

Gain speed and **precision**
with NLP in Solr





Father, Husband, Hamburg
Software Engineer @shopping24
Search Technology



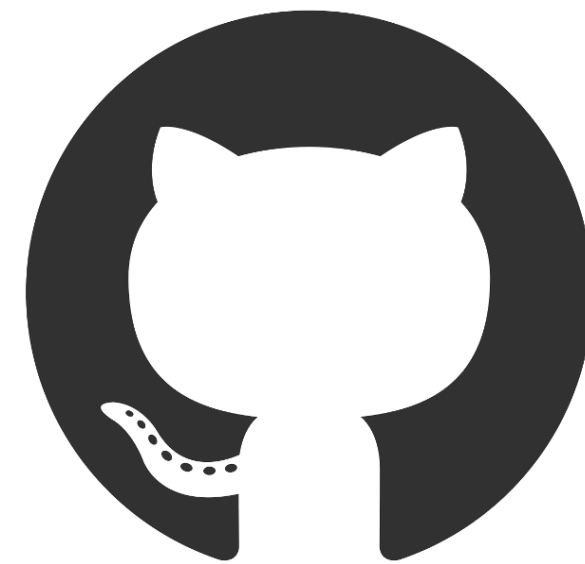
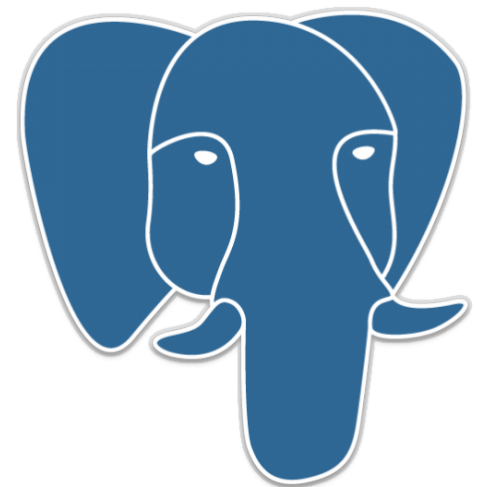
Hamburg

