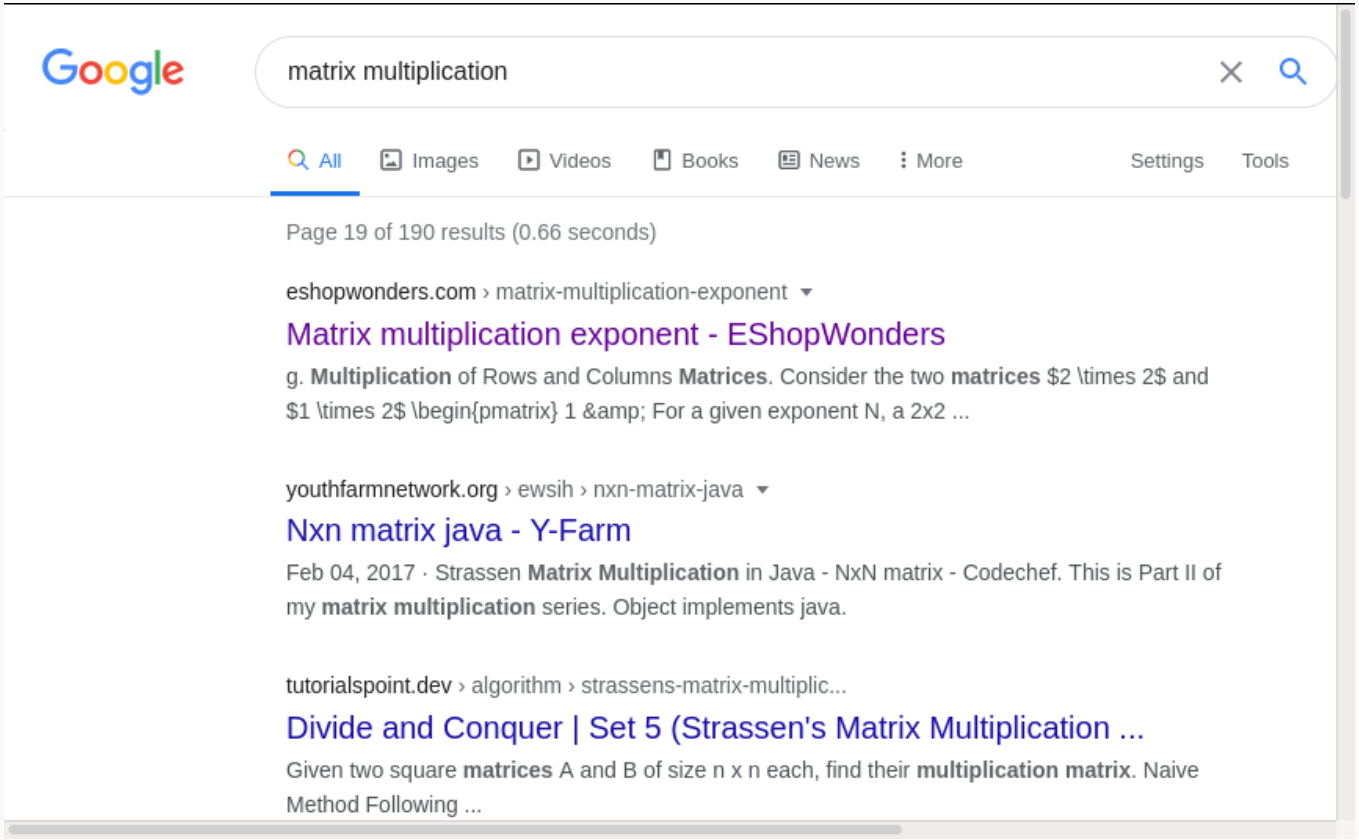
Scan of Modern Information Retrieval

How to asses Website Quality?

Anyone can publish content on the Web...

... how to find quality content?

Problem


Google
matrix multiplication
All
Images
Videos
Books
News
More
Settings
Tools
Page 19 of 190 results (0.66 seconds)
eshopwonders.com › matrix-multiplication-exponent
Matrix multiplication exponent - EShopWonders
g. Multiplication of Rows and Columns Matrices. Consider the two matrices 2×2 and 1×2 \begin{pmatrix} 1 & For a given exponent N, a 2x2 ...
youthfarmnetwork.org › ewsih › nxn-matrix-java
Nxn matrix java - Y-Farm
Feb 04, 2017 · Strassen Matrix Multiplication in Java - NxN matrix - Codechef. This is Part II of my matrix multiplication series. Object implements java.
tutorialspoint.dev › algorithm › strassens-matrix-multiplic...
Divide and Conquer | Set 5 (Strassen's Matrix Multiplication ...
Given two square matrices A and B of size n x n each, find their multiplication matrix. Naive Method Following ...

Early solutions

- Pre-defined set of authoritative content websites?
- Presence in online directories?
- Count incoming links?
- Count incoming links from authorities?

HITS algorithm

General Idea

The idea originate from the time the web had a lot pages serving as directories (hubs here) and a lot of pages providing information (authorities here).

In this model a good reference on a topic should list top quality websites providing information.

HITS algorithm

General Idea

The idea originate from the time the web had a lot pages serving as directories (hubs here) and a lot of pages providing information (authorities here).

In this model a good reference on a topic should list top quality websites providing information.

Algorithm

- Retrieve (e.g. with tf-idf) a set of webpages called "root set of authorities" for the query
- Iterate
  - Determine a set of good hubs for those authorities
  - Determine a set of good authorities for those hubs

Underlying Principles for PageRank

- hypertext documents tend to point to documents that are more important (or serious/credible/trustworthy) than themselves
- the more important a document is, the more important are the documents is points to

Underlying Principles for PageRank

- hypertext documents tend to point to documents that are more important (or serious/credible/trustworthy) than themselves
- the more important a document is, the more important are the documents is points to

General Idea

Compute the probability to arrive at some webpage after following a random walk that teleport randomly with probability α and follow one of the link of the page with probability $(1 - \alpha)$.

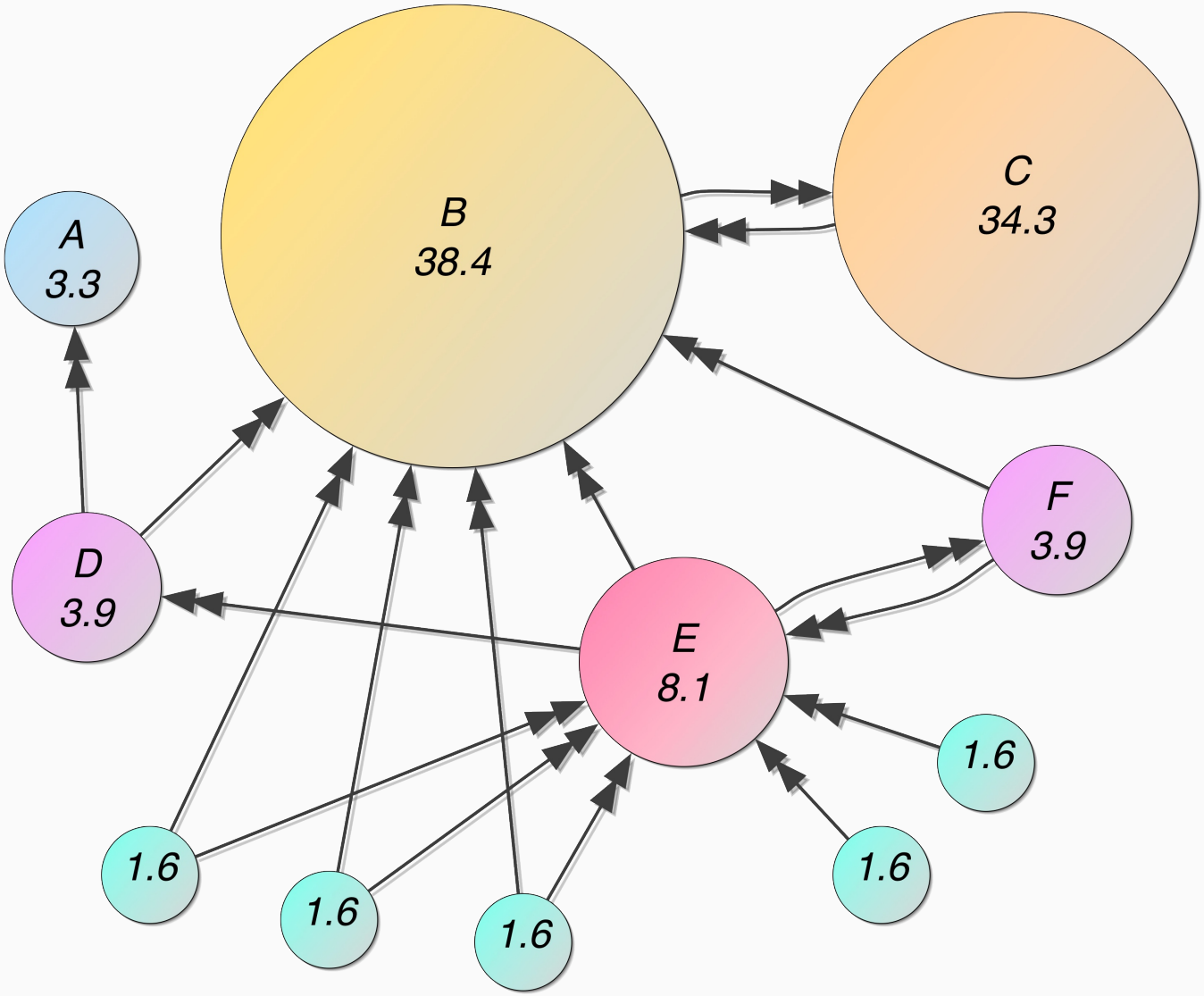
Page Rank

Algorithm

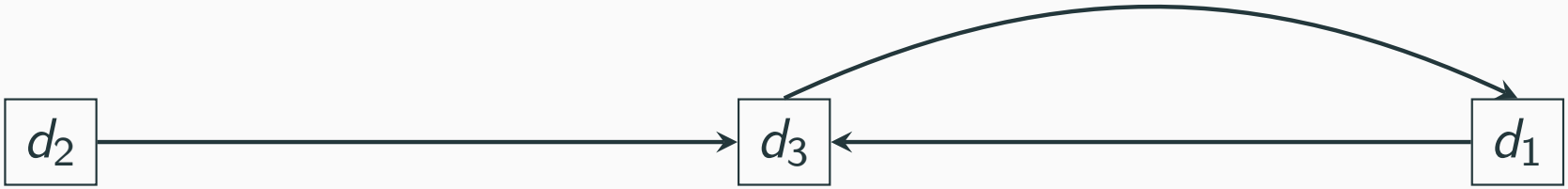
- Compute the matrix $\mathcal{M}$ corresponding to the Markov chain
- Compute $\mathcal{M}^k \times < 1 \dots 1 >^T$ for k big enoug

In practice this algorithm converges exponentially fast in $1 - \alpha$ (a small number of iterations is enough).

