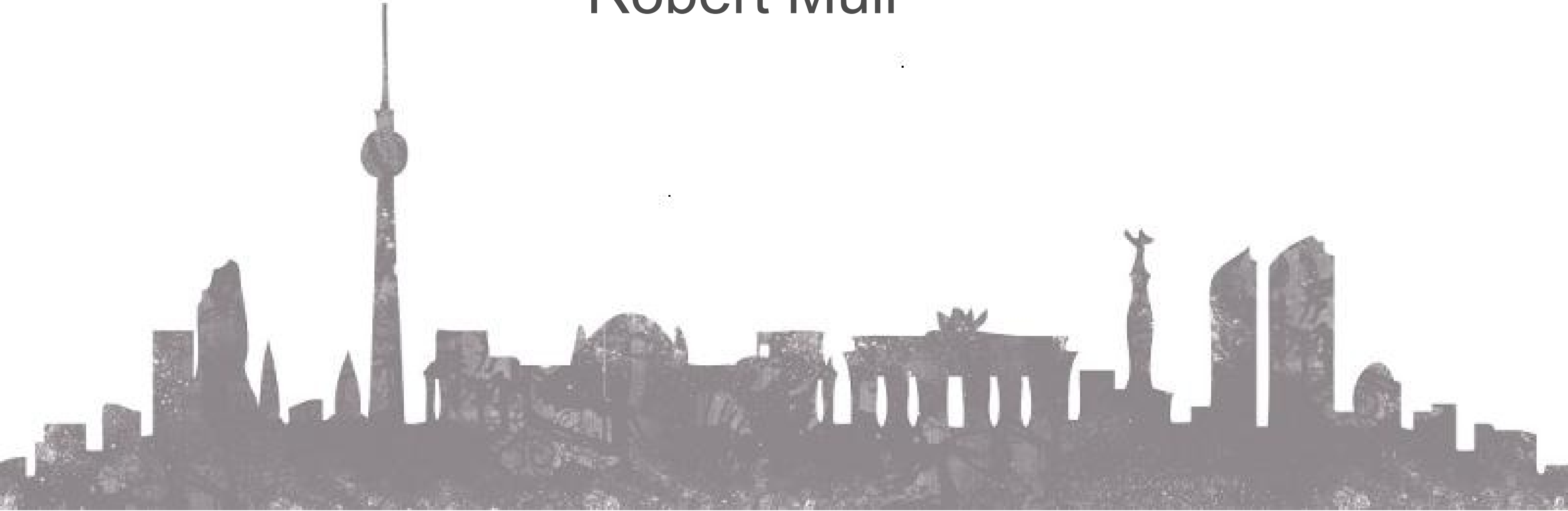
Apache Lucene 4

Robert Muir



Agenda

- Overview of Lucene
- Conclusion
- Resources
- Q & A

Download of Lucene:

core/

analysis/

queryparser/

highlighter/

suggest/

expressions/

join/

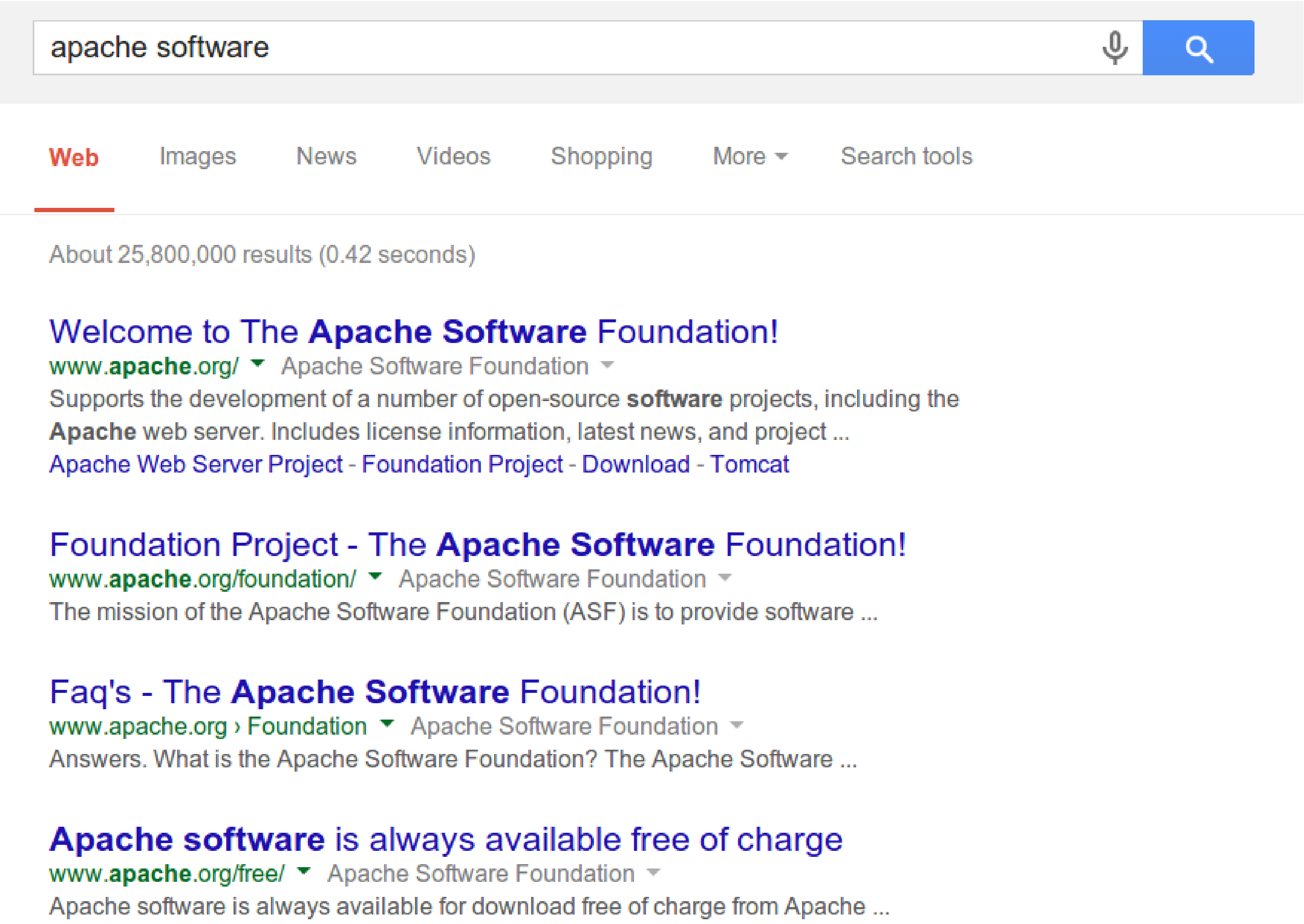
memory/

codecs/

...

core/

Lucene core library



apache software

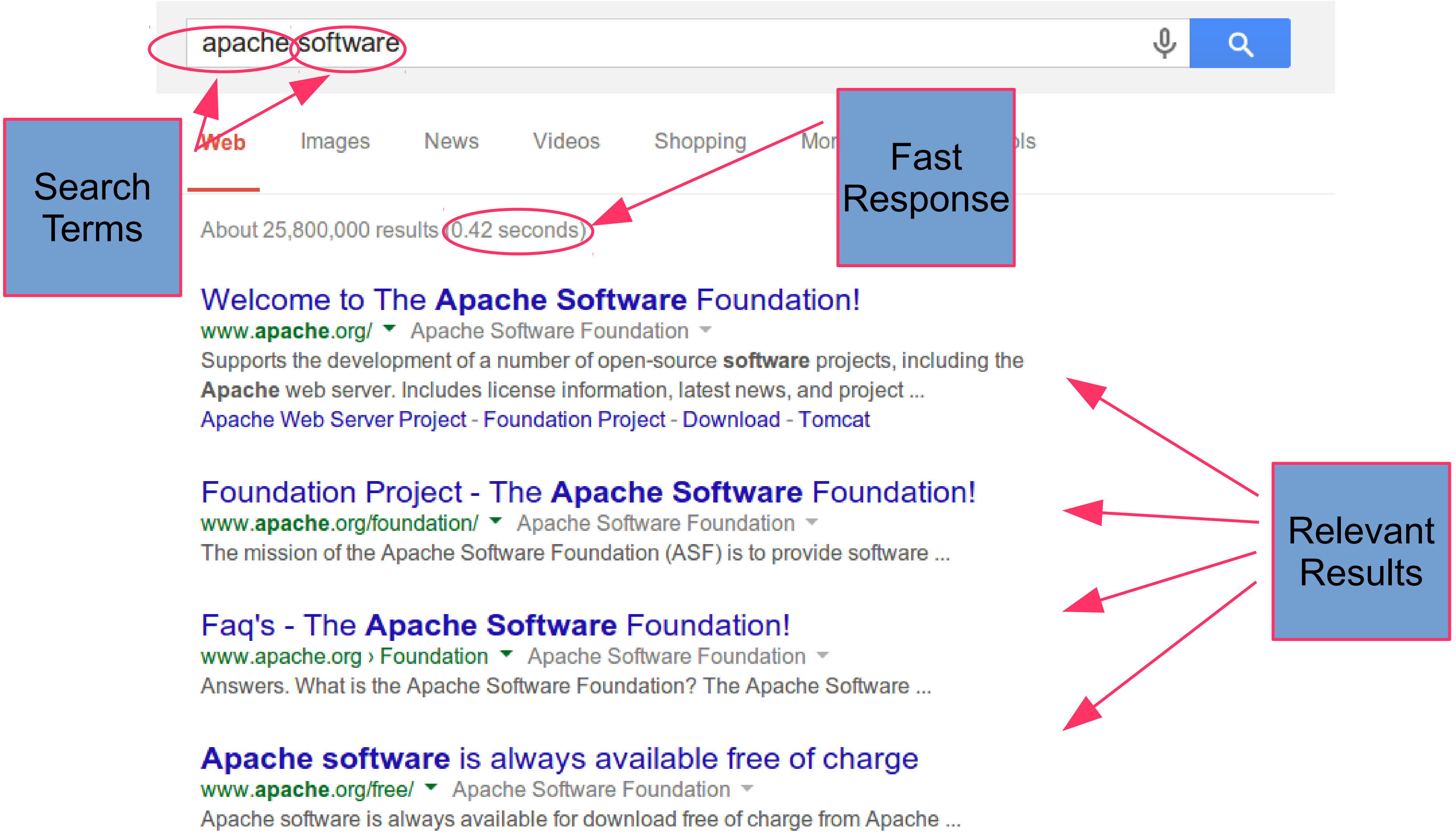
About 25,800,000 results (0.42 seconds)

