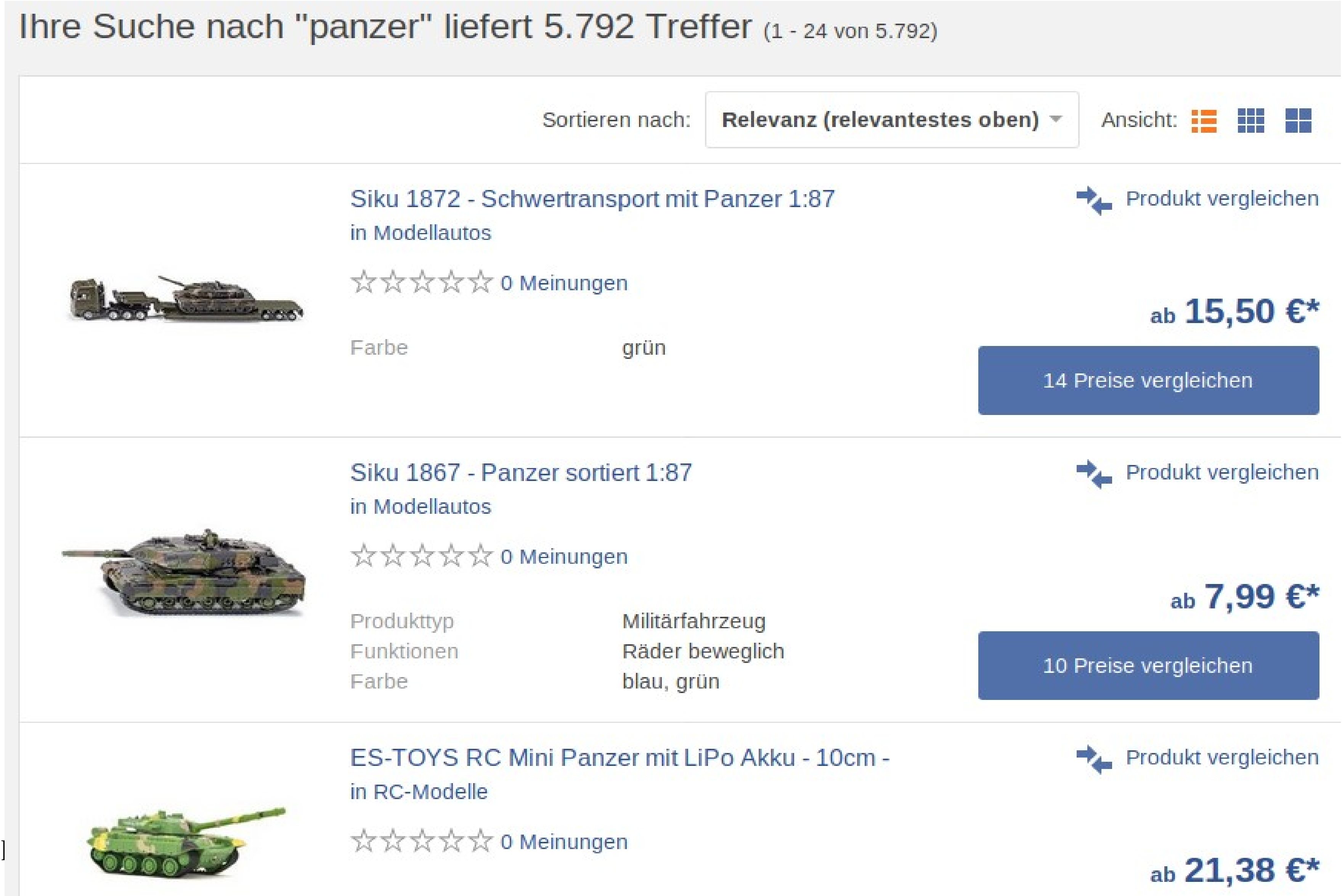
War Stories from the billiger.de Trenches



Ihre Suche nach "panzer" liefert 5.792 Treffer (1 - 24 von 5.792)