PageRank Example



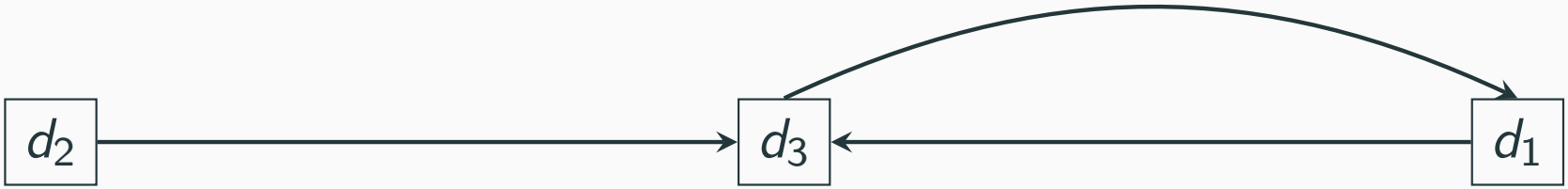
Why α? How to set it?



Exercise

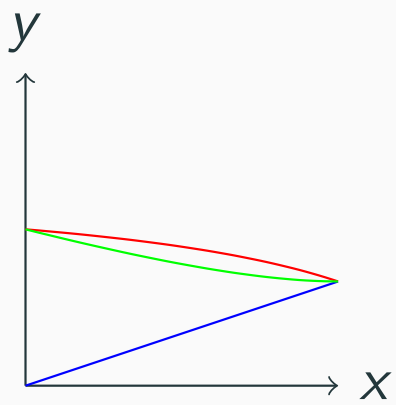
Compute PageRank of d_1, d_2, d_3 for $\alpha = 0, 1, \frac{1}{2}$?

Why α? How to set it?



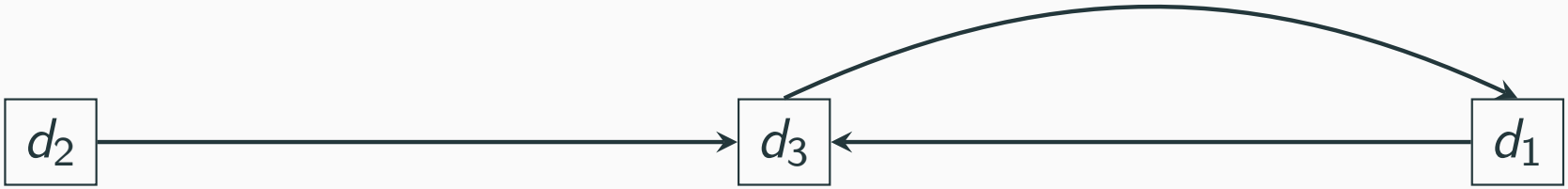
Exercise

Compute PageRank of d_1, d_2, d_3 for $\alpha = 0, 1, \frac{1}{2}$?



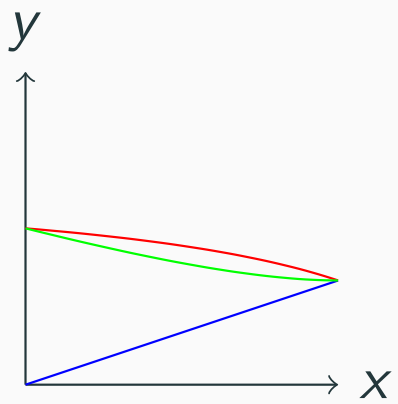
α	d_1	d_2	d_3

Why α? How to set it?



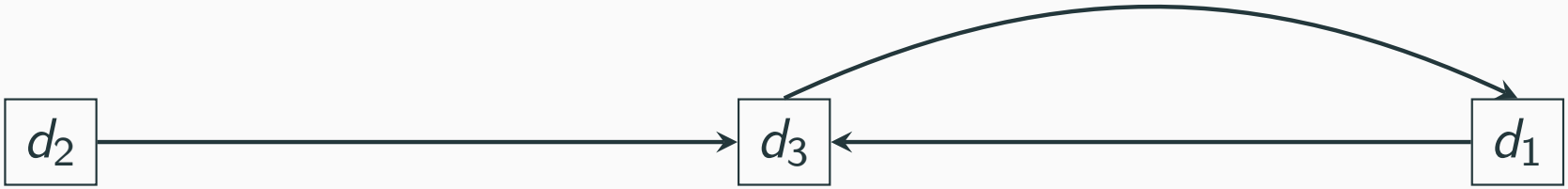
Exercise

Compute PageRank of d_1, d_2, d_3 for $\alpha = 0, 1, \frac{1}{2}$?



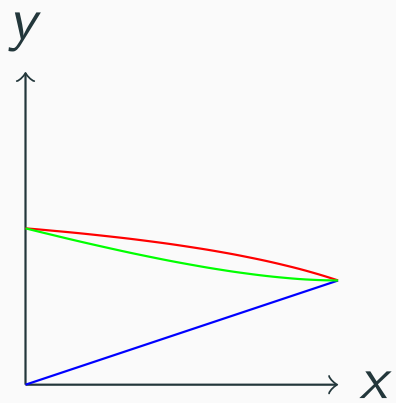
α	d_1	d_2	d_3
0	1/2	1/2	0

Why α? How to set it?



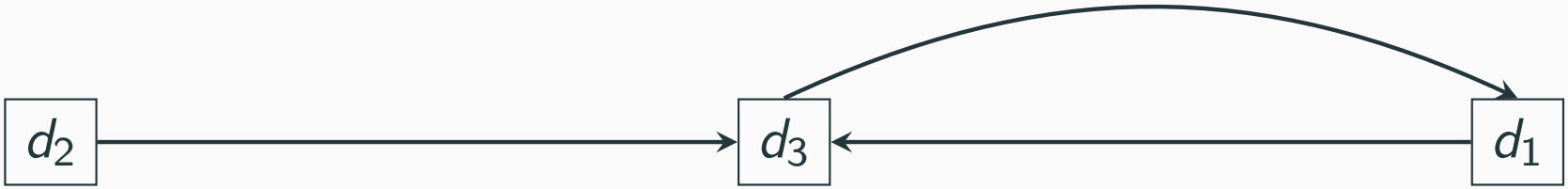
Exercise

Compute PageRank of d_1, d_2, d_3 for $\alpha = 0, 1, \frac{1}{2}$?



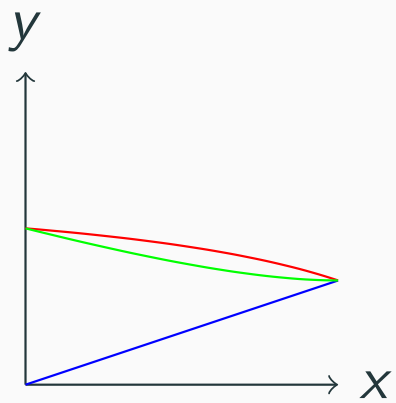
α	d_1	d_2	d_3
0	1/2	1/2	0
1	1/3	1/3	1/3

Why α? How to set it?



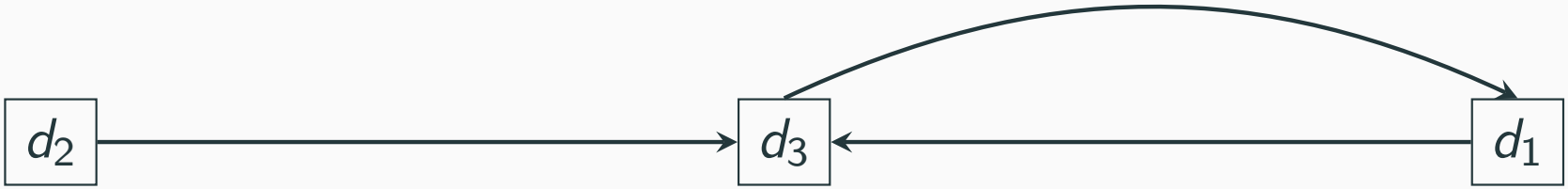
Exercise

Compute PageRank of d_1, d_2, d_3 for $\alpha = 0, 1, \frac{1}{2}$?



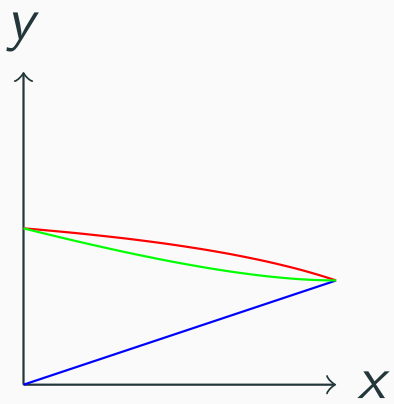
α	d_1	d_2	d_3
0	1/2	1/2	0
1	1/3	1/3	1/3
1/2	4/9	7/18	1/6

Why α? How to set it?



Exercise

Compute PageRank of d_1, d_2, d_3 for $\alpha = 0, 1, \frac{1}{2}$?