Welcome to The **Apache Software** Foundation!
www.**apache**.org/ Apache Software Foundation
Supports the development of a number of open-source **software** projects, including the **Apache** web server. Includes license information, latest news, and project ...
Apache Web Server Project - Foundation Project - Download - Tomcat

Foundation Project - The **Apache Software** Foundation!
www.**apache**.org/foundation/ Apache Software Foundation
The mission of the Apache Software Foundation (ASF) is to provide software ...

Faq's - The **Apache Software** Foundation!
www.apache.org › Foundation Apache Software Foundation
Answers. What is the Apache Software Foundation? The Apache Software ...

Apache software is always available free of charge
www.**apache**.org/free/ Apache Software Foundation
Apache software is always available for download free of charge from Apache ...



apache software

Search Terms

Fast Response

Relevant Results

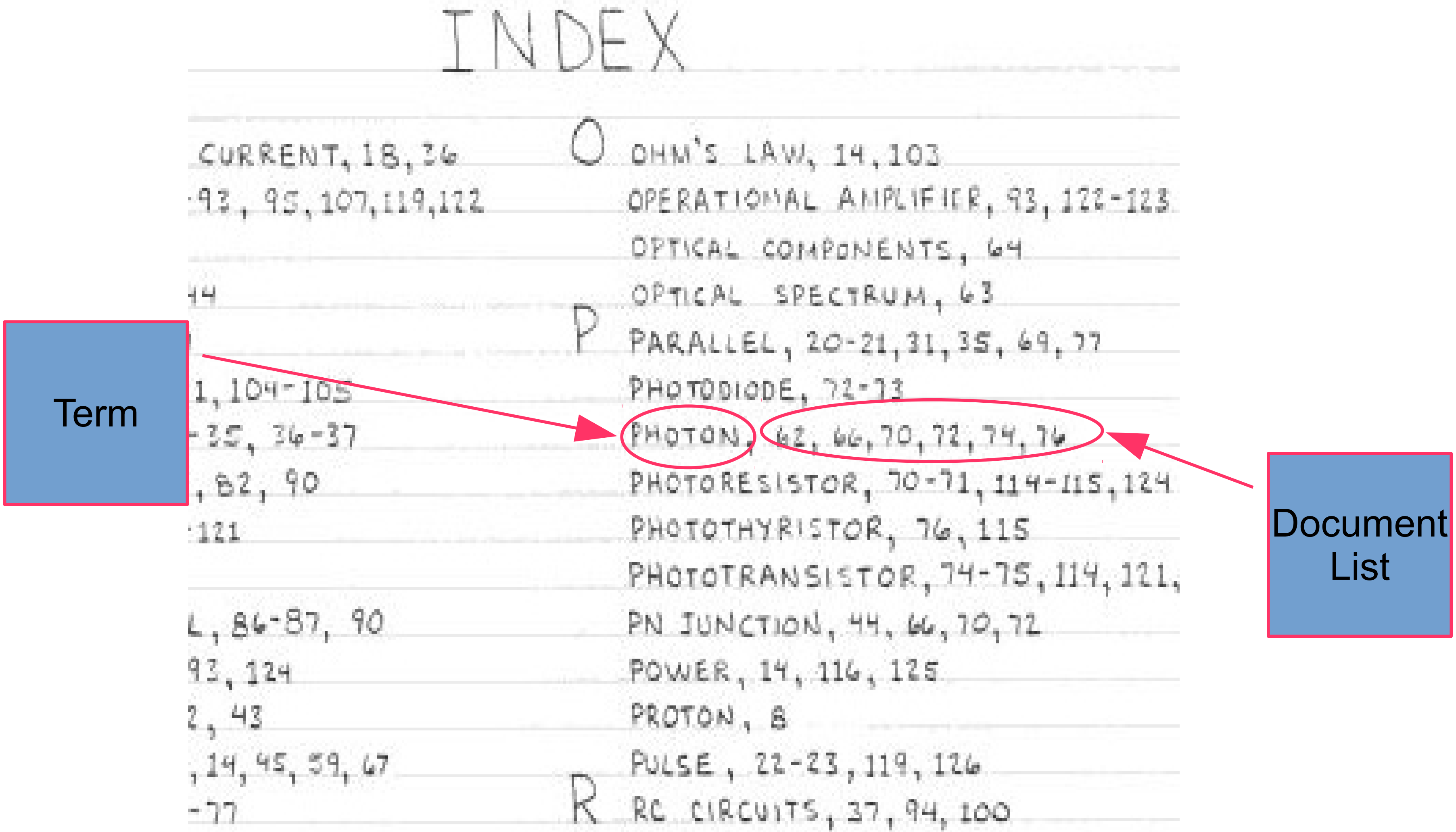
About 25,800,000 results (0.42 seconds)

Welcome to The Apache Software Foundation!
www.apache.org/ ▾ Apache Software Foundation ▾
Supports the development of a number of open-source **software** projects, including the **Apache** web server. Includes license information, latest news, and project ...
Apache Web Server Project - Foundation Project - Download - Tomcat

Foundation Project - The Apache Software Foundation!
www.apache.org/foundation/ ▾ Apache Software Foundation ▾
The mission of the Apache Software Foundation (ASF) is to provide software ...

Faq's - The Apache Software Foundation!
www.apache.org › Foundation ▾ Apache Software Foundation ▾
Answers. What is the Apache Software Foundation? The Apache Software ...

Apache software is always available free of charge
www.apache.org/free/ ▾ Apache Software Foundation ▾
Apache software is always available for download free of charge from Apache ...



INDEX

CURRENT, 18, 36
93, 95, 107, 119, 122
44
1, 104-105
-35, 36-37
, 82, 90
121
L, 86-87, 90
93, 124
2, 43
, 14, 45, 59, 67
-77

O
OHM'S LAW, 14, 103
OPERATIONAL AMPLIFIER, 93, 122-123
OPTICAL COMPONENTS, 64
OPTICAL SPECTRUM, 63

P
PARALLEL, 20-21, 31, 35, 69, 77
PHOTODIODE, 72-73
PHOTON, 62, 66, 70, 72, 74, 76
PHOTORESISTOR, 70-71, 114-115, 124
PHOTOTHYRISTOR, 76, 115
PHOTOTRANSISTOR, 74-75, 114, 121,
PN JUNCTION, 44, 66, 70, 72
POWER, 14, 116, 125
PROTON, 8
PULSE, 22-23, 119, 126

R
RC CIRCUITS, 37, 94, 100

Term

Document List

Indexing with Lucene

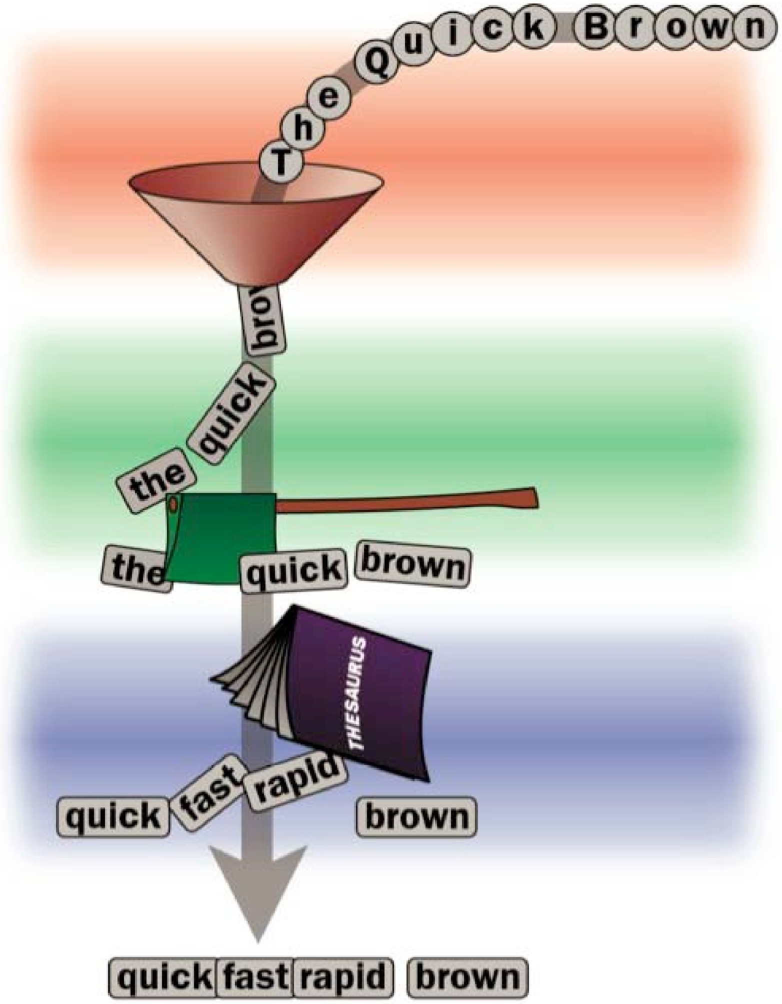
- Fast: over 200GB/hour
- Incremental and “near-realtime”
- Multi-threaded
- Beyond full-text: numbers, dates, binary, ...
- Customize what is indexed (“analysis”)
- Customize index format (“codecs”)

analysis/

text analysis for search

Analysis

- Breaks field into words.
- Normalizes words:
  - Remove accents.
  - Convert to lowercase.
  - Remove plurals.
  - ...
- Depends on language.
- Depends on use case.