Sortieren nach: Relevanz (relevantestes oben)

Siku 1872 - Schwertransport mit Panzer 1:87
in Modellautos

☆☆☆☆☆ 0 Meinungen

Farbe	grün

Siku 1867 - Panzer sortiert 1:87
in Modellautos

☆☆☆☆☆ 0 Meinungen

Produkttyp	Militärfahrzeug
Funktionen	Räder beweglich
Farbe	blau, grün

ES-TOYS RC Mini Panzer mit LiPo Akku - 10cm -
in RC-Modelle

☆☆☆☆☆ 0 Meinungen

Search War Stories

- Taking Shops Offline
  or: 500k Docs: Now You See Them – Now You Don't
- Stemming Pitfalls
- Stopwords Traps
- Using the Users: Implicit Feedback
- Painful (but proper) Quality Control
- Of Products and Offers

Architectural Overview

- Portal billiger.de & Syndication API
- Separate Loadbalancers for Portal, API
- Search Service (Python, Pyramid) talks to localhost
- localhost: SOLRCloud (→ see prev. talk)
- Indices: main index, brands, categories, shops
- Identical indices for portal, API → can reassign servers
- Updater for each index, fed JSON files (deltas for main index, full for the others)

Taking Shops Offline

- Problem: Shops can go “offline”, i.e. *all* their offers must disappear from the site
- Deleting them is easy:
  `{“delete”: {“filter_shops”: 123}}`
- Taking them back online can be expensive & slow for large shops: O(100k+) offers
- Can we avoid deleting/re-indexing them?
- Cannot use one index per shop (because IDF)

Taking Shops Offline: Solr

- Sounds like an SQL JOIN (and we have a shop ID field in the index) → Solr *Cross Core Join*
- Shop index w/ shop IDs JOINed against Main index <field name=”filter_shops”> field:

```
<requestHandler name="/search" …
...
<lst name="appends">

       <str name="fq">{!join fromIndex=shops_online from=id
to=filter_shops}*:*</str>

     </lst>
</requestHandler>
```

- Allows for taking individual shops offline for a partner

Taking Shops Offline: SolrCloud

- Cross Core Join does not work in cloud mode!

- New way:
  1. get all shop Ids from shops_online index (*:* query)
  2. huuuuge terms filter with 2500+ terms:

```
fq=+{!terms f=filter_shops}1,2,3,...
```