α	d_1	d_2	d_3
0	1/2	1/2	0
1	1/3	1/3	1/3
1/2	4/9	7/18	1/6
α	$\frac{3-2\alpha}{6-3\alpha}$	...	$\alpha/3$

PageRank Improvements for assessing page quality

- Take domain into consideration and not just pages
- Some links are better than others
- Take good design into consideration
  - mobile friendliness
  - user friendliness
  - lots of ads
  - https readiness
  - domain age
  - reverse page rank
  - grammar
  - page age
  - sitemap

Spamdexing: Attacks and Counter attacks

The Web is relatively open ...

and people have huge incentives manipulate pageranks

Spamdexing: Attacks and Counter attacks

Attacks

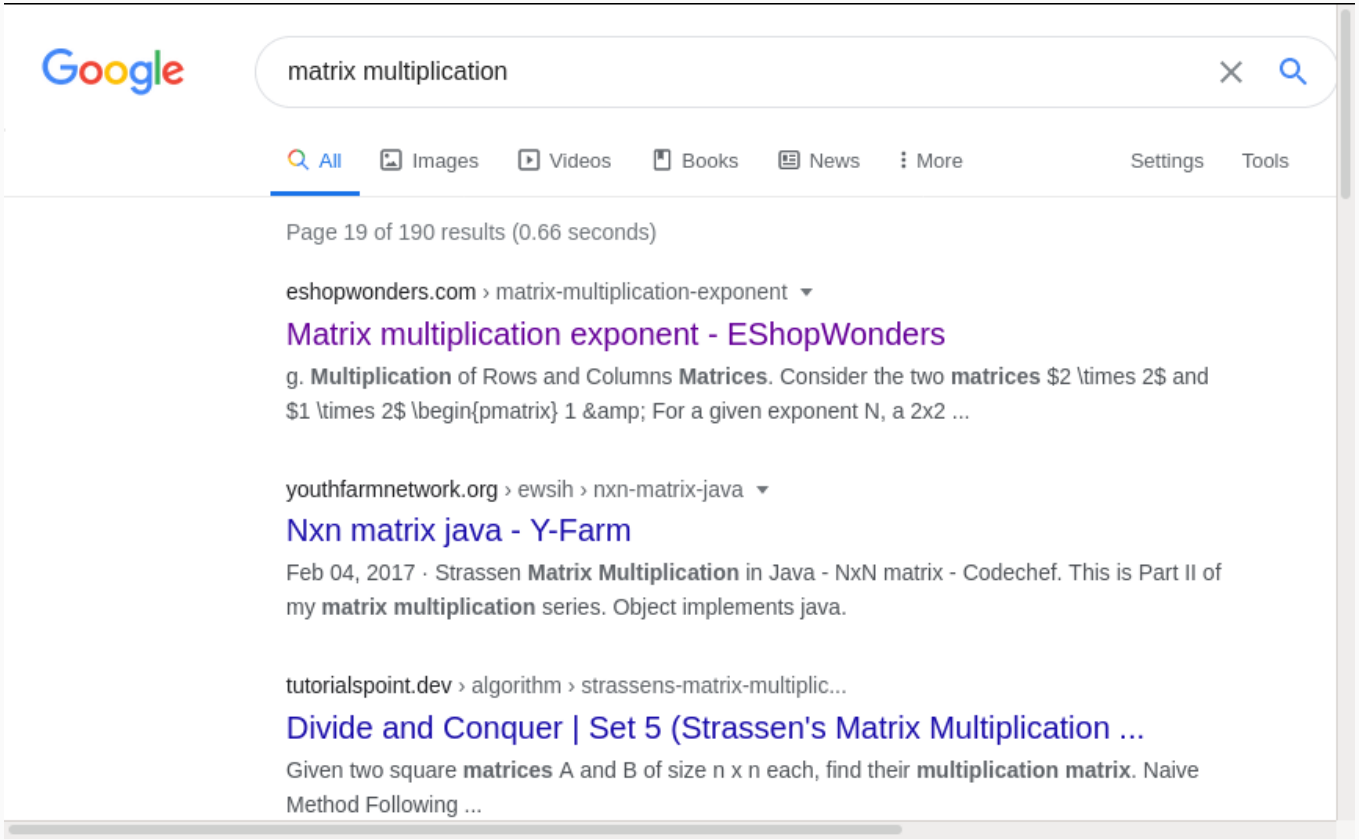
Cloaking (Hidden Content)

Crawlers announce themselves

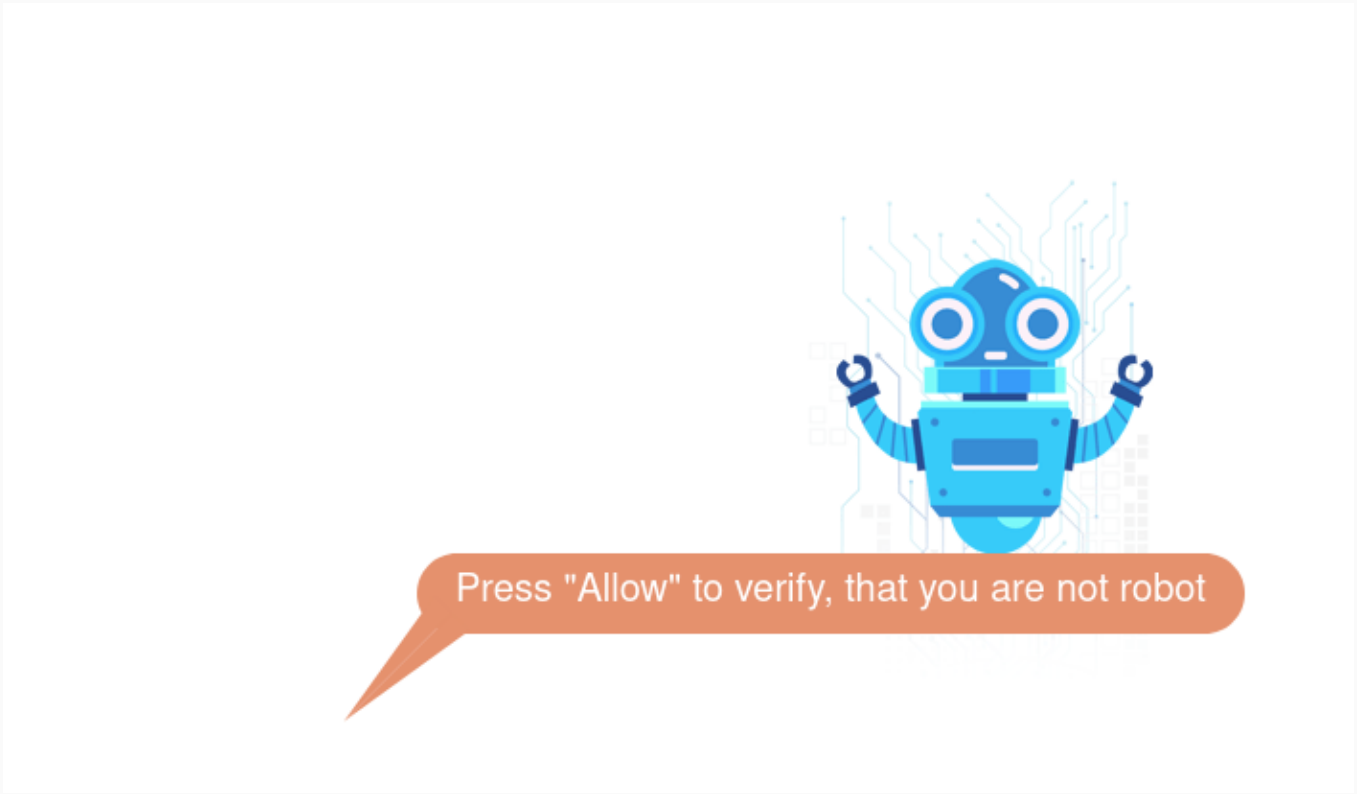
```
User-Agent: Mozilla/5.0 (compatible; AhrefsBot/6.1;
                                        +http://ahrefs.com/robot/)
User-Agent: Googlebot/2.1 (+http://www.google.com/bot.html)
```

Attacker can serve different contents to Crawlers and to Users!