Apache

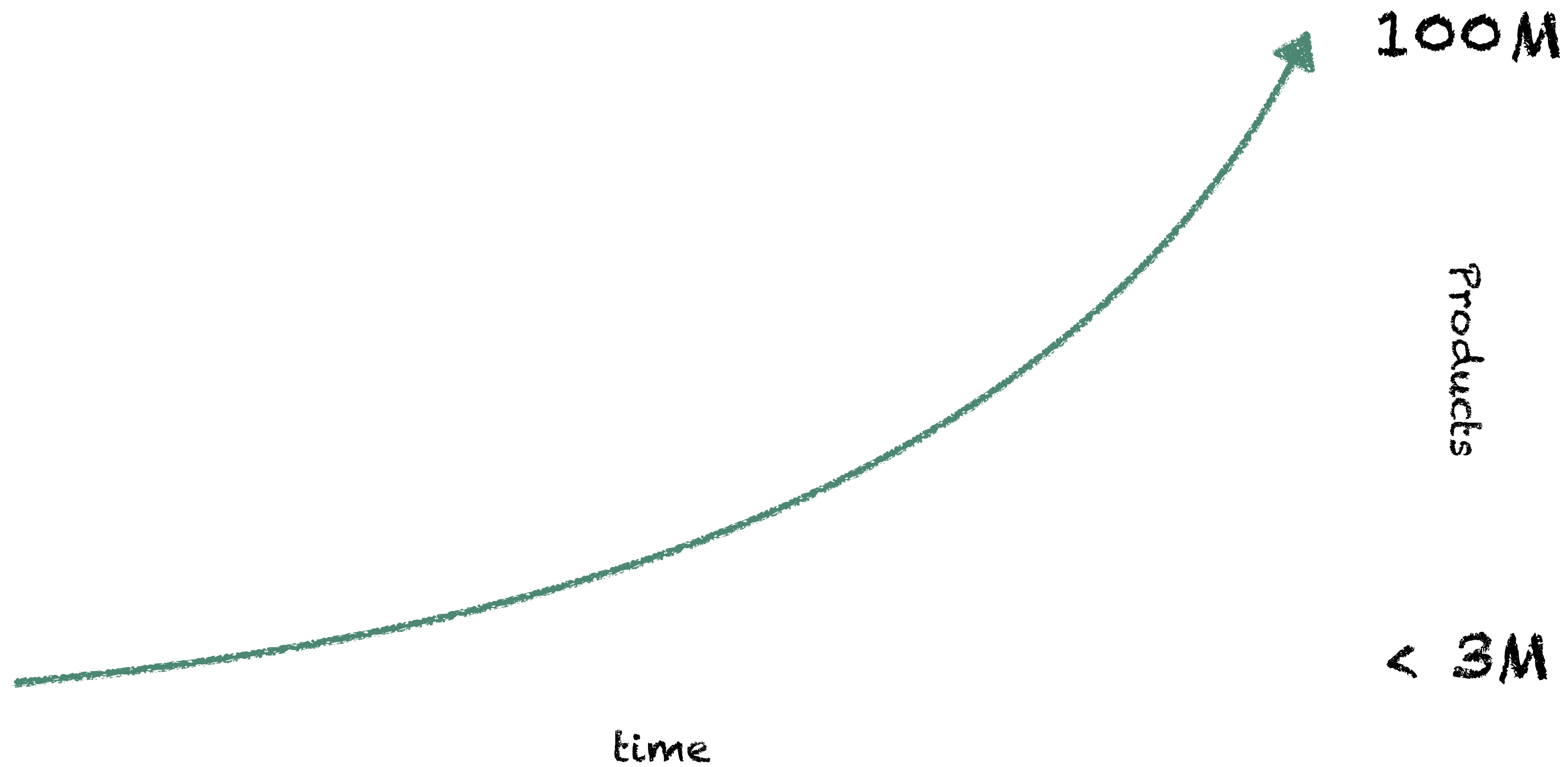
Solr



spring



The screenshot displays the shopping24 website interface. At the top, there's a search bar with 'fernseher' entered. Below it, a list of products is shown, including Samsung LED TVs. A sidebar on the left provides filters for categories, brands, and colors. Below the main product list, there's a section for 'living24.de' with search results for 'lampe'. To the right, a 'YA LOOK' advertisement for sneakers is visible, showing various styles and brands like Polo Sport, DKNY, and BOSS GREEN.



#discspace (indexsize)

#precision

#nosharding

1. Reduce the indexed data / text!

simple winnings:
speed & space

Solutions for:
Product descriptions
vs.
usual text

Dieser Multifunktionsschuh lässt Ihnen die Wahl: Wandern, Trailrun oder einfach mal eine Runde Laufen! Ihm ist nichts zu anspruchsvoll. Das Obermaterial wurde aus wasserabweisendem und verstärktem Nylonmesh mit Synthetikbesätzen gefertigt. Innen erwartet Sie ein angenehmes Textilfutter und eine herausnehmbare Innensohle. Das antibakterielle OrthoLite-Fußbett sorgt für ein optimales Fußklima. Die doppelte EVA-Zwischensohle mit zweitem Härtegrad, absorbiert mit dem zusätzlichen Dämpfungselement im Rückfuß Stöße für ein angenehmeres Laufgefühl. Durch das innovative Schnürsystem mit Feststeller und Verstaumöglichkeit auf der Zungenoberseite sind Sie noch schneller Einsatzbereit. Das Contagrip-Sohlensystem sorgt für eine hohe Stabilität im ganzen Fuß und schützt vor Umknicken. Mit dem griffigen Noppenprofil finden Sie stets guten Halt. Durch Reflektoren sind Sie im

SEO / IRRELEVANT TEXT

Dunkeln besser zu sehen. **RELEVANT TEXT** 3. L37107600 //

Wasserabweisend bedeutet durch Imprägnierung bedingter Nässeschutz (Regen, Tauwasser etc.). Das Eindringen des Wassers wird nur minimiert, kein kompletter Nässeschutz. Wenn Sie wasserfeste Schuhe möchten empfehlen wir Ihnen unseren Filter - Wasserschutz.

SEO / IRRELEVANT TEXT

descriptions: lots of useless text / signals

SEO text = evil = irrelevant

...combine it with...

...will bring you a lot of fun on the road...

...fits pretty well to...

...it will give a really good feeling...

...enjoy it at home on your couch...

"We can use fancy stuff like ..."