Not so elegant, but performance is similar

Stemming

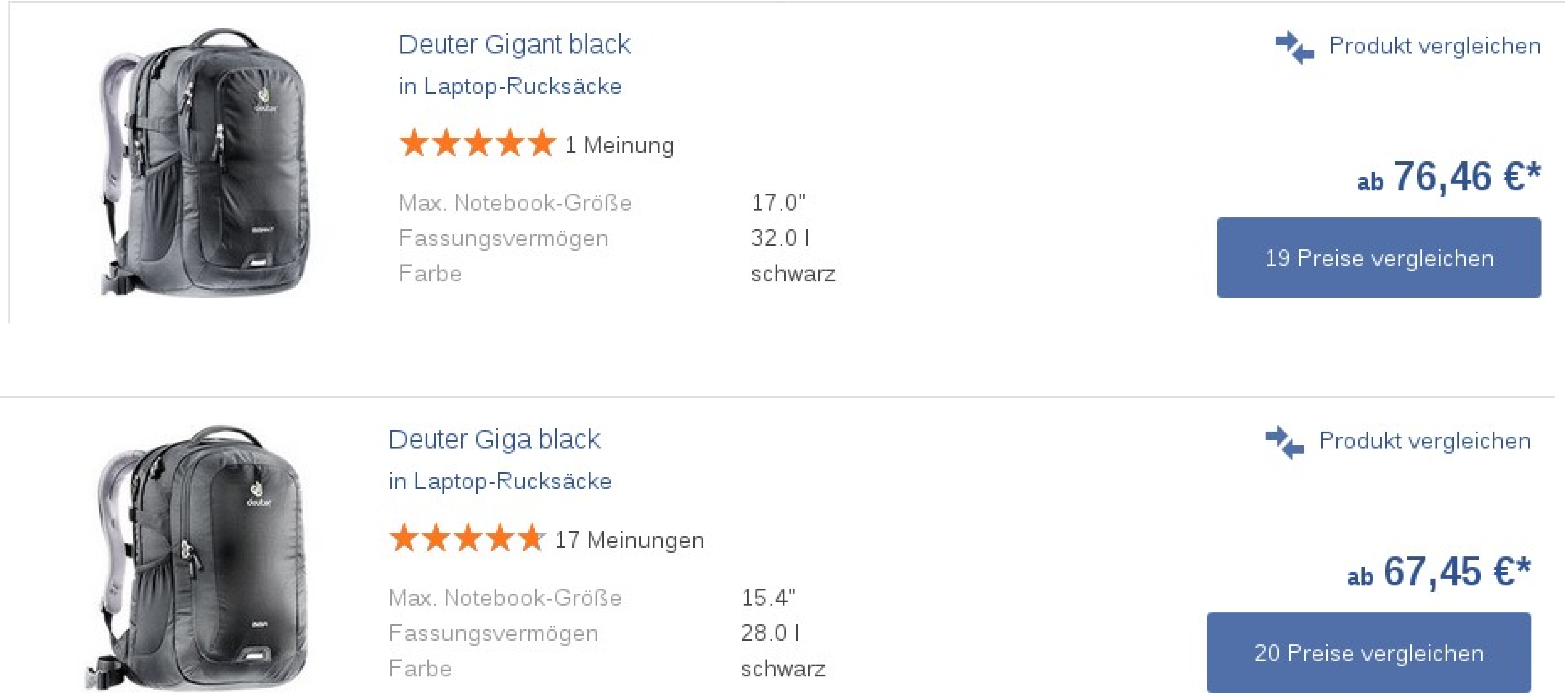
- What is Stemming:

  Reducing words to their base form

  coins → coin, had → have
- Why we do it? **Recall problems**: (singular v. plural mostly)
  “T-Shirts” → “T-Shirt”
  “Schuhe” → “Schuh”
  “Hemden” → “Hemd”
- German stemming: *really* nasty! 8 forms of plural, strong v. weak flexion...
  gehabt → haben, Häuser → Haus,
  => Gehäuse → Haus?
- Four (five) German stemmers available:
  GermanStemFilter, GermanLightStemFilter, GermanMinimalStemFilter, SnowballPorterFilter(“German”, “German2”)

Stemming

- Danger of Overstemming!



Deuter Gigant black

in Laptop-Rucksäcke

★★★★★ 1 Meinung

Max. Notebook-Größe	17.0"
Fassungsvermögen	32.0 l
Farbe	schwarz

Deuter Giga black

in Laptop-Rucksäcke

★★★★★ 17 Meinungen

Max. Notebook-Größe	15.4"
Fassungsvermögen	28.0 l
Farbe	schwarz

Stemming

ACTIVATES STEMMING TO IMPROVE RECALL
FUCKS UP PRECISION
imgflip.com

Stemming

- Endless room for experimentation...
- “This website no verbs”
- Minimal stemming only (GermanMinimalStemFilter)
- Long list of stemming exclusions (KeywordMarkerFilter + protwords_de.txt)
- Reactive: find & add new exclusion, must re-index
- Understemming → synonyms