Cloaking Example

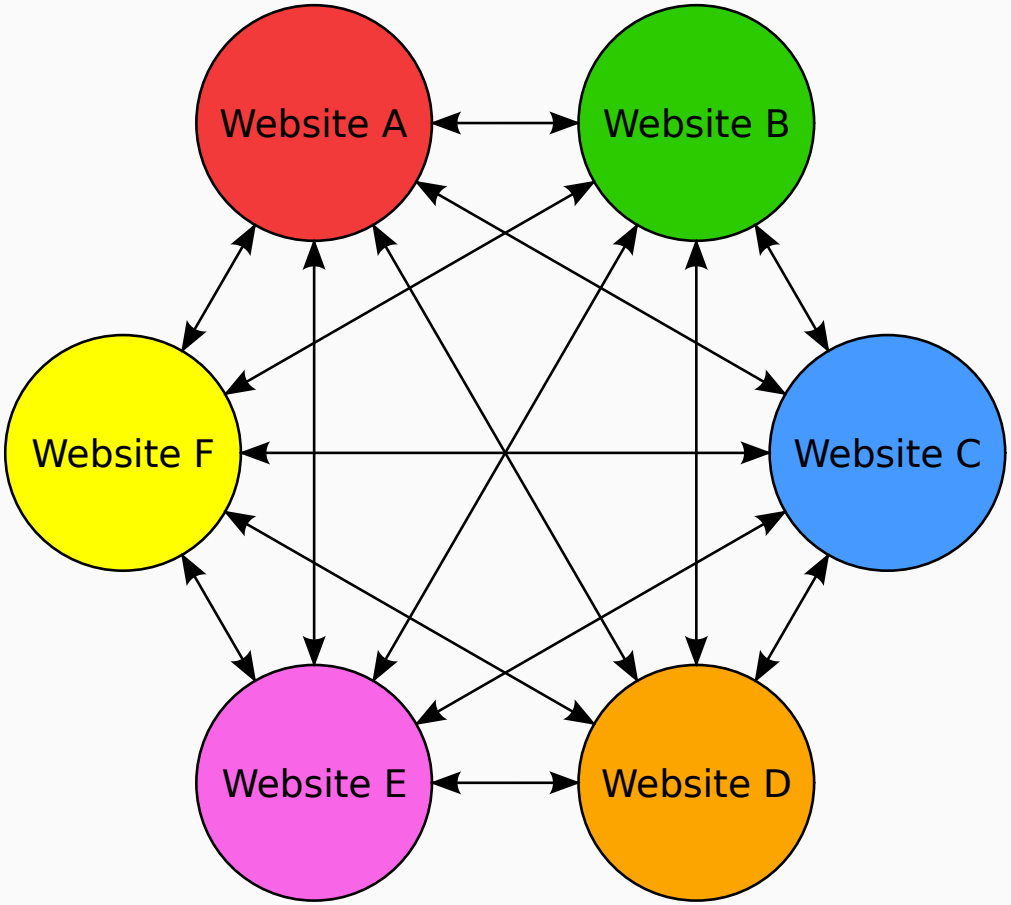


Cloaking Example



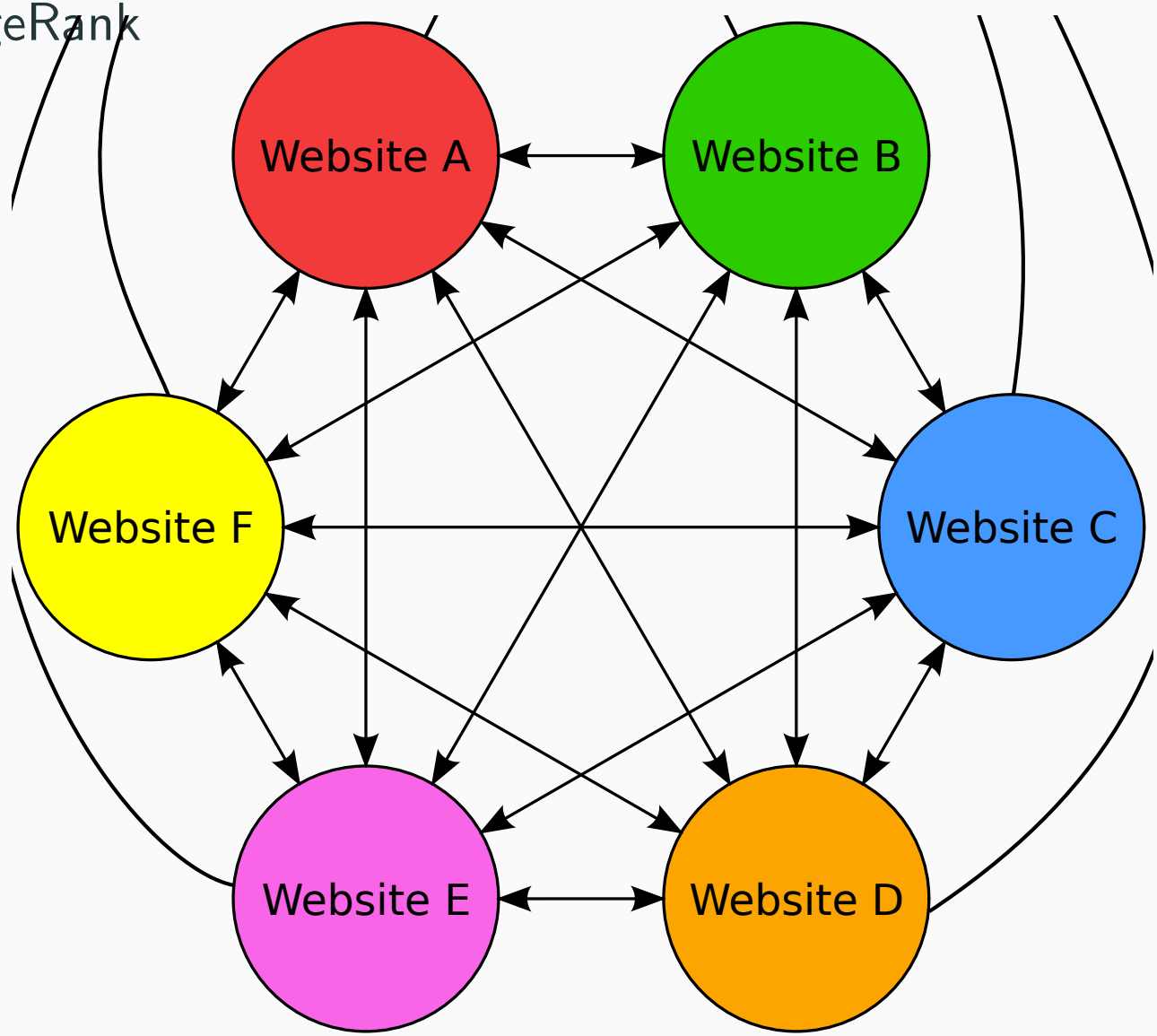
Link Farm

Lots of fraudulent webpages linking to each other.



Link Farm

For PageRank



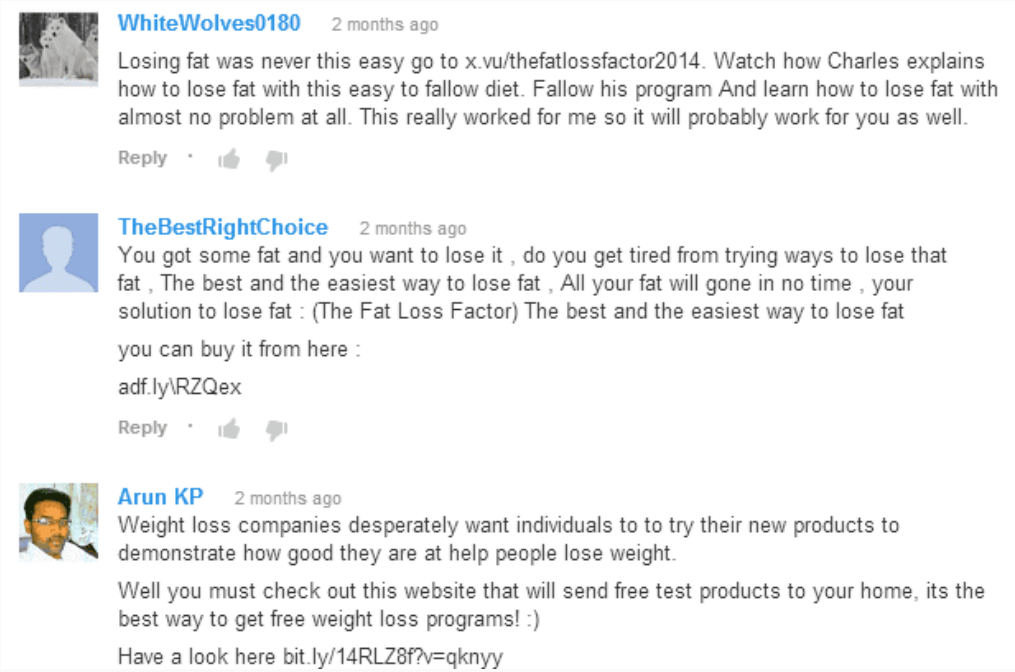
Word Selection

Repeat (in an hidden way) lots of time some keyword to improve tf-idf

```
<div style="display:none;">
Bank Money Bitcoin
Bank Money Bitcoin
Bank Money Bitcoin
Bank Money Bitcoin
Bank Money Bitcoin
</div>
```

“Fake” backlinks

Put comments on trustworthy websites to improve backlink quality



WhiteWolves0180 2 months ago

Losing fat was never this easy go to x.vu/thefatlossfactor2014. Watch how Charles explains how to lose fat with this easy to fallow diet. Fallow his program And learn how to lose fat with almost no problem at all. This really worked for me so it will probably work for you as well.

Reply

TheBestRightChoice 2 months ago

You got some fat and you want to lose it , do you get tired from trying ways to lose that fat , The best and the easiest way to lose fat , All your fat will gone in no time , your solution to lose fat : (The Fat Loss Factor) The best and the easiest way to lose fat

you can buy it from here :

adf.ly\RZQex

Reply

Arun KP 2 months ago

Weight loss companies desperately want individuals to to try their new products to demonstrate how good they are at help people lose weight.

Well you must check out this website that will send free test products to your home, its the best way to get free weight loss programs! :)

Have a look here bit.ly/14RLZ8f?v=qknyy

Spamdexing: Attacks and Counter attacks

Defense

PageRank protection

- Analyze the graph to find strange patterns

  cliques, two-way links, etc.

- Detect copied or machine generated content of low interest

  e.g. Mirror websites

- Curating a list of top quality websites

  define PageRank semi-manually there

Keyword protection

- Use the linker to determine linkee topic

  Click here for more information on Plumbus.

- Detect copied or machine generated content of low interest

  e.g. Mirror websites

- Curating a list of top quality websites

  define PageRank semi-manually from there

Link protection

- Determine link interest

  Links near the top are most interesting than links from the comment section.

- Incite webmaster to best practices

  enforce `rel="nofollow"` on external links

- Determine links from user generated content

Improving the Information Retrieval

Stemming

Stemming consists in replacing a word with its "root form".

It can:

- reduce synonymy

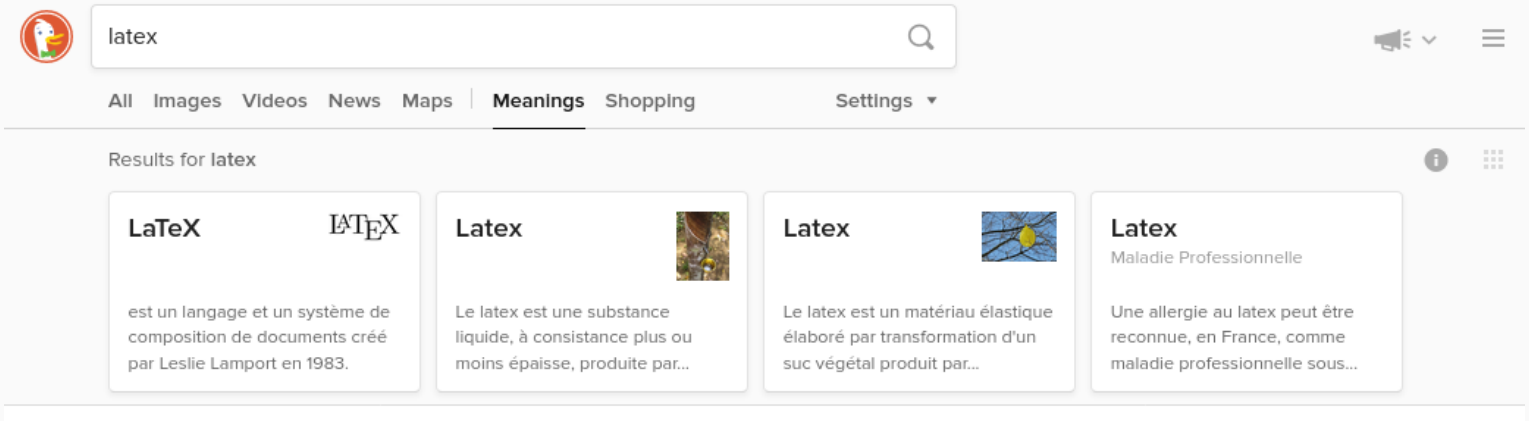
  car, *auto*, *automobile*

- remove conjugation

  dream, *dreamt*, *dreamed*

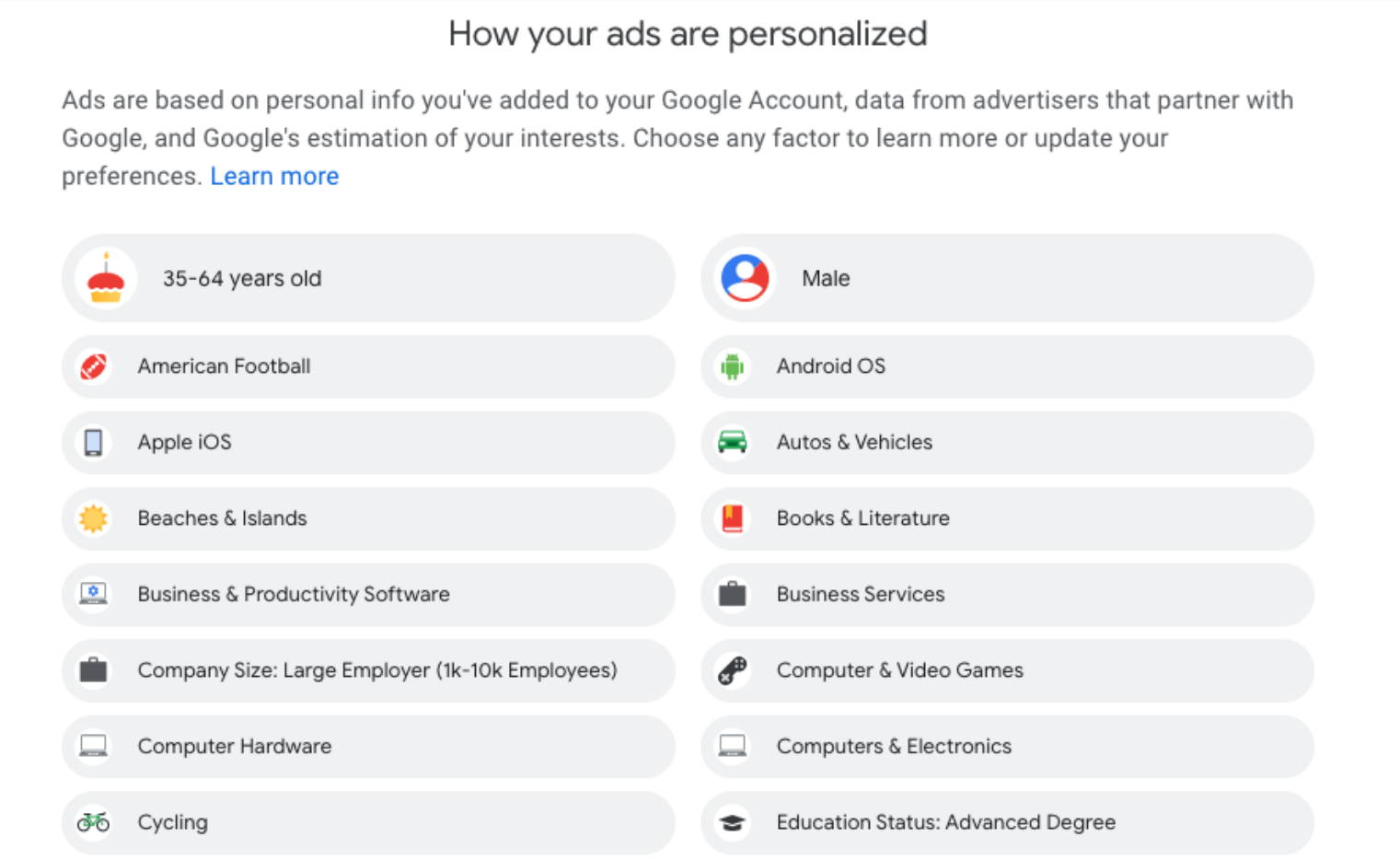
Word Embeddings

Using Latent semantic analysis to either have fuzzy matchs or disambiguate.