Analysis (cont)

- Customizable “chain” or “pipeline”
- **Tokenizer**: split text into words.
- **TokenFilter**: modify words.
- **Analyzer**: complete chain.
  - Tokenizer + some TokenFilters

Analysis: Tokenizer

- Example document:
  - “Hope there are beers at Buzzwords”
- WhitespaceTokenizer:
  - [Hope, there, are, beers, at, Buzzwords]
- 21 tokenizers out of box.
  - Or use your own.

Analysis: TokenFilter

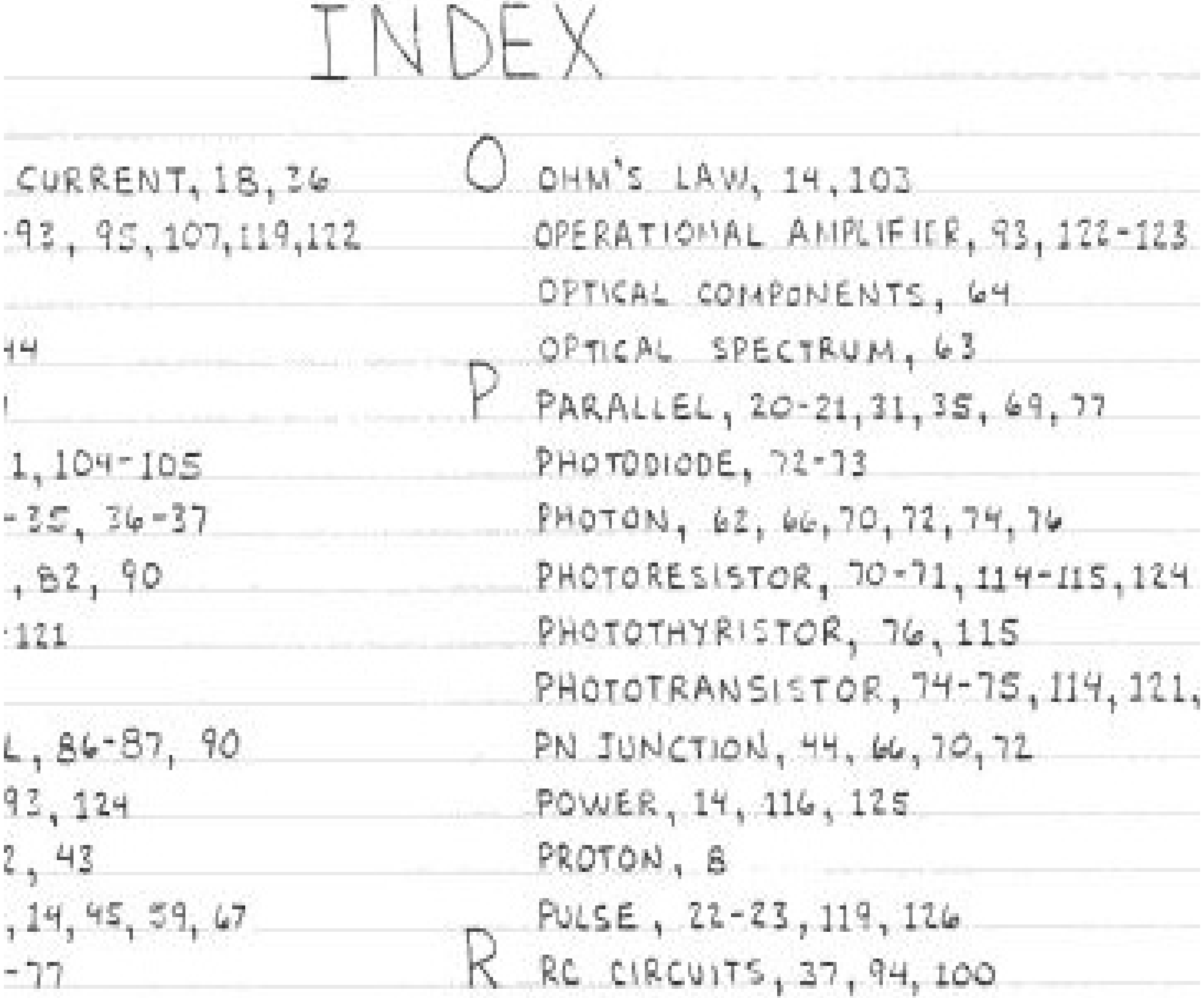
- [Hope, there, are, beers, at, Buzzwords]
- LowerCaseFilter:
  - [**hope**, there, are, beers, at, **buzzwords**]
- EnglishMinimalStemFilter:
  - [hope, there, are, **beer**, at, **buzzword**]
- 99 tokenfilters out of box.
  - Or use your own.

Analysis: Analyzer

- Chain of Tokenizer + some TokenFilters.
- Used at both **index** and **query** time.
- Analyzers for 35 languages out of box.
    - Or use your own.

queryparser/

query languages



INDEX

CURRENT, 18, 36
93, 95, 107, 119, 122

44
1
1, 104-105
-35, 36-37
, 82, 90
121

L, 86-87, 90
93, 124
2, 43
, 14, 45, 59, 67
-77

O
OHM'S LAW, 14, 103
OPERATIONAL AMPLIFIER, 93, 122-123
OPTICAL COMPONENTS, 64
OPTICAL SPECTRUM, 63

P
PARALLEL, 20-21, 31, 35, 69, 77
PHOTODIODE, 72-73
PHOTON, 62, 66, 70, 72, 74, 76
PHOTORESISTOR, 70-71, 114-115, 124
PHOTOTHYRISTOR, 76, 115
PHOTOTRANSISTOR, 74-75, 114, 121,
PN JUNCTION, 44, 66, 70, 72
POWER, 14, 116, 125
PROTON, 8
PULSE, 22-23, 119, 126

R
RC CIRCUITS, 37, 94, 100

Queries

circuit: 5, 6, 8

parallel: 4, 5, 6

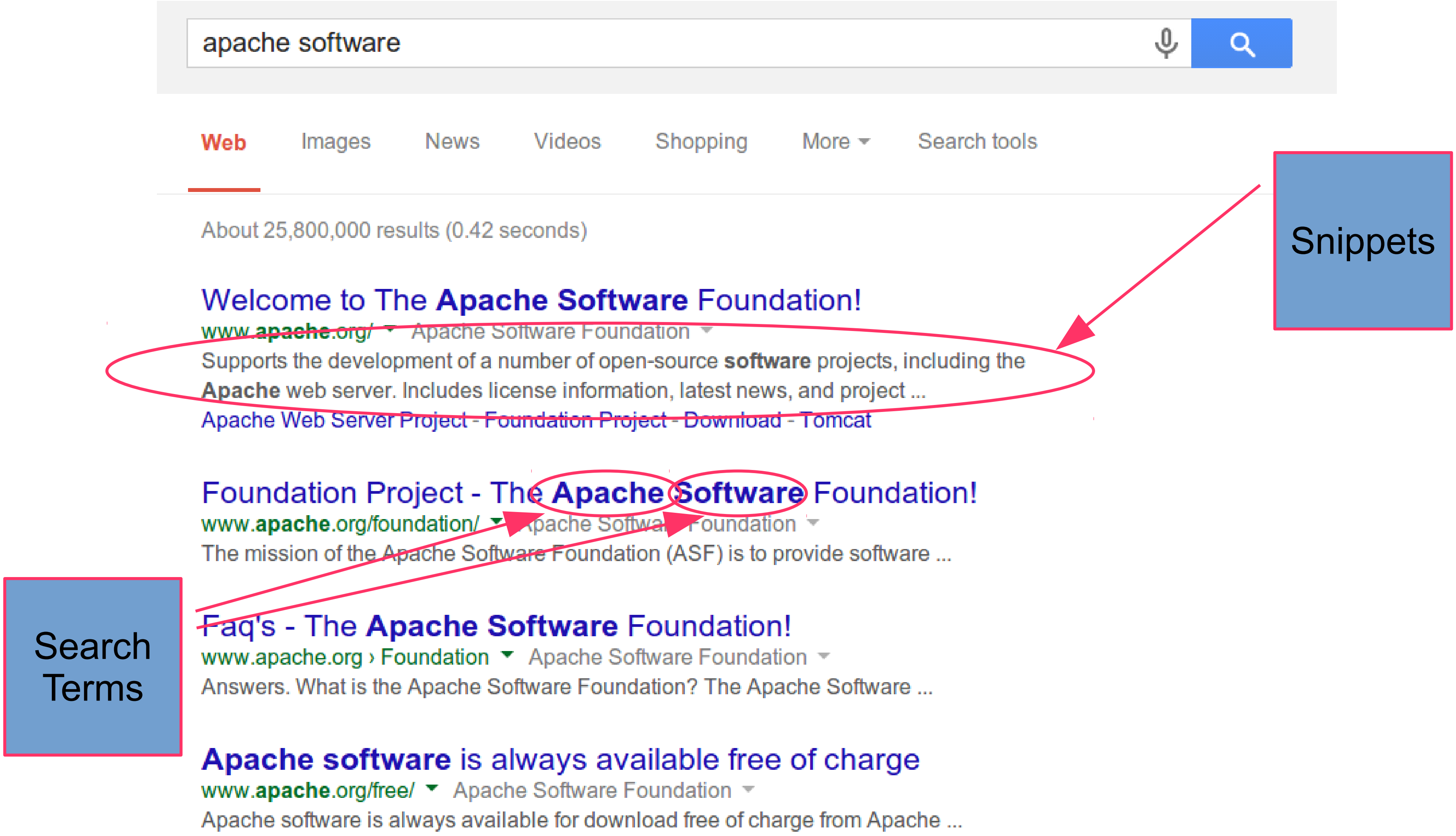
- circuit **OR** parallel: 4, 5, 6, 8 (union)
- circuit **AND** parallel: 5, 6 (intersection)
- circuit **NOT** parallel: 8 (subtraction)