`https://wiki.apache.org/solr/LanguageAnalysis`

Stopwords

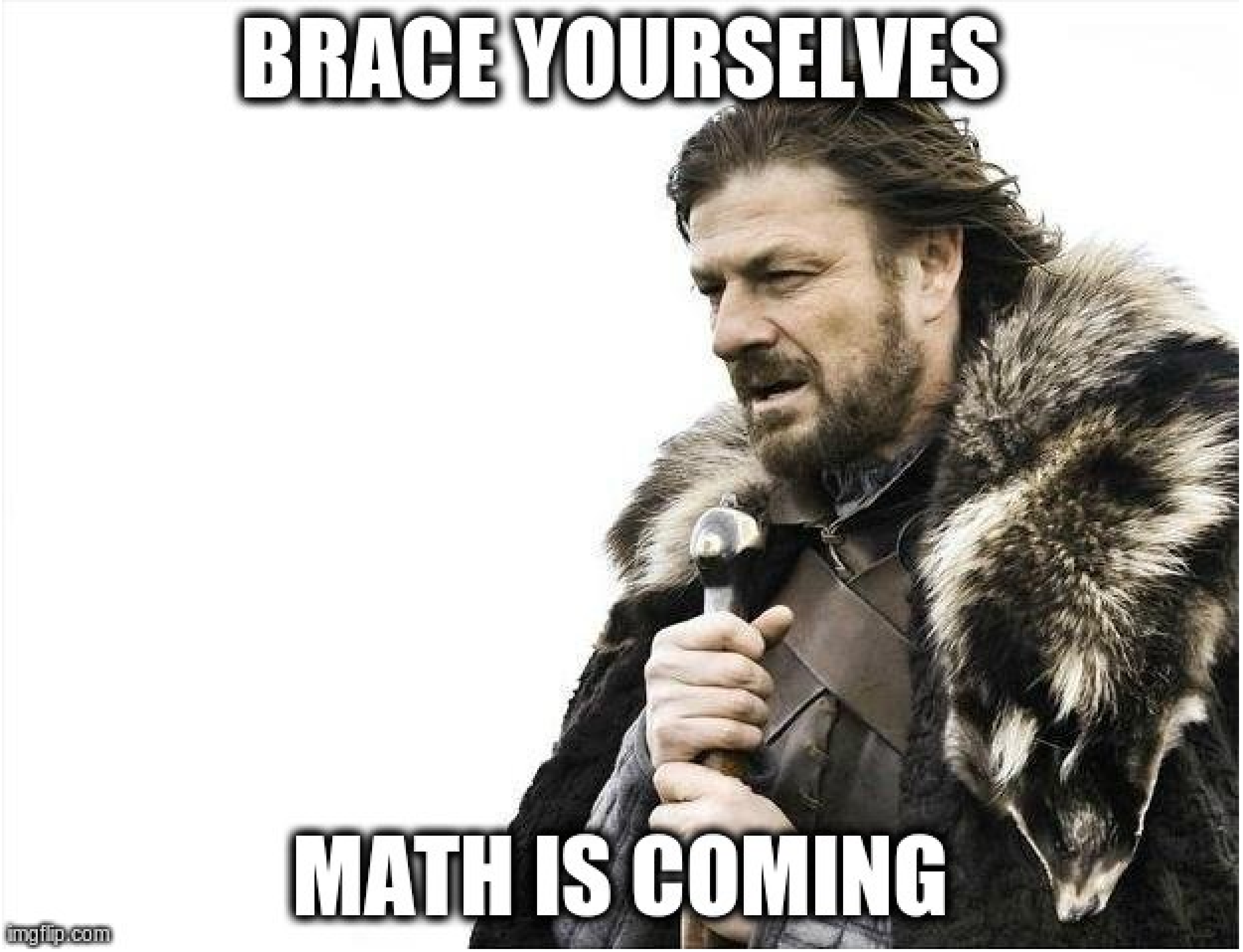
- Stopwords are words we strip from both docs and queries
- Smaller index
- More precise document length norm, good when users rarely include stopwords
- Controls recall on fuzzy queries:
  “asdfghjk is all qwertzu”
- “to be or not to be” → empty query, no hits
- Stopwords stripped from index and query:

```
<filter class="solr.StopFilterFactory" ignoreCase="true"
        words="lang/stopwords_de.txt" format="snowball" />
```