OpenNLP Framework

Keyword extraction

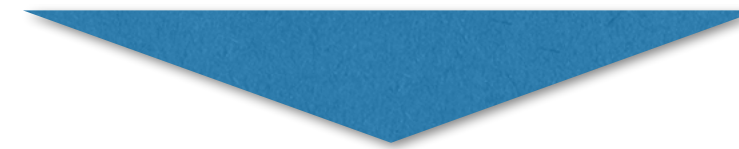
SVM's

Neural Networks

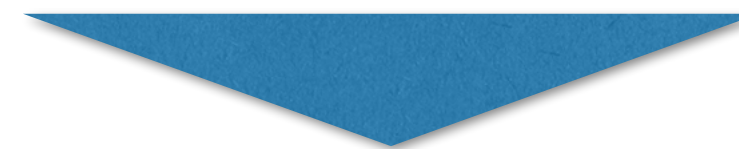
DeepLearning

Product pipeline

Gathering / Preprocessing / Normalization



Named entity recognition



Apache
Solr
... indexing

```
<fieldType name="text" class="solr.TextField" positionIncrementGap="100">
  <analyzer type="index">
    <!-- Remove HTML tags and decode HTML character entities like &#039; in the input -->
    <charFilter class="solr.HTMLStripCharFilterFactory" />

    <!-- Normalize all whitespace to a single space character -->
    <charFilter class="solr.PatternReplaceCharFilterFactory"
      pattern="\s+"
      replacement=" " />

    <!-- Use the sentence tokenizer, which removes "noise" sentences and keeps only "signal" -->
    <tokenizer class="com.s24.search.solr.analysis.AnalyzingSentenceTokenizerFactory"
      stopwordfile="etc/generated/list_stopwords.de.txt"
      filter="true" />
```

Key assumption:

"Sentences in product descriptions with useful information do not contain a lot of stop words"
(from the view of a search engine)

```
List sentences = splitToSentences(cleanUp(description));  
  
for(s : sentences){  
  
    if(s contains a lot of signs){  
        // split again  
    }  
  
    int stopwordCount = countStopwords(s);  
    int wordCount = countWords(s);  
  
    if(stopwordCount / wordCount < threshold){  
        // sentence is interesting  
    }  
}
```



~21%

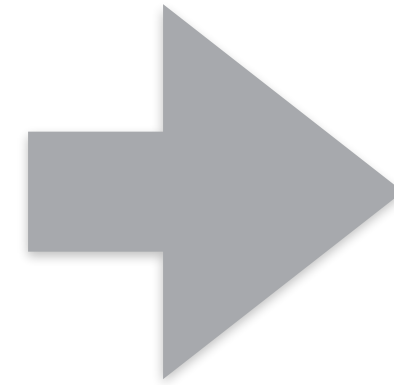
Learnings

- Power of RegExp
(libs with heuristics are not as fast)

```
Pattern.compile("(?<=[.?!\\|;-])\\s+(?=\\p{Lu})");
```

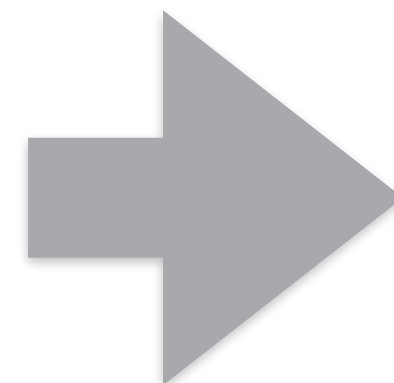
- Do not analyze text that is just one sentence
- Split sentences again

Trendige Jeans bei Herreenausstatter: Jeans von Replay – Waitom Eine modische Jeans aus der Laserblast-Reihe: Die lässige Optik und die markante Waschung entstehen durch ein speziell von Replay entwickeltes Verfahren. Der authentische Look wird durch einen besonderen Waschvorgang erreicht, der zum einen die Konturen verfeinert, und dazu noch umweltfreundlich ist. Gut zu wissen: Die Wasserersparnis liegt bei diesem Verfahren bei ca. 85%! Schnitt/Verarbeitung: Regular Slim – Five-Pocket-Schnitt. Etwas schmaler geschnittenes, gerades Bein, reguläre Bundhöhe. Button Fly: Verdeckte Knopfleiste mit drei Knöpfen, plus ein Knopf im Bund. Bund mit 5 Gürtelschlaufen für eine Gürtelbreite von max. ca. 5,5 cm. Vorne 2 Eingriffstaschen und eine schräg aufgesetzte Feuerzeugtasche rechts. Zwei Gesäßtaschen. Hüftsattel im Rückenteil. Die Vorderen Taschen sind mit Nieten gesichert. Nähte Dunkelblau. Logo-Patch hinten rechts am Bund und in Dreiecksform am linken Tascheneingriff, rot gesticktes Logo in der rechten Gesäßtasche. Maße: Bei W33 / L34 Beinlänge ca. 110 cm, Fußweite ca. 40 cm. Material: 100% Baumwolle in Denim-Qualität. Farbe/Dessin: Helles Denimblau – stark gewaschener Denim. Vereinzelte Destroyed-Effekte an den Tascheneingriffen, am Oberschenkel und an Saumkanten, Sitzfalten in der Hüftbeuge und hellere Abriebe an Kanten, Nähten, Oberschenkel und Gesäß.