Queryparsers

- “Classic”: strict syntax.
  - cheap AND (buzzwords OR beer)
  - +cheap +(buzzwords beer)
- “Simple”: lenient syntax.
  - +cheap +(buzzwords beer
- Or build queries via API

highlighter/

query-biased summaries



apache software

About 25,800,000 results (0.42 seconds)

Welcome to The **Apache Software** Foundation!
www.**apache**.org/ Apache Software Foundation
Supports the development of a number of open-source **software** projects, including the **Apache** web server. Includes license information, latest news, and project ...
Apache Web Server Project - Foundation Project - Download - Tomcat

Foundation Project - The **Apache Software** Foundation!
www.**apache**.org/foundation/ Apache Software Foundation
The mission of the Apache Software Foundation (ASF) is to provide software ...

Faq's - The **Apache Software** Foundation!
www.apache.org › Foundation Apache Software Foundation
Answers. What is the Apache Software Foundation? The Apache Software ...

Apache software is always available free of charge
www.**apache**.org/free/ Apache Software Foundation
Apache software is always available for download free of charge from Apache ...

Snippets

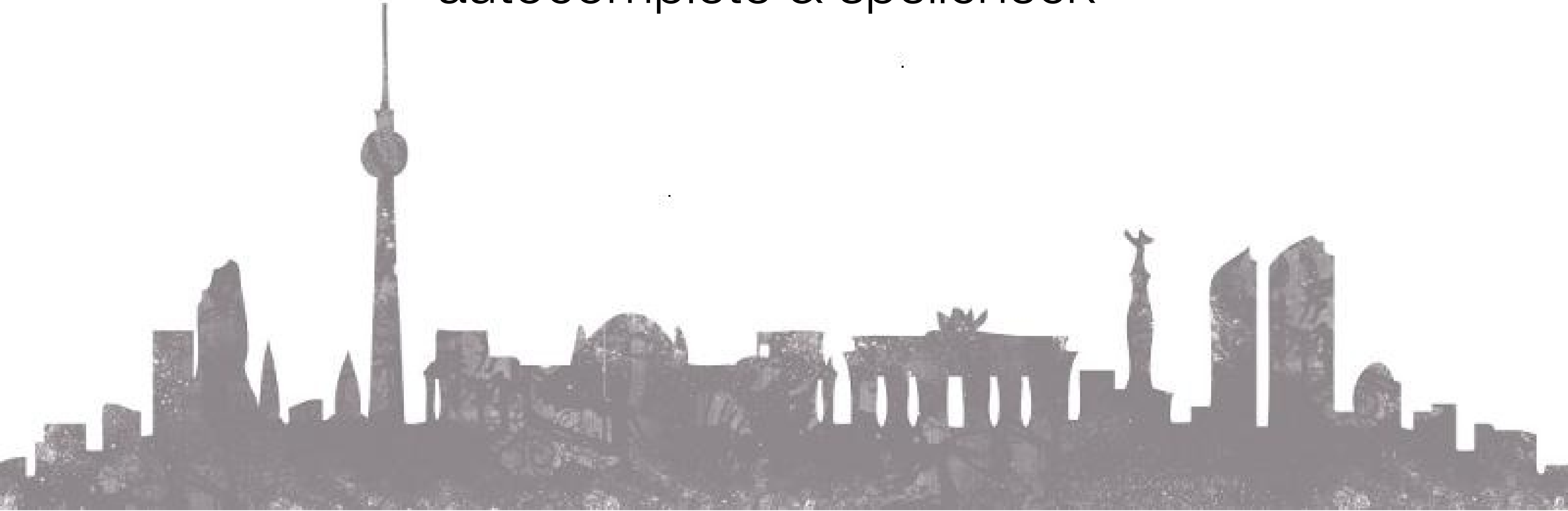
Search Terms

Highlighting

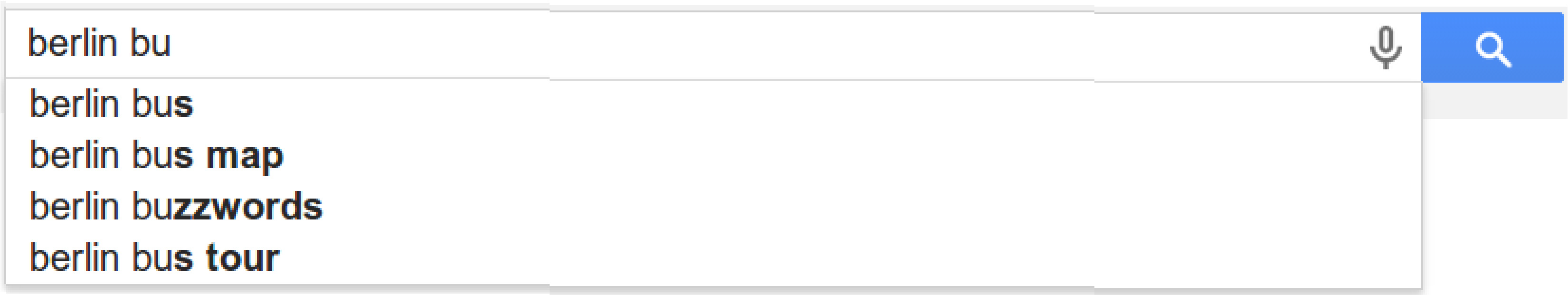
- Three choices in Lucene
  - Different algorithms and data structures
- Customize snippets
  - e.g. sentence boundary, regex, ...
- Control search term highlighting
  - Bold, colors, etc.

suggest/

autocomplete & spellcheck



Suggest



berlin bu

berlin bu**s**

berlin bu**s map**

berlin bu**zzwords**

berlin bu**s tour**

Suggest: Autocomplete

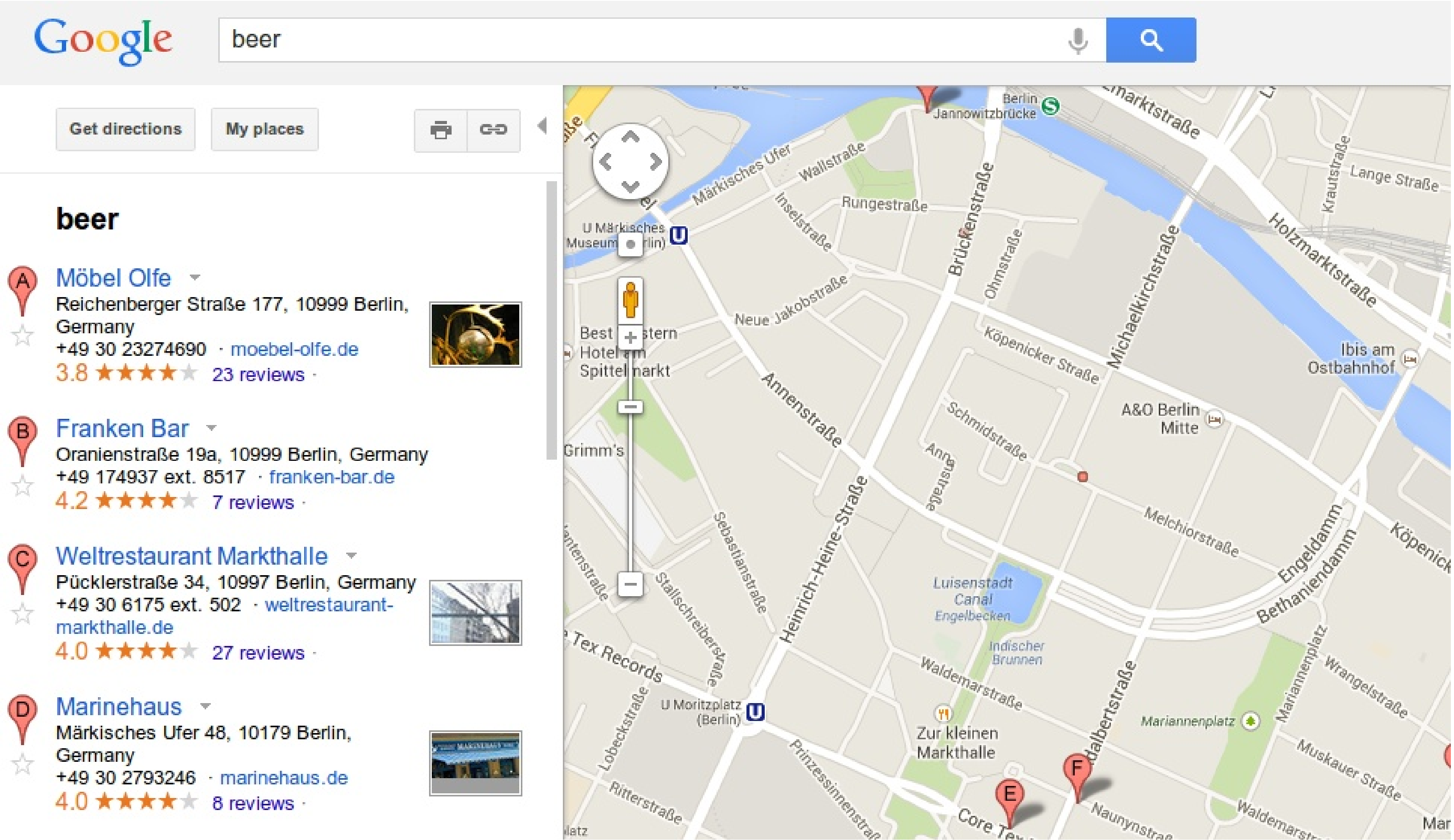
- Customize with analysis chain
    - Ex: accent-insensitive
- Typo correction / edit distance
- Infix suggestions
- Attach “payloads”
- Customize ranking with expressions