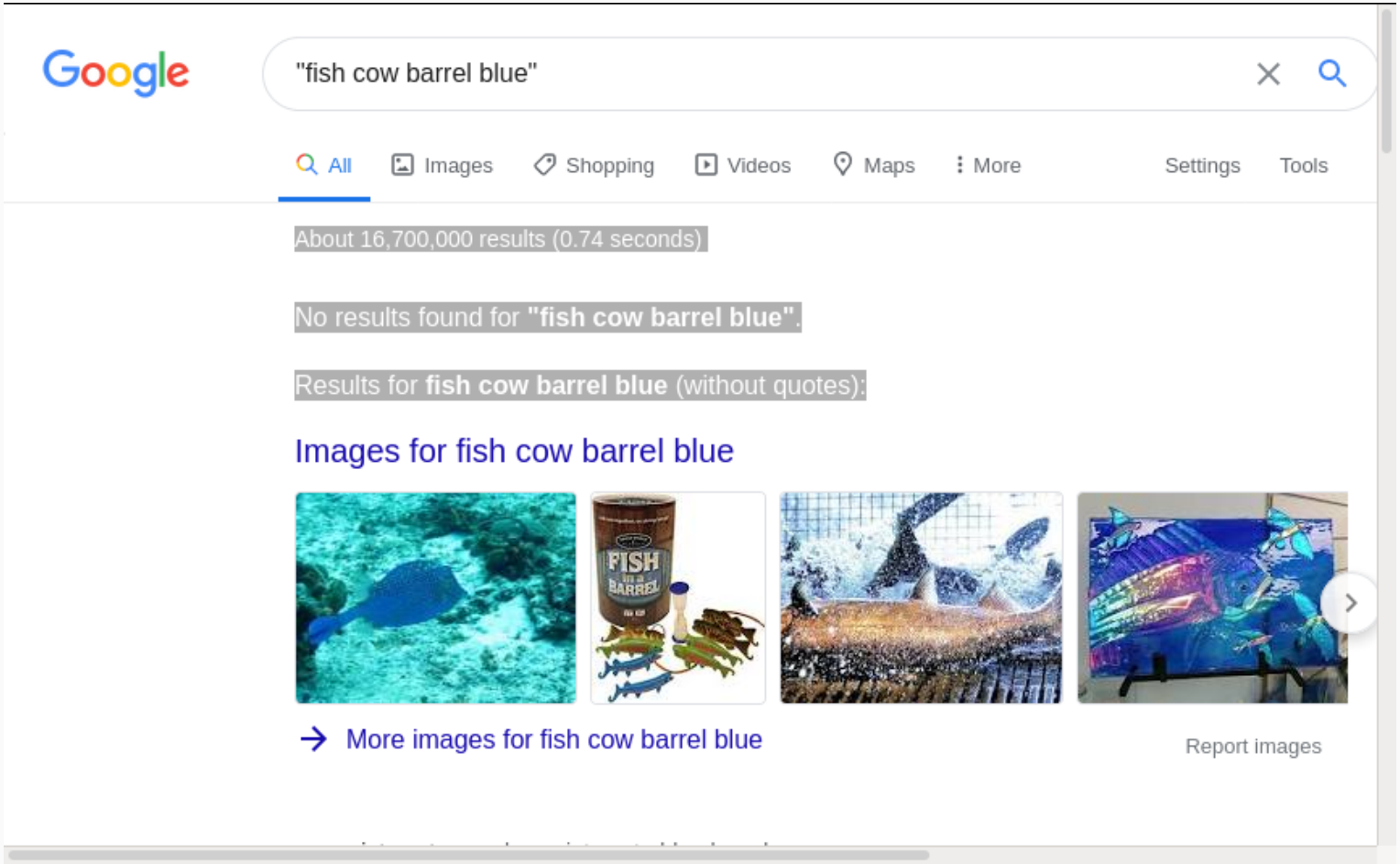


Topic model

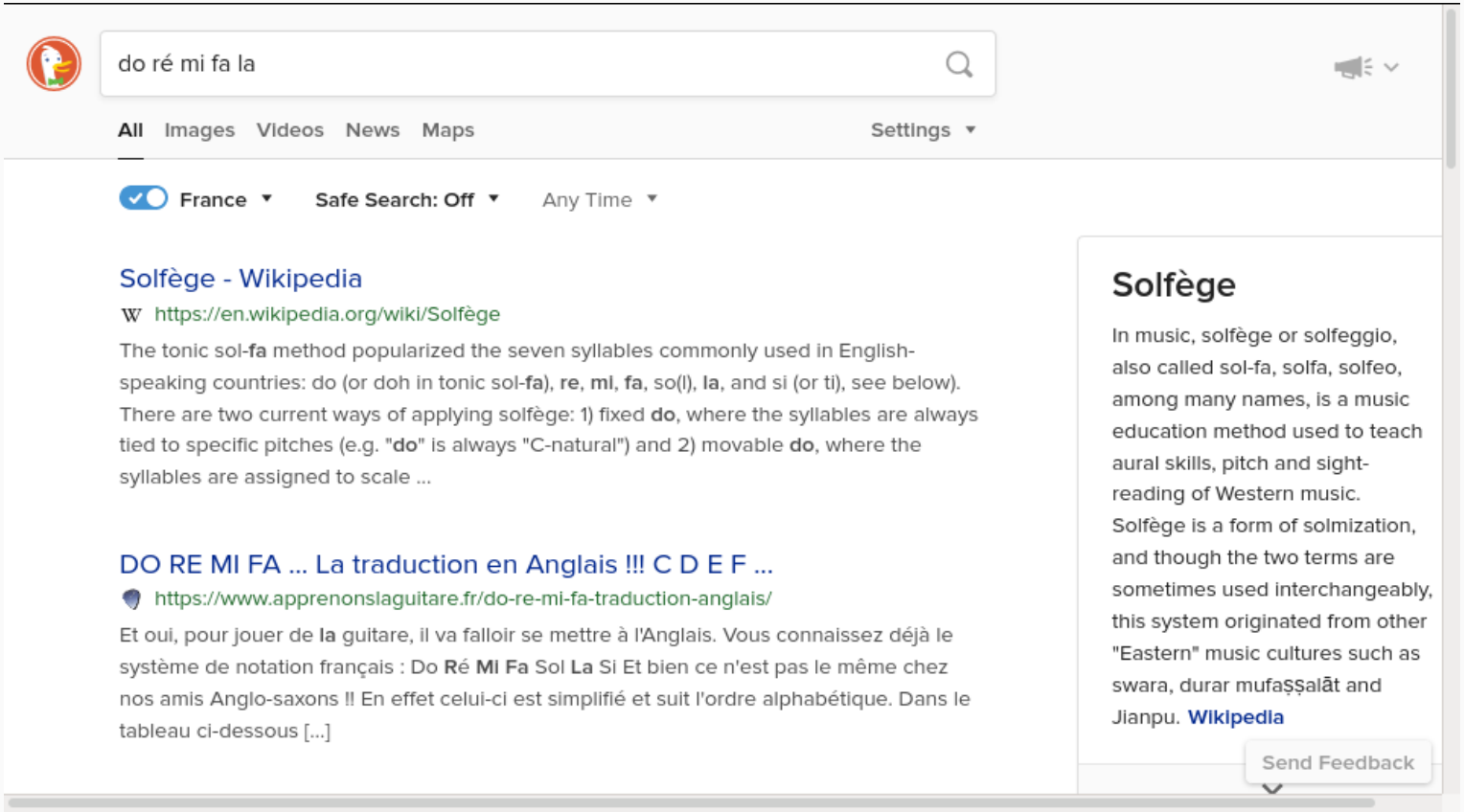
Determine a set of topics for a user to personalize results



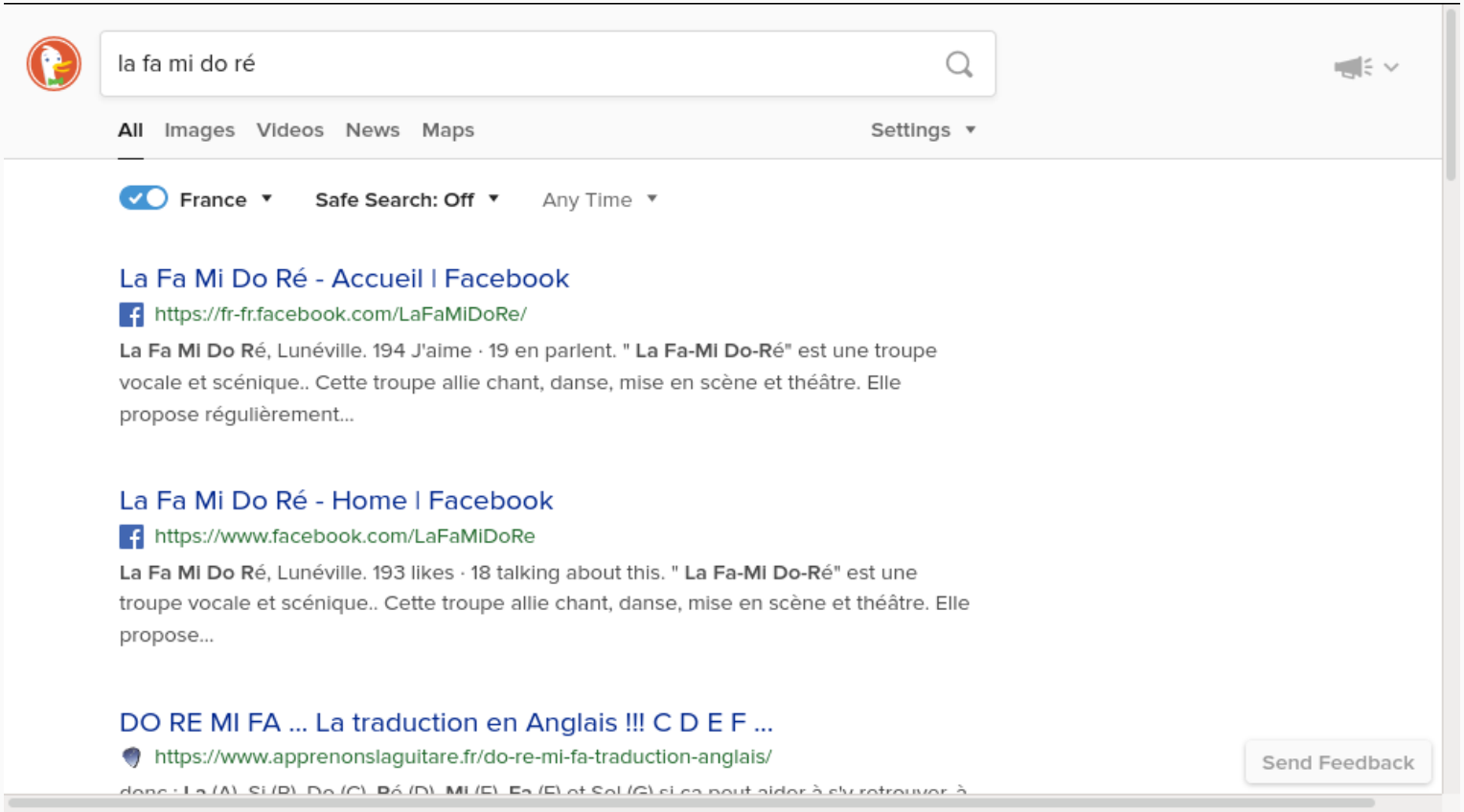
Force some words or groups of words to be there



