Lucene&Co

Zbyszko Papierski
mephiss@gmail.com
@ZPapierski

Lucene - indexing

Analysis

- CharFilter
- Tokenizer
- Filter
- … in any combination

Lucene - querying

- Analysis process - identical as indexing
- …or not
- Relevancy scoring
  - TF-IDF
  - there are others, less used

Almighty TF-IDF

$$score(q,d) = coord(q,d) * queryNorm(q) * \sum_{t\ in\ q} (tf(t\ in\ d) * idf(t)^2 * t.getBoost() * norm(t,d))$$

Lucene in Action

Solr

- Developed concurrently with Lucene
- Very popular
- Sharding/replication
- Cloud support as an afterthought

Solr in Action

Elasticsearch

- Exteremaly powerful out-of-the-box
- … but fully configurable without losing it's flexibility

Elasticsearch in Action

Why Solr?

- Better support
- More popular
  - since March 2014, Amazon's CloudSearch is fueled by Solr
- Same team as Lucene
- Shard splitting capabilities

Why Elasticsearch?

- Better data integration & visualisation
- Very rapid development
- Very flexible
- Simple REST API
- Automatic shard rebalancing
- better data representation

Links

https://github.com/maeph/JUG-Lucene - examples

http://lucene.apache.org/solr/ - Solr

http://lucene.apache.org/ - Lucene

http://www.elasticsearch.org/ - Elasticsearch

http://solr-vs-elasticsearch.com/ - feature comparision - updated!

Queries?