Intermission: **Thoughts on Scoring**

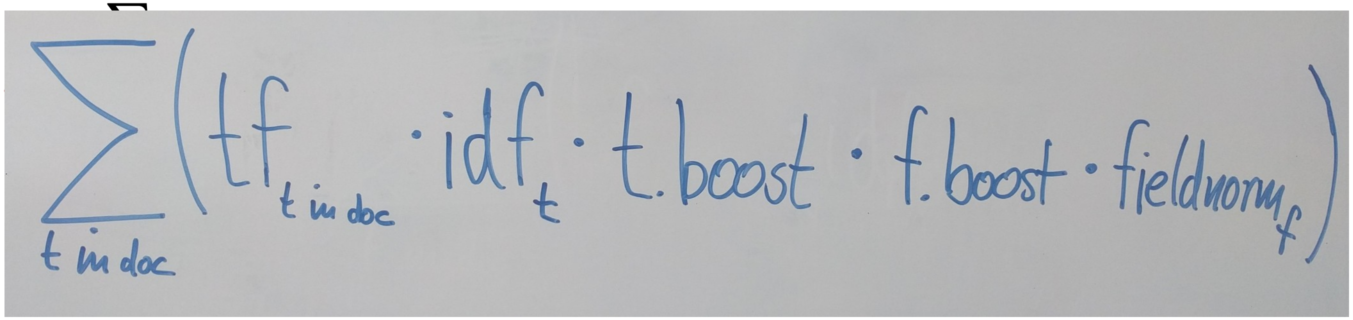
- If a term occurs often in a doc, it must be really important for that doc!
- If a term occurs in only a few docs, these docs must be really special (w.r.t. this term)
- Corollary: If a term occurs in virtually every doc, it doesn’t tell us much.
- Some fields are inherently more important than others (knowledge of doc structure).
- Some terms are inherently more important than others (knowledge/semantics of terms)
- One term in a long doc is less significant than one term in a short doc Case in point: a doc consisting of a single term.

Scoring



Lucene Scoring

- Score proportional to



$$\sum_{t \text{ in doc}} \left(tf_{t \text{ in doc}} \cdot idf_{t} \cdot t.boost \cdot f.boost \cdot fieldnorm_{f} \right)$$

- Term frequency: how often a term occurs in doc terms that occur often in the doc are more important
- Inverse doc frequency: in how many docs the term occurs