Suggest: Did you mean

- User ignored your fantastic autocomplete.
- Don't just return 0 results!
- N-gram or edit distance.
- Word spacing errors (“berlin buzz words”)

expressions/

custom ranking



beer

A **Möbel Olfe**
Reichenberger Straße 177, 10999 Berlin, Germany
+49 30 23274690 · moebel-olfe.de
3.8 ★★★★☆ 23 reviews ·

B **Franken Bar**
Oranienstraße 19a, 10999 Berlin, Germany
+49 174937 ext. 8517 · franken-bar.de
4.2 ★★★★☆ 7 reviews ·

C **Weltrestaurant Markthalle**
Pücklerstraße 34, 10997 Berlin, Germany
+49 30 6175 ext. 502 · weltrestaurant-markthalle.de
4.0 ★★★★☆ 27 reviews ·

D **Marinehaus**
Märkisches Ufer 48, 10179 Berlin, Germany
+49 30 2793246 · marinehaus.de
4.0 ★★★★☆ 8 reviews ·

Expressions

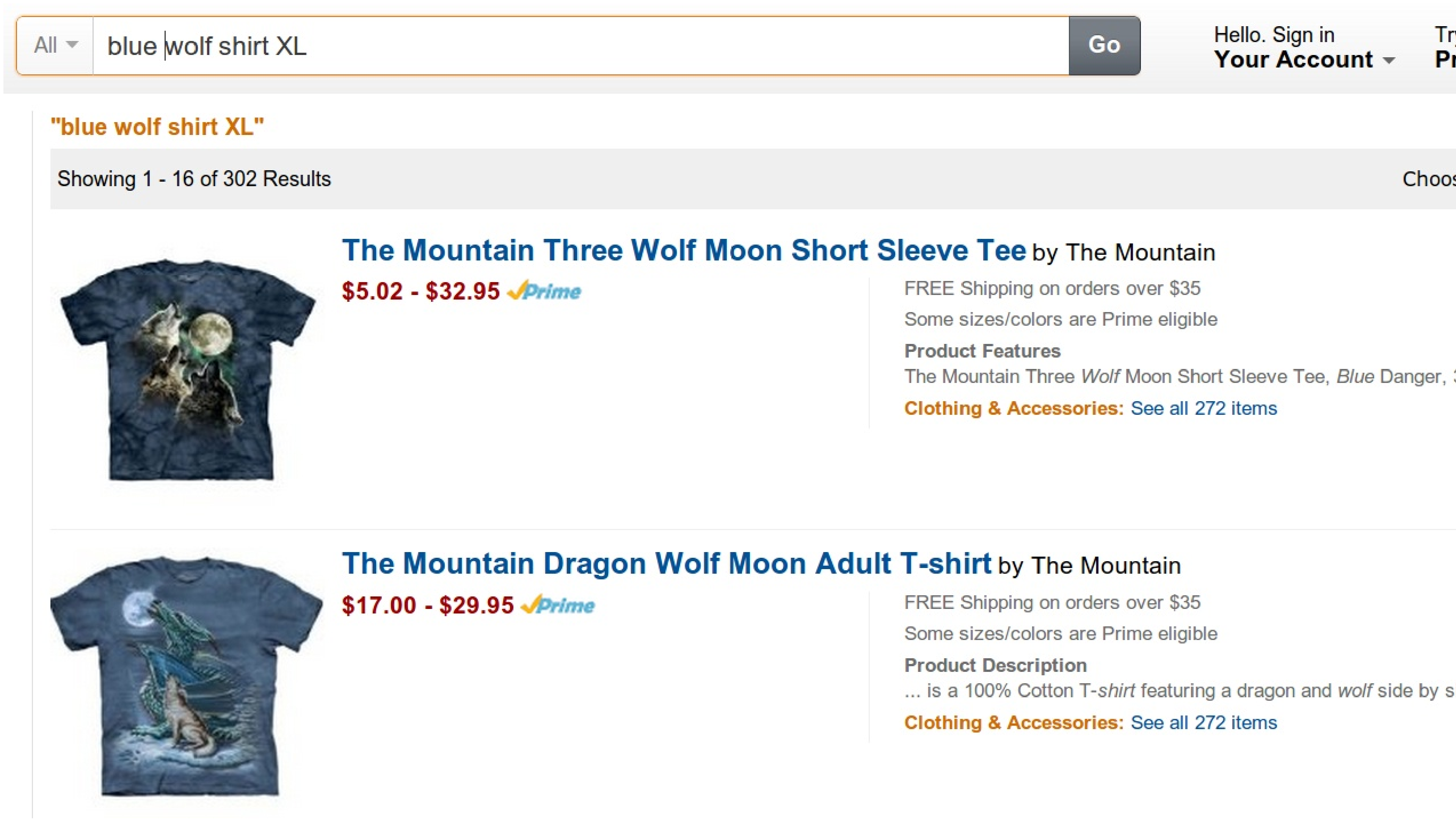
- Ranking function in Javascript subset.
  - Compiled to native bytecode.
- Combine different features:
  - Lucene's text score
  - numeric fields
  - distance functions

Expressions (cont)

- Sort by geographic distance:
  - “haversin(39.75, -104.99, lat, lon)”
- Incorporate ratings and popularity:
  - “avgRating * ln(numRatings) / distance”
- Pluggable functions and variable bindings

join/

nested documents



blue wolf shirt XL

"blue wolf shirt XL"