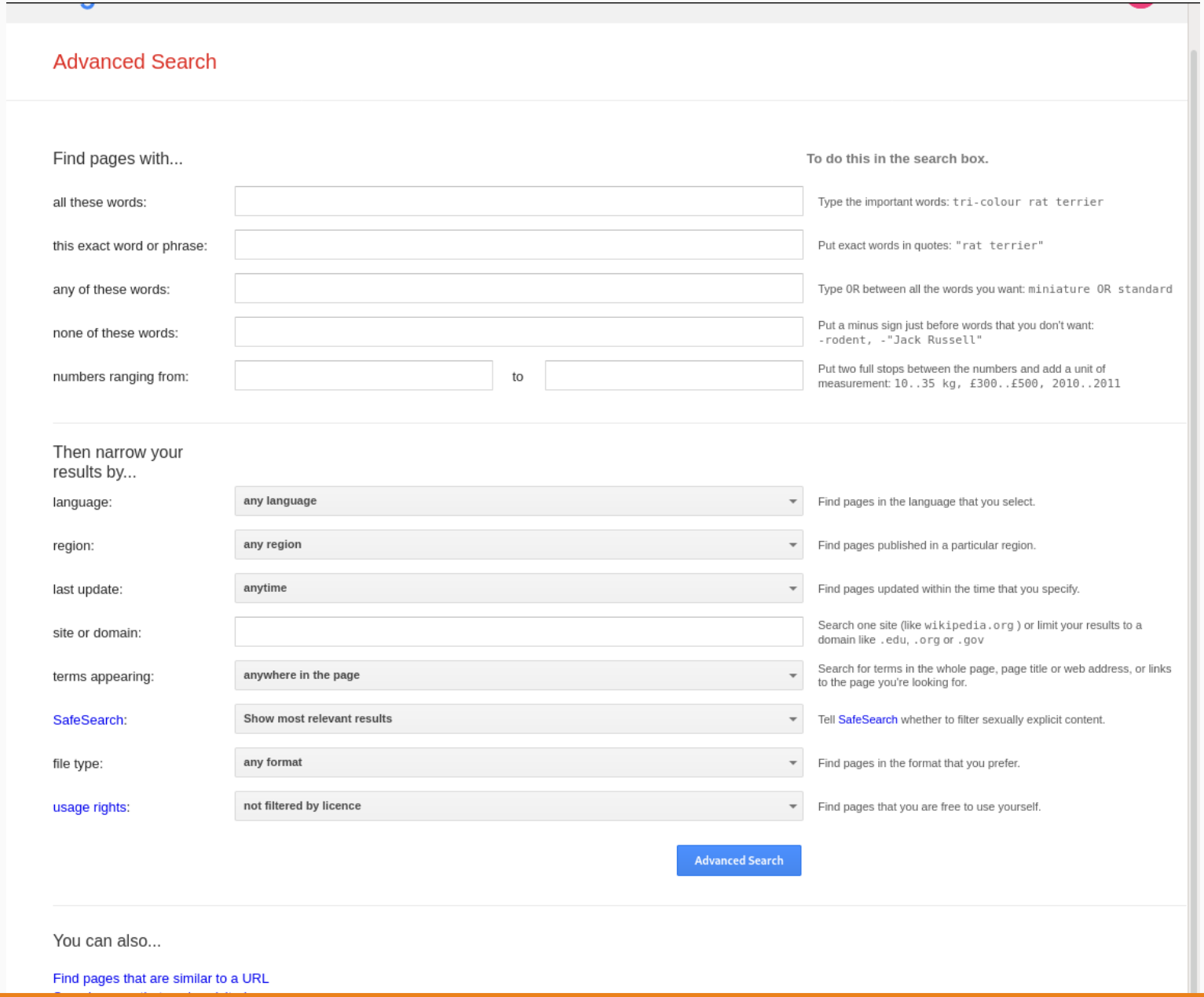
Force locality



Force locality

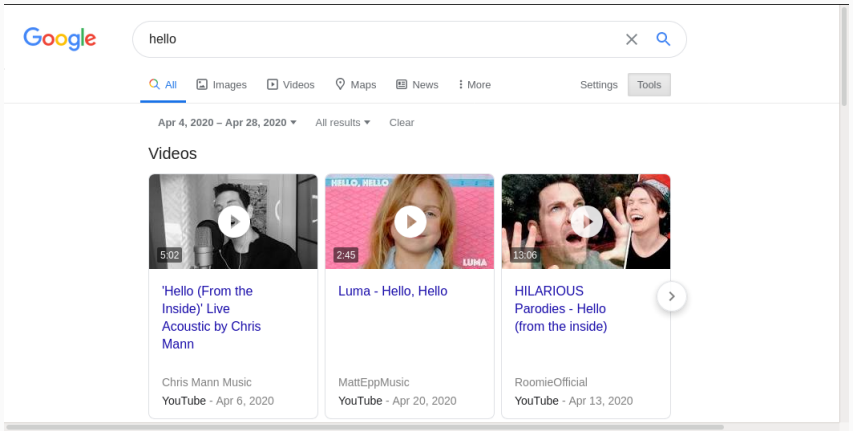


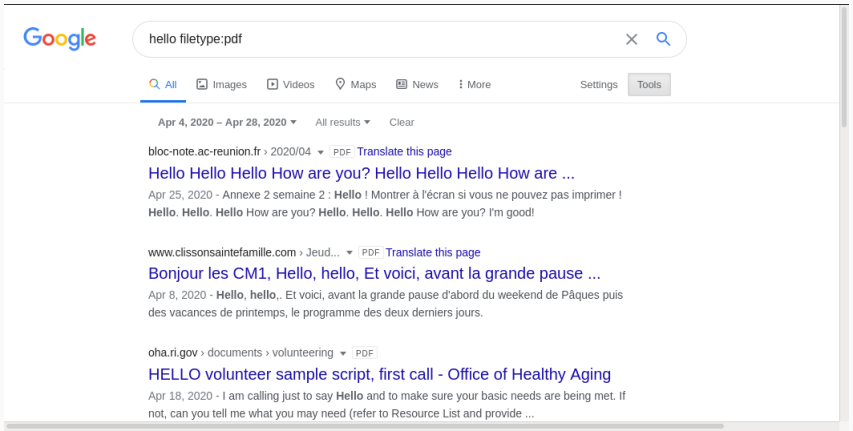
Advanced Search



Metadata search

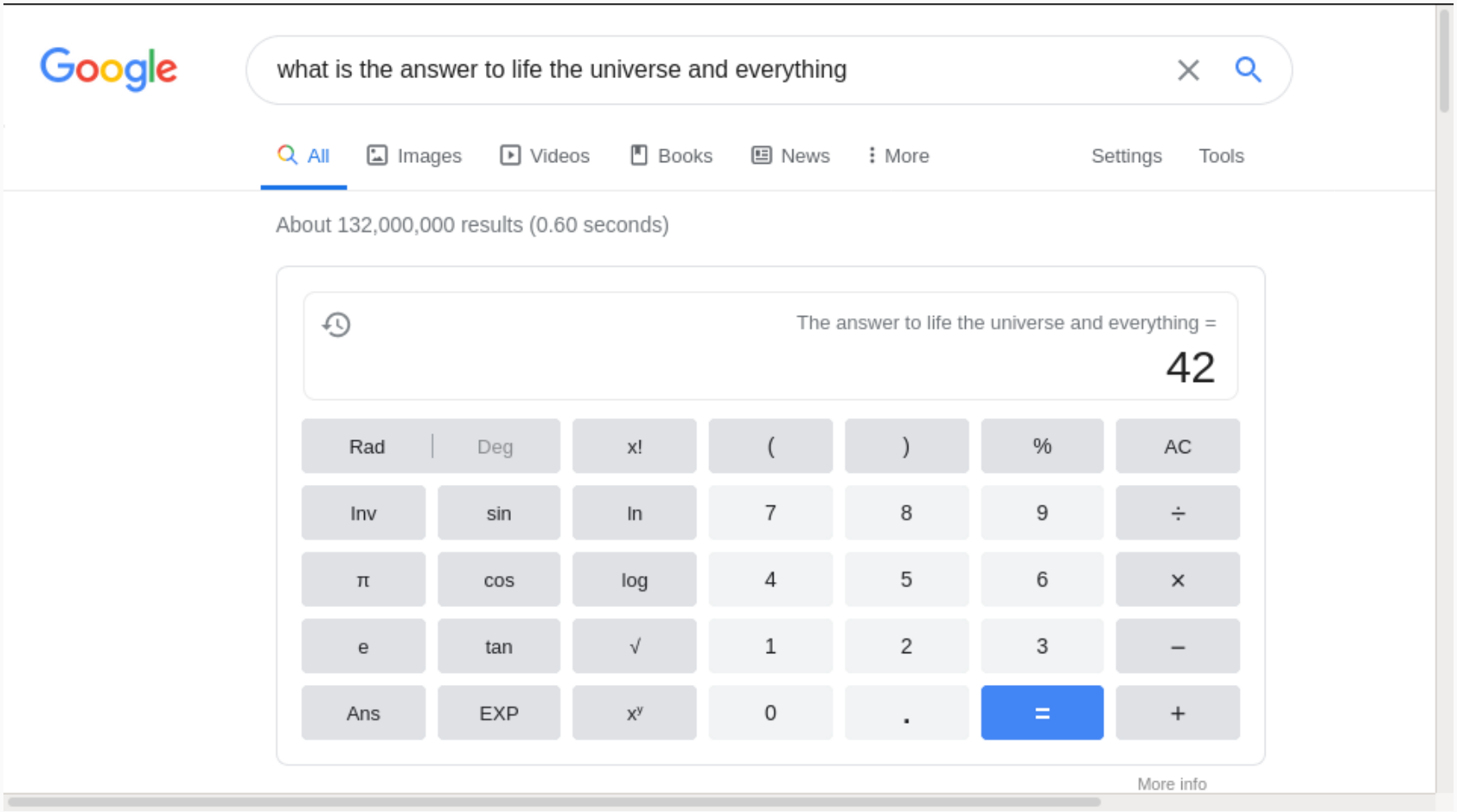
For instance, determine lang, period of time, information on documents, etc.