Incorporating User Feedback

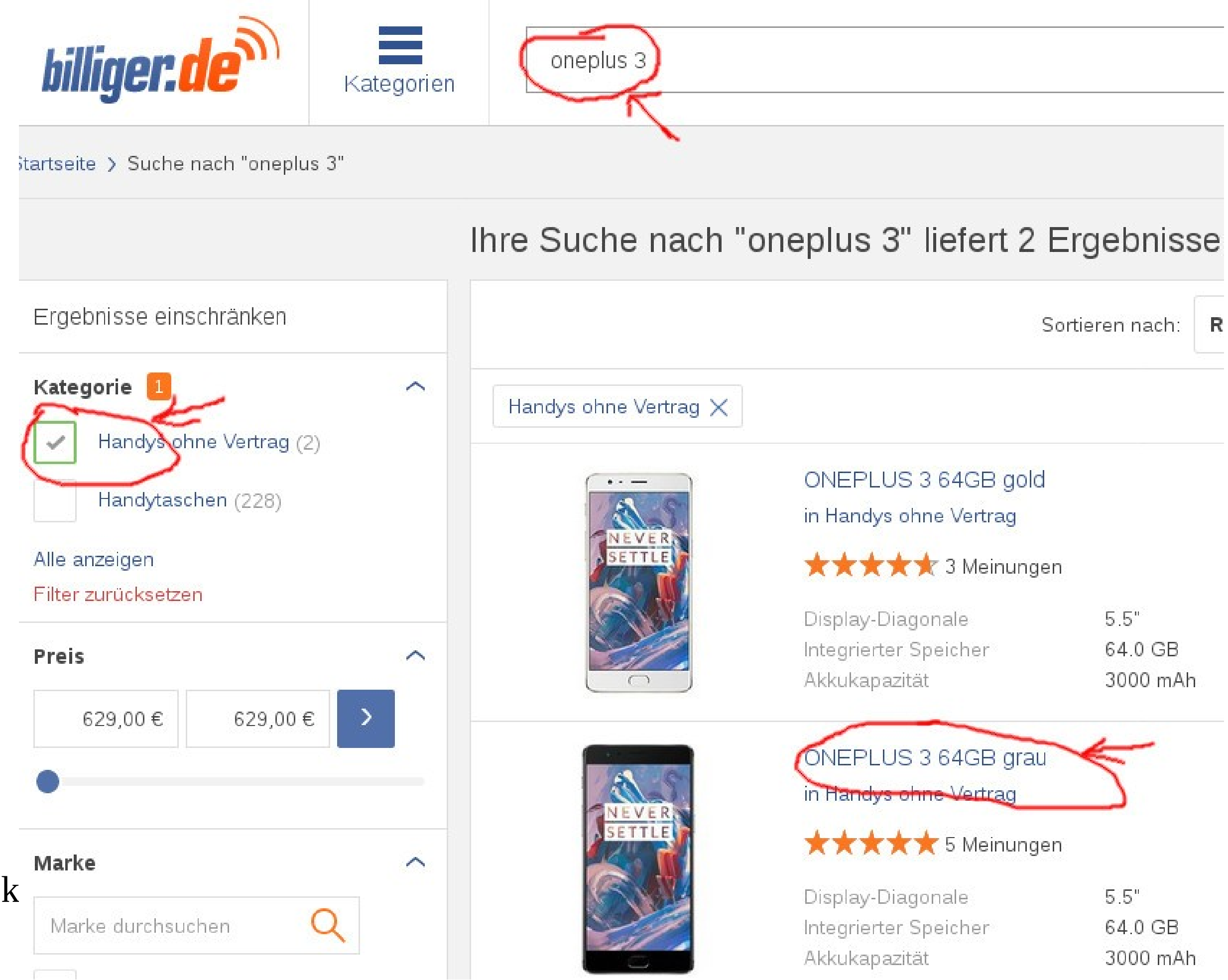
- Explicit user feedback:

$$\sum_{¿t \in q>¿<(tf*idf^2)>¿¿} ¿$$

- Implicit user feedback? ...

Incorporating Implicit User Feedback

- “Implicit Feedback” = user clicking on results (products, offers, categories & other filters)



billiger.de

Kategorien

oneplus 3

Startseite › Suche nach "oneplus 3"

Ihre Suche nach "oneplus 3" liefert 2 Ergebnisse

Ergebnisse einschränken

Kategorie 1

☑ Handys ohne Vertrag (2)

□ Handytaschen (228)

Alle anzeigen

Filter zurücksetzen

Preis

629,00 € 629,00 €

Marke

Marke durchsuchen

Sortieren nach: R

Handys ohne Vertrag ✕

ONEPLUS 3 64GB gold
in Handys ohne Vertrag
★★★★★ 3 Meinungen

Display-Diagonale	5.5"
Integrierter Speicher	64.0 GB
Akkukapazität	3000 mAh

ONEPLUS 3 64GB grau
in Handys ohne Vertrag
★★★★★ 5 Meinungen

Display-Diagonale	5.5"
Integrierter Speicher	64.0 GB
Akkukapazität	3000 mAh

Patrick

Incorporating Implicit User Feedback

- Harvested from frontend logs (docs)...

```
{"ts": 1496212351, "query_id":
"8118d1c33b1141268e5fa0a4e65392f3", "query_string":
"emser halspastillen", "results": ["1:82706493",
"1:82262858", "1:515286705", "1:82268643", "1:81028376",
"1:132832088", "1:82349225", "1:82271347", "1:82281586",
"1:324484307", "1:324485523", "1:82358049", "1:82247198",
"1:82236122", "1:286990644", "1:745607317",
"1:324485504", "1:553963013", "1:746122391",
"1:82414054", "1:887951887", "1:82465108", "1:745762420",
"1:887951018"], "offset": 0}

{"ts": 1496212362, "query_id":
"8118d1c33b1141268e5fa0a4e65392f3", "clicked_offset": 0}
```