Showing 1 - 16 of 302 Results

Choos

The Mountain Three Wolf Moon Short Sleeve Tee by The Mountain

$5.02 - $32.95 Prime

FREE Shipping on orders over $35

Some sizes/colors are Prime eligible

Product Features

The Mountain Three *Wolf* Moon Short Sleeve Tee, *Blue* Danger,

Clothing & Accessories: See all 272 items

The Mountain Dragon Wolf Moon Adult T-shirt by The Mountain

$17.00 - $29.95 Prime

FREE Shipping on orders over $35

Some sizes/colors are Prime eligible

Product Description

... is a 100% Cotton T-*shirt* featuring a dragon and *wolf* side by s

Clothing & Accessories: See all 272 items

Nested documents

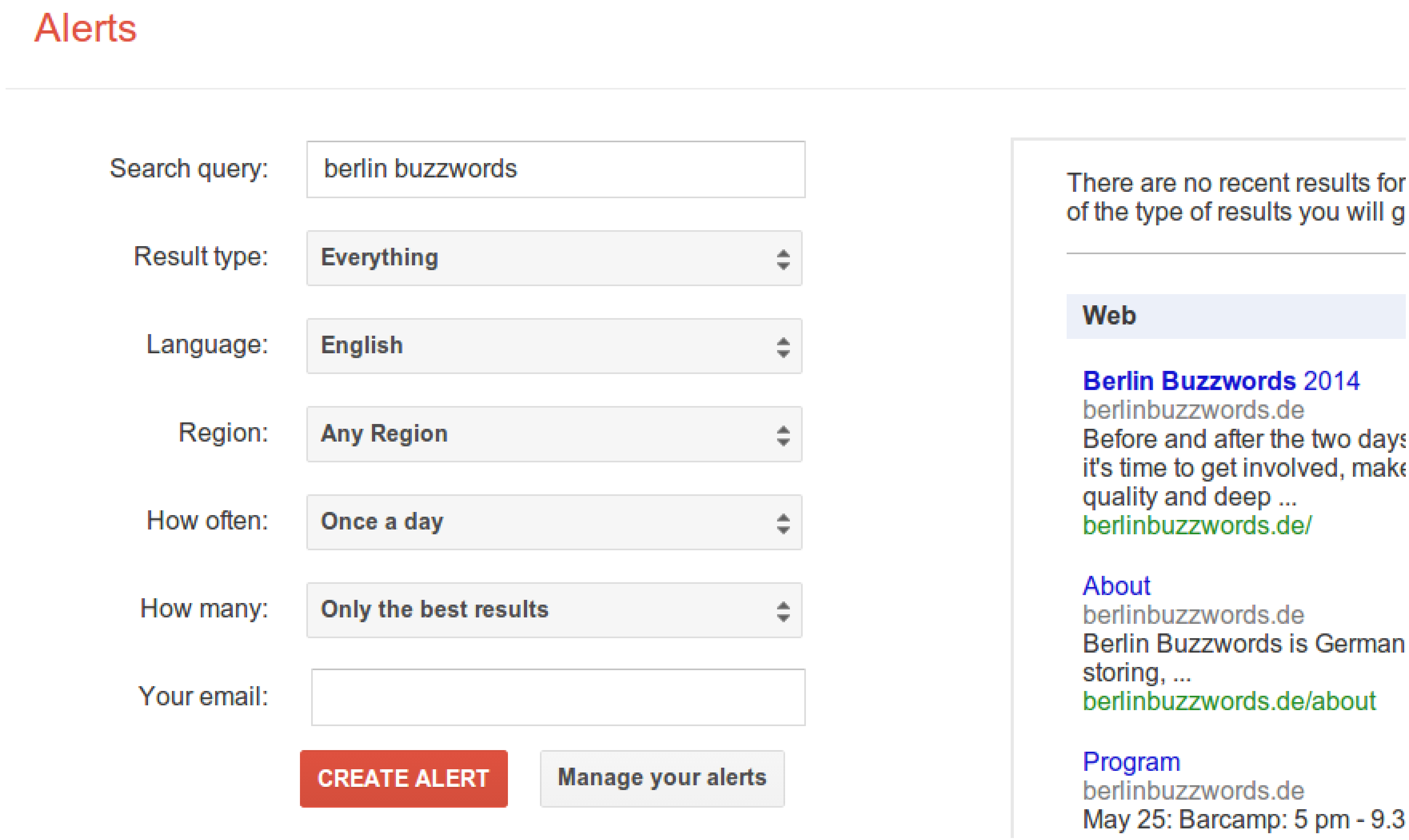
```
{
   “name”: “Wolf Shirt”,
   “sku” : [
        {  “color” : “blue”,  “size”   : “XL” },
        {  “color” : “red”,    “size”   : “S” }
    ]
}
```

Nested documents (cont)

- Type of “join” from child to parent.
- More intuitive for nested structures.
- Alternative to denormalization.

memory/

prospective search



Alerts

Search query: berlin buzzwords

Result type: Everything

Language: English

Region: Any Region

How often: Once a day

How many: Only the best results

Your email:

There are no recent results for
of the type of results you will g

Web

Berlin Buzzwords 2014
berlinbuzzwords.de
Before and after the two days
it's time to get involved, make
quality and deep ...
berlinbuzzwords.de/

About
berlinbuzzwords.de
Berlin Buzzwords is German
storing, ...
berlinbuzzwords.de/about

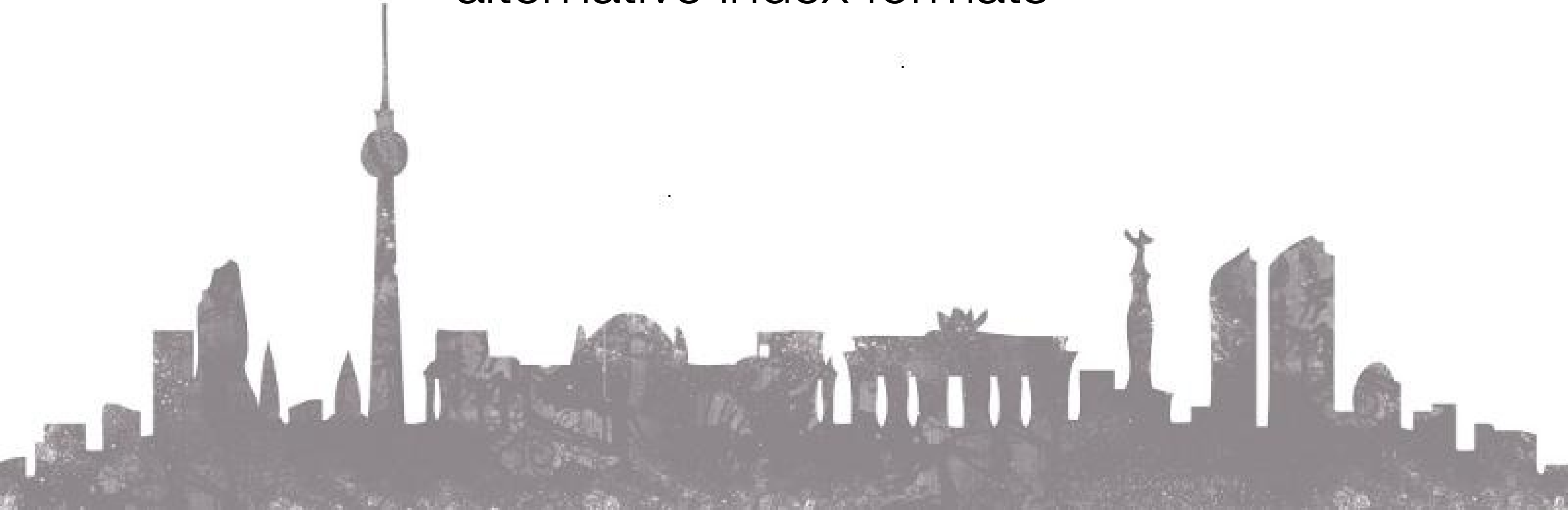
Program
berlinbuzzwords.de
May 25: Barcamp: 5 pm - 9.3

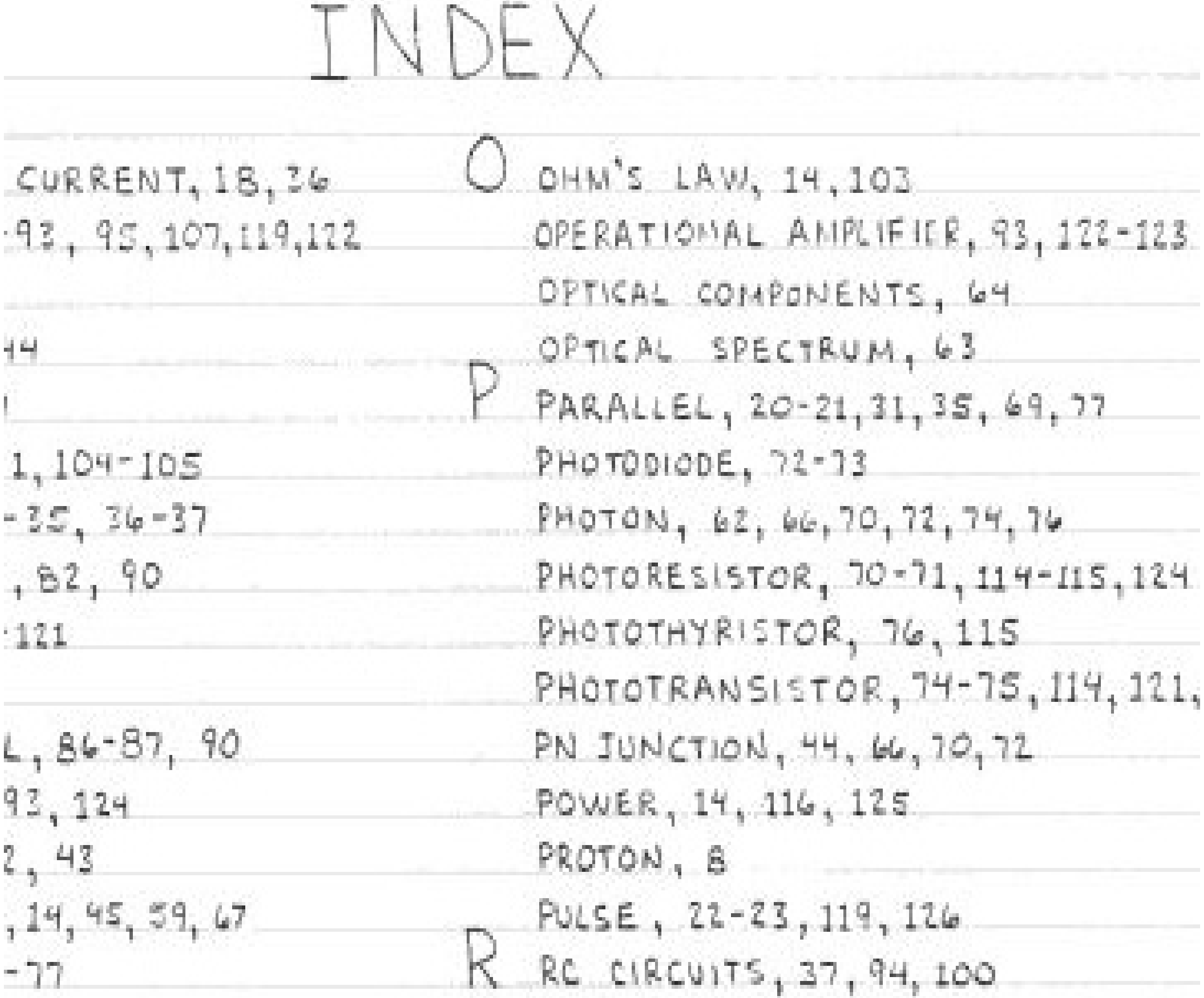
Memory: turned upside down

- When document comes in:
  1. Create single document index in RAM.
  2. Run all the queries (this is very fast)
  3. Throw away the index.

codecs/

alternative index formats





INDEX

CURRENT, 18, 36
-93, 95, 107, 119, 122
44
1
1, 104-105
-35, 36-37
, 82, 90
-121
L, 86-87, 90
93, 124
2, 43
, 14, 45, 59, 67
-77

O
OHM'S LAW, 14, 103
OPERATIONAL AMPLIFIER, 93, 122-123
OPTICAL COMPONENTS, 64
OPTICAL SPECTRUM, 63

P
PARALLEL, 20-21, 31, 35, 69, 77
PHOTODIODE, 72-73
PHOTON, 62, 66, 70, 72, 74, 76
PHOTORESISTOR, 70-71, 114-115, 124
PHOTOTHYRISTOR, 76, 115
PHOTOTRANSISTOR, 74-75, 114, 121,
PN JUNCTION, 44, 66, 70, 72
POWER, 14, 116, 125
PROTON, 8
PULSE, 22-23, 119, 126

R
RC CIRCUITS, 37, 94, 100

Index compression

- O
  - Ohm: 5,6
  - Optical: 1,4
  - ...
- P
  - Parallel: 3
  - Photon: 2,6
  - ...