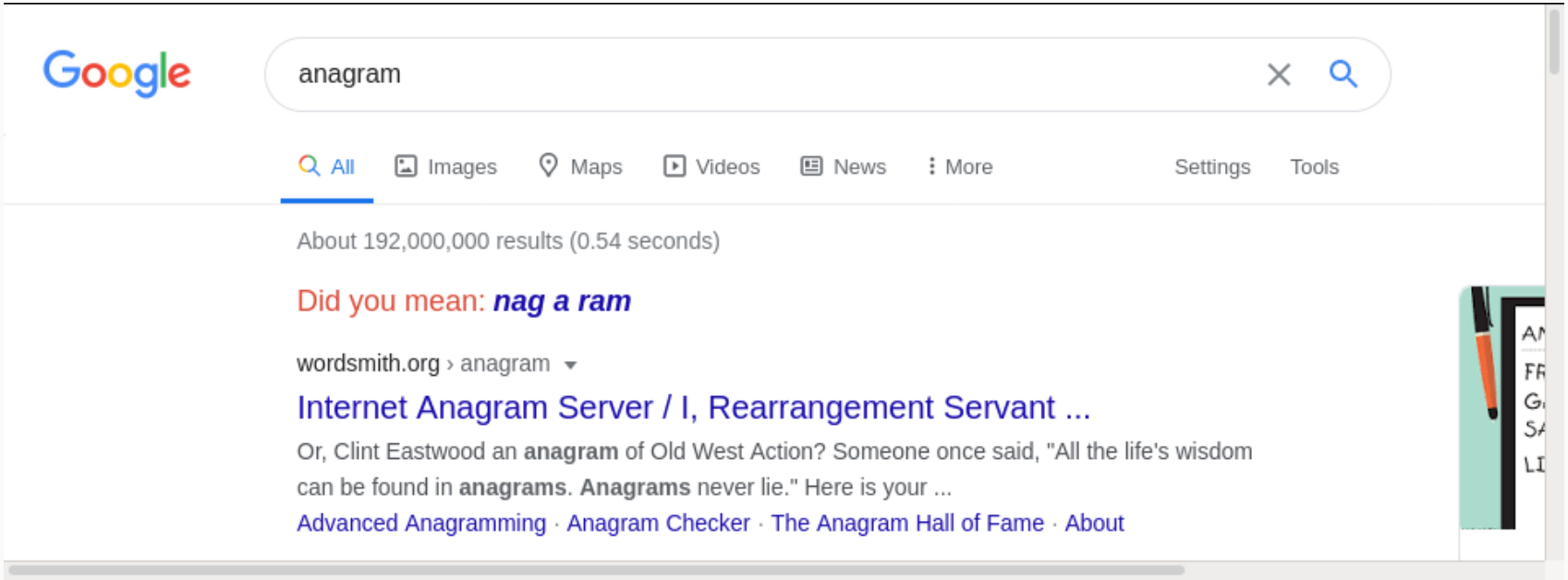
Manually curated search

Manually (or semi-automatically) enforce results for some queries



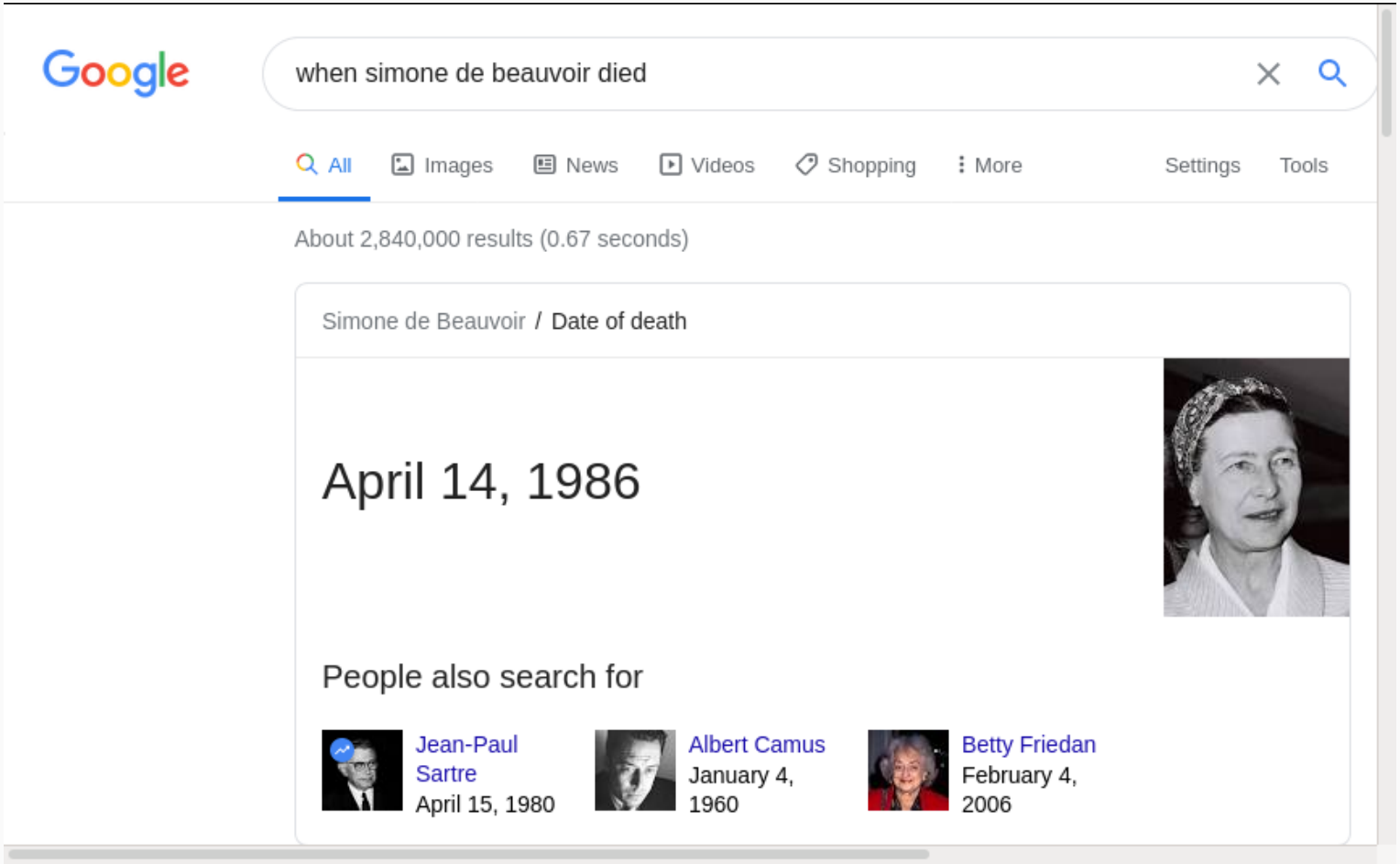
Manually curated search

Manually (or semi-automatically) enforce results for some queries

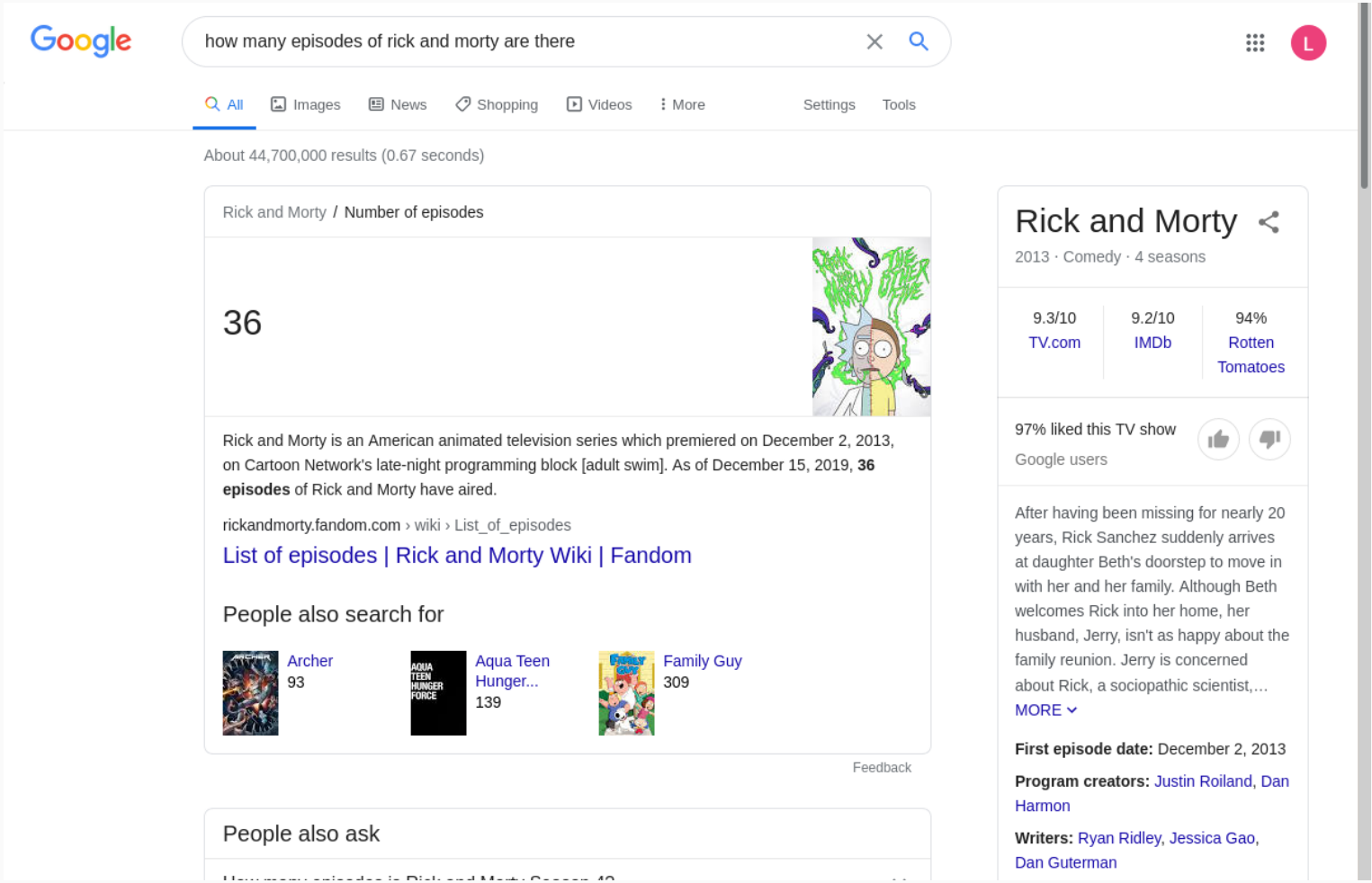


Beyond webpages

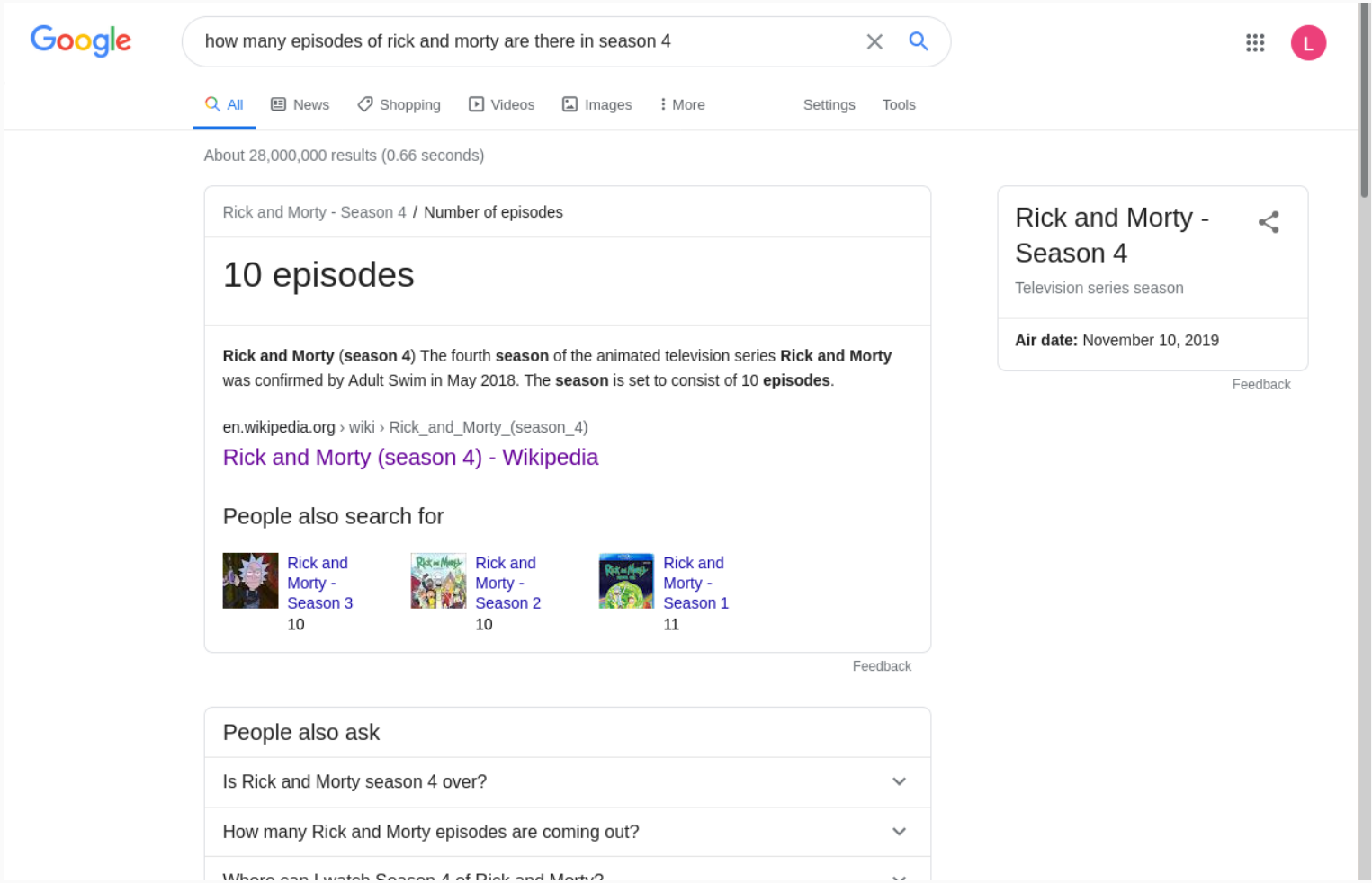
Semantic querying



Semantic querying



Semantic querying



Semantic querying