Incorporating Implicit User Feedback

- ... and from search service logs (filters):

```
{"query":"sat receiver",
 "filters":{
   "type":[1,0], "f_2308":[124784], "f_18":[16463,23704],
   "categories":[106886]},
 "page_no":0,"page_size":24,
 "boosts":{"has_image":[[1],10.0]},
 "sort_mode":["score,desc","clickout_relevance,desc","id,desc"],
 "options":{
   "fuzzy":true,"facet_mode":"multiselect","client_tag":"search"
},
"_duration":0.208976984,"_total_hits":75}
```

Incorporating Implicit User Feedback

- Map/Reduce jobs to distill boosts from logs (using Nokia Disco)
- Web interface for product managers to tune boosts (for new docs)
- *Query-Local Term Boosts* (QLTB): qltb.xml contains the resulting boosts:

```
<query text="zelte">
  <term boost="1.9" field="filter_brands" value="3778545"/>
     <!--Mc Kinley-->
  <term boost="1.1" field="filter_brands" value="1027815"/>
     <!--High Peak-->
  <term boost="300.0" field="filter_categories"
     value="103377"/> <!--Zelte-->
</query>
<query text="zerkleinerer">
  <term boost="1.1" field="filter_brands" value="7621"/>
    <!--Moulinex-->
</query>
<query text="zimmerantenne">
  <term boost="100.0" field="filter_categories"
    value="103145"/> <!--DVB-T2 Antennen-->
</query>
```

Incorporating Implicit User Feedback

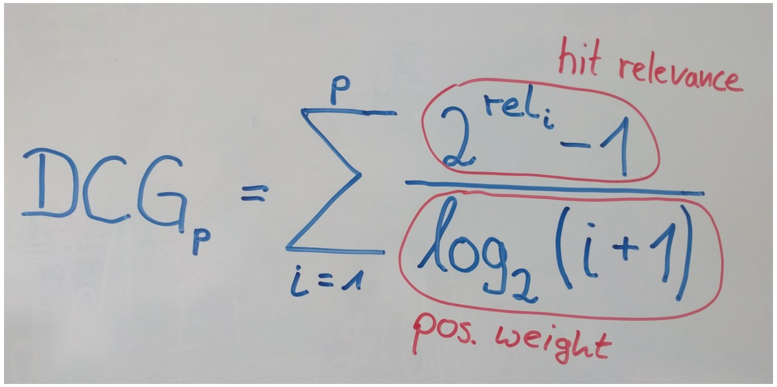
- Open Source: QltbComponent for Solr https://github.com/solute/qltb
- SolrCloud: re-written in Python b/c ZooKeeper
- Threshold: at least *n* clicks to be included in QLTB.xml
- ca. 7000 queries with boost terms → major quality improvement for more frequent queries

Intermission: **On Quality**

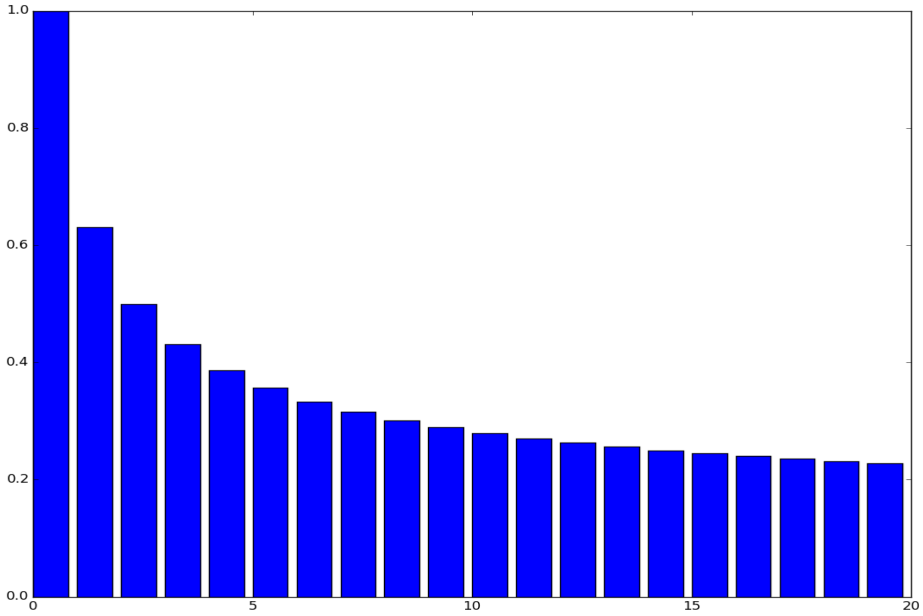
- How do we measure quality?
- Doc can be good, mediocre, or bad w.r.t. query
- Assume enough good docs for page 1
- Bad docs on pos 1 are *really* bad
- Bad docs on pos 24 are not as bad as pos 1
- Position of bad or mediocre docs matters