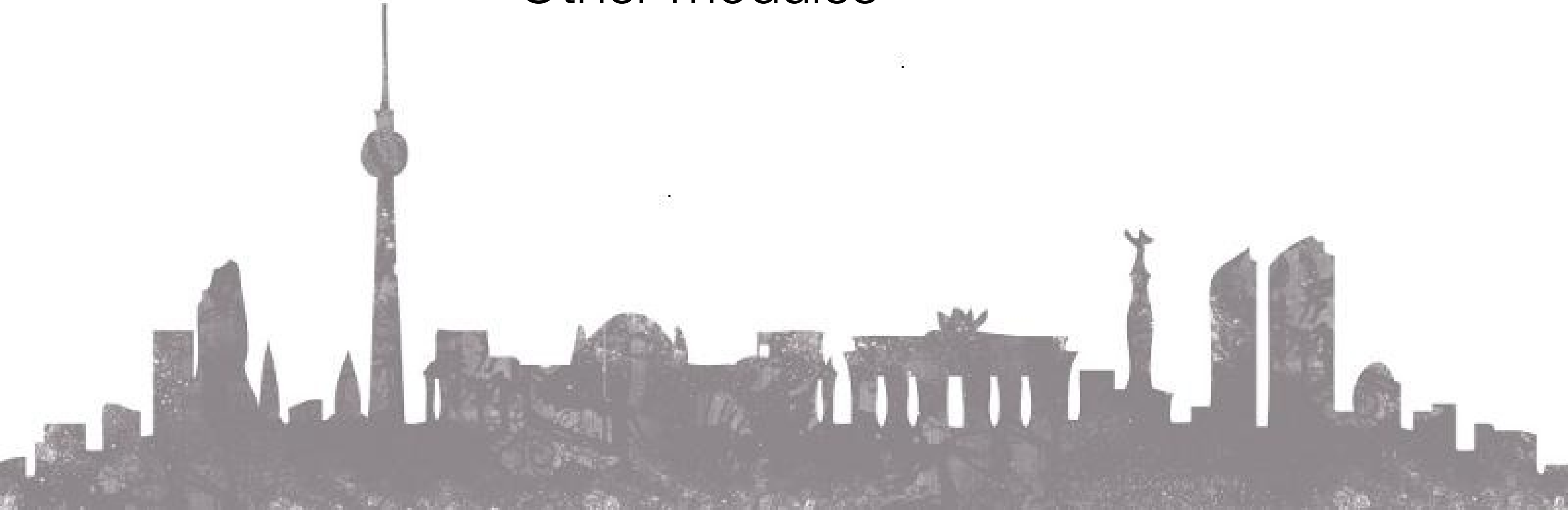
Index compression

- O
  - hm: 5,6
  - ptical: 1,4
  - ...
- P
  - arallel: 3
  - hoton: 2,6
  - ...

Codecs

- Index formats for different use cases
- Different compression options
  - e.g. slower but smaller
- Different datastructures for different data
  - e.g. terms in a Trie

Other modules



Other modules

- **benchmark/**
  - Measure performance and relevance.
- **classification**/
  - Classify documents based on the index.
- **demo/**
  - Simple example code for Lucene.

Other modules (cont)

- **facet/**
  - Navigational Search
- **grouping**/
  - Related search results
- **misc/**
  - Index tools (split, sort, examine, etc)

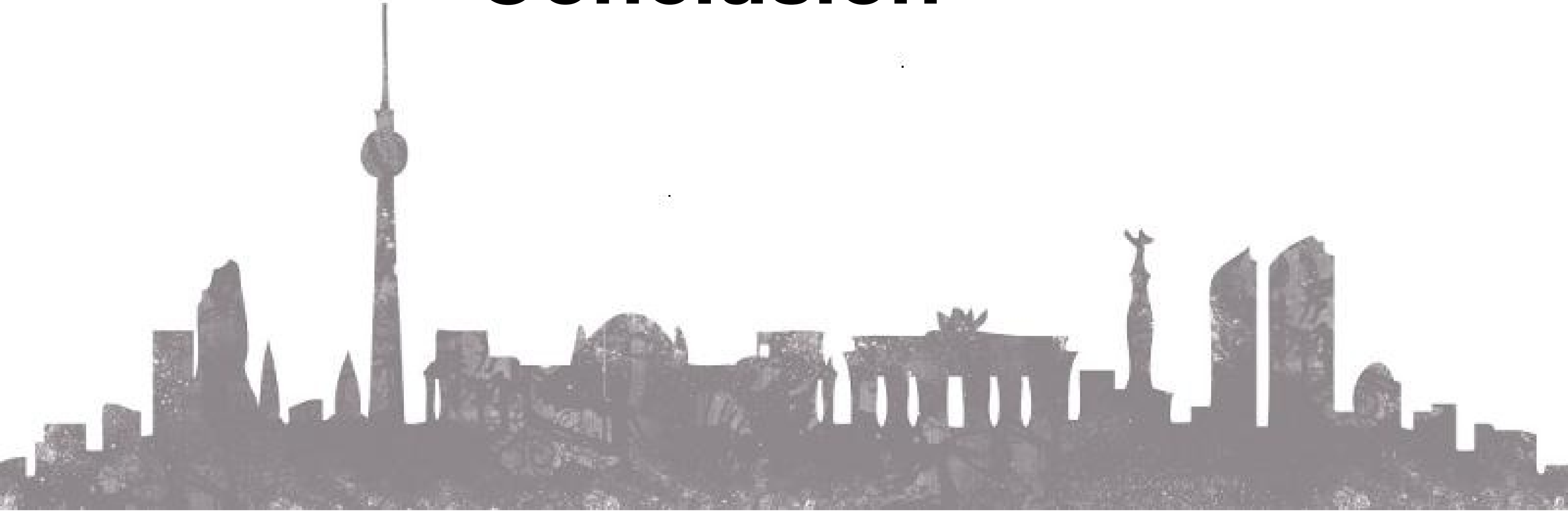
Other modules (cont)

- **queries**/
  - Additional queries and filters.
- **replicator/**
  - Synchronize index files across machines.

Other modules (cont)

- **sandbox/**
    - Experimental or “sandy” code.
- **spatial**/
    - Advanced geospatial support (e.g. shapes).
- **test-framework/**
    - Test harnesses for Lucene extensions.

Conclusion



Conclusion

- Lucene has components for many use-cases.
- We briefly touched on some of these.
- For a deeper look...