Schnitt/Verarbeitung: Regular Slim – Five-Pocket-Schnitt.
Etwas schmaler geschnittenes, gerades Bein, reguläre Bundhöhe.
Zwei Gesäßtaschen.
Hüftsattel im Rückenteil.
Nähte Dunkelblau.
Maße: Bei W33 / L34 Beinlänge ca. 110 cm, Fußweite ca. 40 cm.
Material: 100% Baumwolle in Denim-Qualität.
Farbe/Dessin: Helles Denimblau – stark gewaschener Denim.

Glücksgriff: Wir kennen ihn, den Griff zur Garderobe, wenn wir das Haus verlassen wollen. Dieser geht meistens blind in eine Richtung und zu einer bestimmten Jacke, oder zu einem Mantel. Dieser Parka könnte der neueste Glücksgriff werden, denn dank des herausnehmbaren Innenfutters ist er das ganze Jahr über tragbar. Die Farbe ist ein Multistylor und der Schnitt ein Trend, der uns noch lange erhalten bleibt. Details: abnehmbare Kapuze, zwei Reißverschlusstaschen, Ziernieten, Tunnelzug zur Taillenregulierung, zwei seitliche Eingriffstaschen und zusätzliche Reißverschlusstaschen, Riegel an den Armabschlüssen, Tunnelzug am Saum, herausnehmbares Kunstfellfutter, Material: 56% Polyester, 33% Baumwolle, 11% Nylon, Futter: 100% Polyester, Füllung: 100% Polyester, Maße: Länge ca. 76 cm.



Details: abnehmbare Kapuze,
zwei Reißverschlusstaschen,
Ziernieten,
Tunnelzug zur Taillenregulierung,
zwei seitliche Eingriffstaschen und zusätzliche Reißverschlusstaschen,
Riegel an den Armabschlüssen,
Tunnelzug am Saum,
herausnehmbares Kunstfellfutter,
Material: 56% Polyester,
33% Baumwolle,
11% Nylon,
Futter: 100% Polyester,
Füllung: 100% Polyester,
Maße: Länge ca. 76 cm.

Word reduction of ~60-80%

No SEO text or product description?

Keywordextraction: RAKE-algorithm

Apache Solr

From Wikipedia, the free encyclopedia

Solr (pronounced "solar") is an [open source enterprise search](#) platform, written in [Java](#), from the **Apache Lucene** project. Its major features include [full-text search](#), hit highlighting, [faceted search](#), real-time indexing, dynamic clustering, database integration, [NoSQL](#) features^[1] and rich document (e.g., Word, PDF) handling. Providing distributed search and index replication, Solr is designed for scalability and [Fault tolerance](#).^[2] Solr is the second-most popular enterprise search engine after [Elasticsearch](#).^[3]

Solr runs as a standalone full-text search server. It uses the [Lucene](#) Java search library at its core for full-text indexing and search, and has [REST-like HTTP/XML](#) and [JSON](#) APIs that make it usable from most popular programming languages. Solr's external configuration allows it to be tailored to many types of application without Java coding, and it has a plugin architecture to support more advanced customization.

Apache [Lucene](#) and Apache Solr are both produced by the same Apache Software Foundation development team since the two projects were merged in 2010. It is common to refer to the technology or products as Lucene/Solr or Solr/Lucene.

Rake: How it works

- RAKE-algorithm:
 - Uses stopwords and punctuation as boundaries
 - Calculates a score for each candidate
 - Returns 1/3 of the top candidates as keyword result

Solr is an open source enterprise search platform,
written in Java, from the Apache Lucene project.

● candidate ● boundary

Rake:

- Enhancements:
 - Define domain specific "stop words":
additional word types, more signs, urls...
 - Propagate score from overlapping / related keywords
 - (Additional filtering)



@github: [shopping24/solr-analyzers](#)

mail:

work@tobiaskaessmann.de

dev blog:

developer.s24.com