On Quality: nDCG

- Normalized Distributed Cumulative Gain nDCG



$$DCG_p = \sum_{i=1}^{p} \frac{\boxed{2^{rel_i} - 1}}{\boxed{\log_2 (i+1)}}$$



- Great: re-run queries, re-compute nDCG, get algorithm improvement factor! Fully automatic!

On Quality: nDCG

- Not so great: Need someone to rate all result docs on page 1 as good, bad, or mediocre w.r.t. query
- Changes in algo → new docs on first page.
- Even worse: at billiger.de, docs fluctuate wildly!
- If you can't automate, make it easy to do by hand
- New tool: automate just the process (running query sets against different searchers), and <u>support</u> human “visual diff” against baseline

Products and Offers

- How should we treat products v. offers w.r.t. search?
  Slight preference for products.
- “Proper” way:
  - Offer: use offer title
  - Product: use *product* title (manually edited!)
  - *Boosting*: `filter_type:product^2.0`
- Surprise: Recall problems ☹

Products and Offers, tf and idf

- Problem: User Typos.
  Solution: Shop Typos.
  - Offer: use offer title
  - Product: use product title (manually edited)
    + offer titles
  - “Boosting” via **tf**/idf: Deuter Gigant Black →

```
Deuter Deuter Gigant Black - Laptoprucksack Deuter Deuter Rucksack
Gigant, black, 47 x 35 x 27 cm, 32 Liter, 8042470000 Deuter Deuter
Daypack Gigant Rucksack mit Laptopfach 47 cm - black Deuter Deuter Gigant
Laptoprucksack schwarz Gr.  Deuter Deuter School/Uni Gigant
Laptoprucksack 47 cm - black Deuter Deuter Gigant black - Laptoprucksack
schwarz Deuter Deuter Gigant black Deuter Deuter Rucksack Gigant black
Deuter Deuter Daypack Gigant Rucksack 47 cm Laptopfach black Deuter
Deuter Rucksack Gigant black Deuter Deuter Gigamt Laptoprucksack Schwarz
Deuter deuter Rucksack "Gigant", 32 l Deuter Deuter: Tagesrucksack
Gigant, Schwarz, verfügbar in Größe 0 Deuter Deuter Rucksack Gigant Black
Deuter Deuter Bookpack Laptop-Rucksack Gigant Black Farbe 7000 black
Deuter Deuter Laptoprucksack Gigant 47 cm black Deuter Deuter GIGANT
Rucksack School & Daypack 17,3" black Deuter Deuter Giga
Damen/Herren Gigant
```

Products and Offers, tf and idf

- Drawbacks:
  - tf depends on number of offers in product...
  - Short, concise offer might beat product if a product offer does SEO (bloating the product, → doc length norm)
  - Sweet spot similarity: tf
- Sweet spot similarity: plateau of lengths that should all have a norm of 1.0
- As always: work in progress!

More Topics...

- Suggestions
- (Term-) Fuzzy Search
- Filters and Filter Alternatives
- EAN/ISBN/PZN Handling
- Model Identifier Handling
- Degrading Performance/Progressive Index Growth
- ...