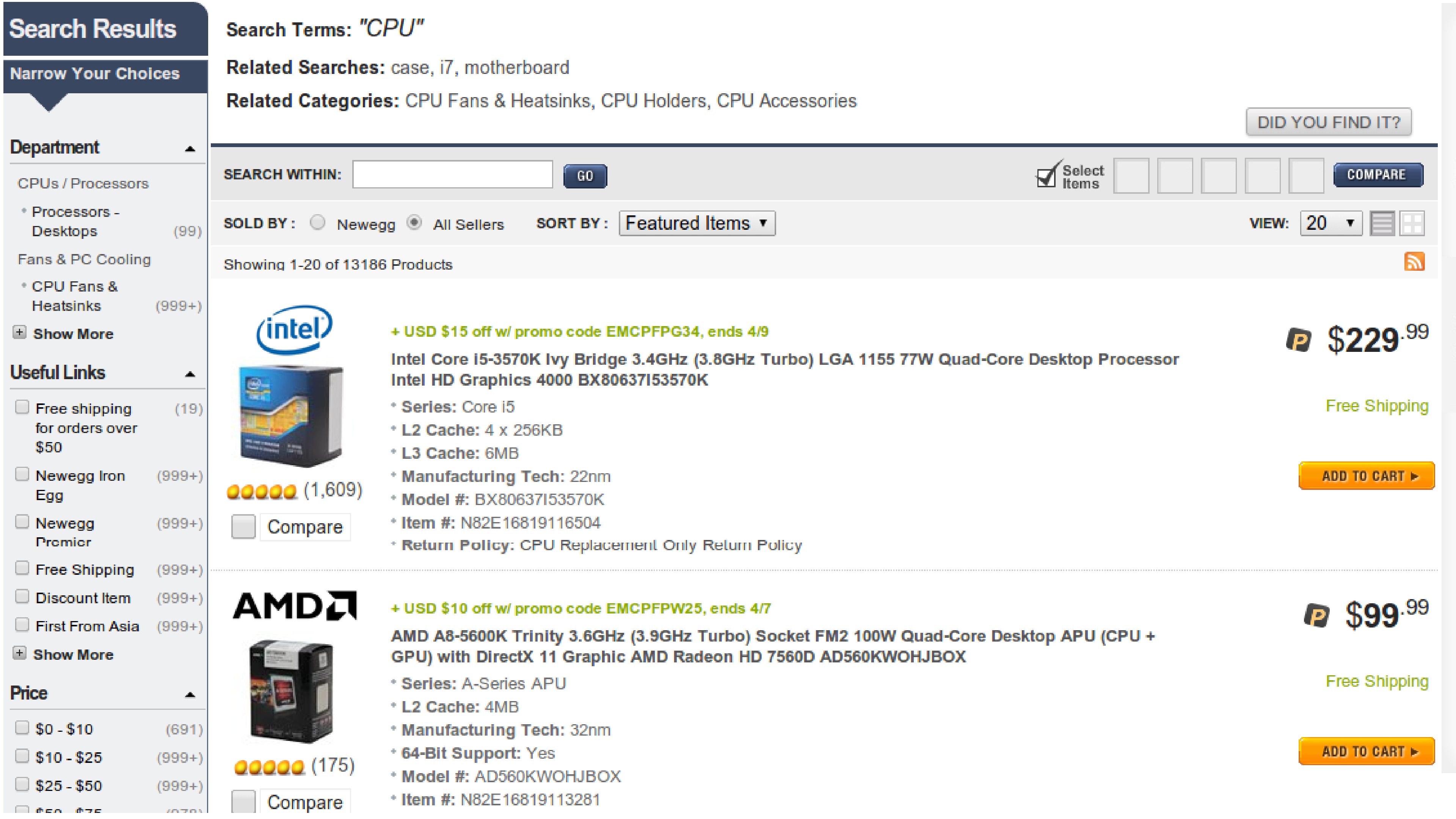
Resources

- Online documentation:
  - http://lucene.apache.org
- Lucene in Action
  - Still a fantastic introduction.

Q & A

facet/

navigational search



Search Results

Narrow Your Choices

Department

CPUs / Processors

- Processors - Desktops (99)

Fans & PC Cooling

- CPU Fans & Heatsinks (999+)

⊞ Show More

Useful Links

□ Free shipping for orders over $50	(19)
□ Newegg Iron Egg	(999+)
□ Newegg Premier	(999+)
□ Free Shipping	(999+)
□ Discount Item	(999+)
□ First From Asia	(999+)

⊞ Show More

Price

□ $0 - $10	(691)
□ $10 - $25	(999+)
□ $25 - $50	(999+)
□ $50 - $75	(978)

Search Terms: *"CPU"*

Related Searches: case, i7, motherboard

Related Categories: CPU Fans & Heatsinks, CPU Holders, CPU Accessories

DID YOU FIND IT?

SEARCH WITHIN: [ ] GO ☑ Select Items COMPARE

SOLD BY : ○ Newegg ● All Sellers SORT BY : Featured Items ▾ VIEW: 20 ▾

Showing 1-20 of 13186 Products

+ USD $15 off w/ promo code EMCPFPG34, ends 4/9

Intel Core i5-3570K Ivy Bridge 3.4GHz (3.8GHz Turbo) LGA 1155 77W Quad-Core Desktop Processor Intel HD Graphics 4000 BX80637I53570K

- **Series:** Core i5
- **L2 Cache:** 4 x 256KB
- **L3 Cache:** 6MB
- **Manufacturing Tech:** 22nm
- **Model #:** BX80637I53570K
- **Item #:** N82E16819116504
- **Return Policy:** CPU Replacement Only Return Policy

●●●●● (1,609) □ Compare

$229.99

Free Shipping

ADD TO CART ▸

+ USD $10 off w/ promo code EMCPFPW25, ends 4/7

AMD A8-5600K Trinity 3.6GHz (3.9GHz Turbo) Socket FM2 100W Quad-Core Desktop APU (CPU + GPU) with DirectX 11 Graphic AMD Radeon HD 7560D AD560KWOHJBOX

- **Series:** A-Series APU
- **L2 Cache:** 4MB
- **Manufacturing Tech:** 32nm
- **64-Bit Support:** Yes
- **Model #:** AD560KWOHJBOX
- **Item #:** N82E16819113281

●●●●● (175) □ Compare

$99.99

Free Shipping

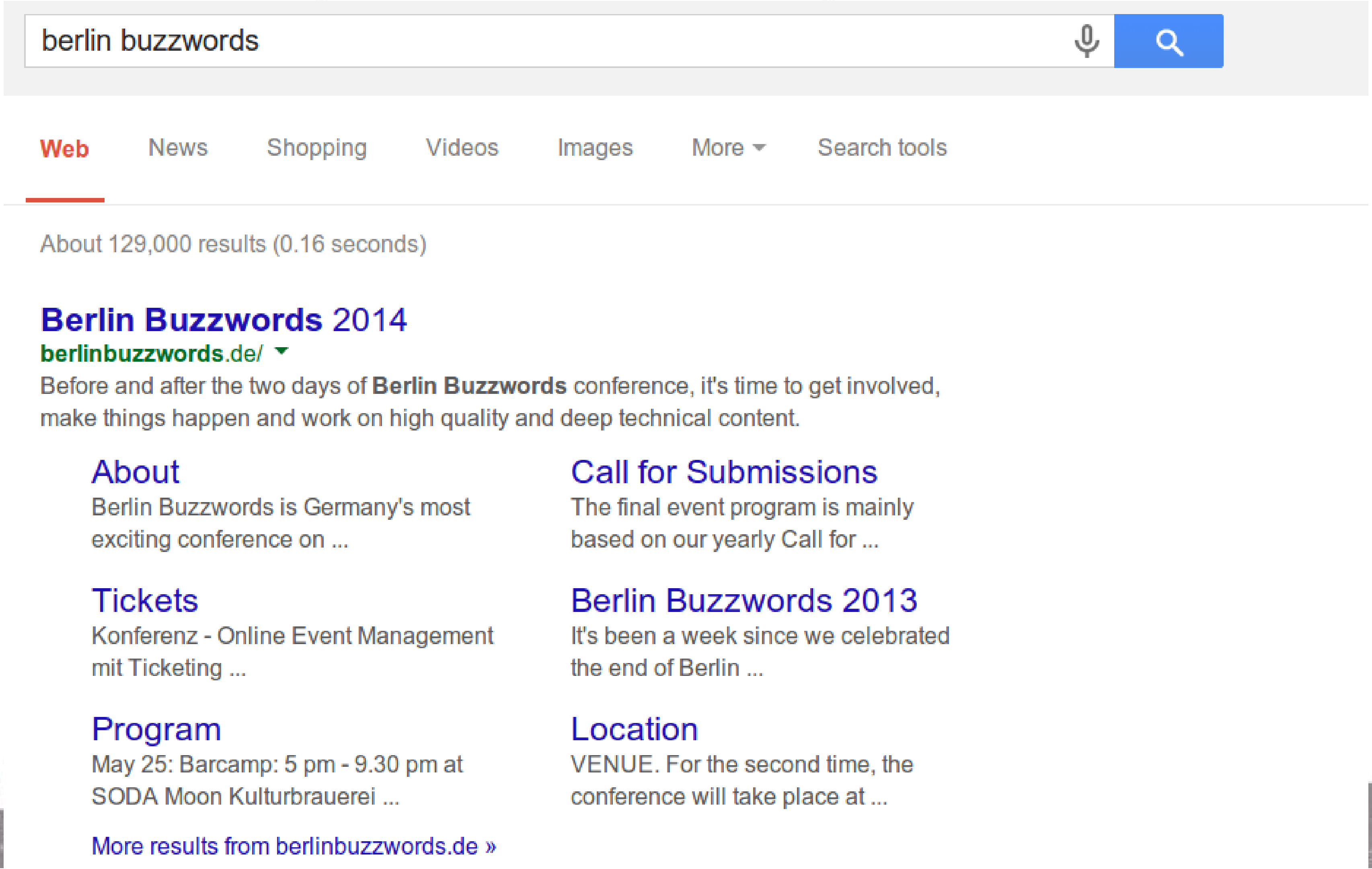
ADD TO CART ▸

Facets

- Flat hierarchy or taxonomy
- Bucket by Terms:
  - e.g. category Electronics, Kitchen, ...
- Bucket by Numeric ranges:
  - e.g. price $10-$20, $20-$30, ...
- Bucket by Expression ranges:
  - e.g. distance 10-20km, 20-30km, ...

grouping/

related search results



About 129,000 results (0.16 seconds)

Berlin Buzzwords 2014
berlinbuzzwords.de/
Before and after the two days of **Berlin Buzzwords** conference, it's time to get involved, make things happen and work on high quality and deep technical content.

About
Berlin Buzzwords is Germany's most exciting conference on ...

Call for Submissions
The final event program is mainly based on our yearly Call for ...

Tickets
Konferenz - Online Event Management mit Ticketing ...

Berlin Buzzwords 2013
It's been a week since we celebrated the end of Berlin ...

Program
May 25: Barcamp: 5 pm - 9.30 pm at SODA Moon Kulturbrauerei ...

Location
VENUE. For the second time, the conference will take place at ...

More results from berlinbuzzwords.de »

Grouping

- Organizes hits into groups (e.g. website)
- Group by field
- Group by function, expression
- Customize how groups are ranked
  - sum, avg, etc