Academic Research to Open Source
Open Source AI Workshop

Andy Hind
April 5, 2019
Who am I?

• Andy Hind - Reformed academic?
  – Oracle
  – Alfresco
  – Campden BRI
  – University of Edinburgh – Chemical Engineering
  – British Antarctic Survey
Agenda

1. Introduction
2. Document Fingerprints
3. Getting it into Lucene and SOLR
4. Vectors are interesting ...
5. The journey
New ideas appearing in Lucene/SOLR

• Learning to rank
  – RankNet 2005/LambdaMART 2010
  – SOLR 2015 (rerank 2014) - Elastic 2017
New ideas appearing in Lucene/SOLR

- (b)kd – trees
  - Paper 2003
  - Lucene 2015
Numeric Types in Lucene

- NumberTools
  1. BCDTypeField
  2. SortableTypeField
  3. TypeField
- UTF-8 terms
  4. NumberTools
- Auto-prefix terms
  5. Auto-prefix terms

- Modified UTF-8 terms
- FieldCache

- Lucene 0.01
  - March 2000
- Lucene 1.0
  - July 2004
- Solr 1.1
  - Dec. 2006
- Lucene 2.9
  - Sept. 2009
- Lucene/Solr 4.0
- Solr 1.4
  - Nov. 2009
- Lucene 2.4
  - Oct. 2008
- Lucene/Solr 5.2
  - June 2015
- Lucene/Solr 6.2
  - Aug. 2016
- Lucene 2.9
  - Feb. 2015
- Lucene/Solr 6.0
  - Apr. 2016
Numeric Types in Lucene
Encoded Strings

• Encode information in tokens

• Multi-lingual indexing
  – Encode locale/analysis chain ... {en}woof

• Many fields
  – Encode field id .... woof:1
  – Salesforce – Activate 2018
New ideas appearing in Lucene/SOLR

• Locality Sensitive Hashing & Minhash
  – AltaVista - 1997
  – Lucene 2016/SOLR 2018
New ideas appearing in Lucene/SOLR

• PID control from the 1920s .....
Agenda

1. Introduction
2. Document Fingerprints
3. Getting it into Lucene and SOLR
4. Vectors are interesting ...
5. The journey
Document Fingerprints

- Document similarity
- “More like this”
  - SOLR term vectors
    - index is 7.8 x larger (http://blog.mikemccandless.com/2012/)

- ???
  - (Near) duplicates
  - Inclusion
  - Query expansion (recall)
  - Feature for LTR (precision)
  - Smaller
Document Fingerprints – LSH – Minhash

- Mining of Massive Datasets - http://www.mmds.org
  - Chapter 3 “Finding Similar Items”

- Jaccard similarity of documents - BOW
- Similarity does not have to be high to be significant
  - Character N-grams
  - Word Shingles
- Minhash
- Locality Sensitive Hashing – approximate nearest neighbour search
  - Data dependent or independent
Document Fingerprints – LSH – Minhash - Timeline


• 2012 - Mining of Massive Datasets - [http://www.mmds.org](http://www.mmds.org)


• 2016 – [https://issues.apache.org/jira/browse/LUCENE-6968](https://issues.apache.org/jira/browse/LUCENE-6968)
Document Fingerprints – LSH – Minhash

• Mining of Massive Datasets - http://www.mmds.org
  – Chapter 3 “Finding Similar Items”

– Jaccard similarity of documents - BOW
– Similarity does not have to be high to be significant
  • Character N-grams
  • Word Shingles
– Minhash
– Locality Sensitive Hashing – approximate nearest neighbour search

5 word shingle
Document Fingerprints – LSH – Minhash

• Mining of Massive Datasets - [http://www.mmds.org](http://www.mmds.org)
  – Chapter 3 “Finding Similar Items”

  – Jaccard similarity of documents - BOW
    – *Similarity does not have to* be high to be significant
      • Character N-grams
      • Word Shingles
  – Minhash
  – Locality Sensitive Hashing – approximate nearest neighbour search
Document Fingerprints - Example

A  
CMIS 1.0  5 word n-grams

The Content Management Interoperability Services (CMIS) standard defines a domain model and Web Services and Restful AtomPub bindings that can be used by applications to work with one or more Content Management repositories/systems.

B  
CMIS 1.1  5 word n-grams

The Content Management Interoperability Services (CMIS) standard defines a domain model and Web Services, Restful AtomPub and browser (JSON) bindings that can be used by applications to work with one or more Content Management repositories/systems.

\[ C(A, B) = \frac{23}{30} \approx 77\% \]
\[ C(B, A) = \frac{23}{32} \approx 72\% \]
\[ J(A, B) = \frac{23}{39} \approx 59\% \]
Min Hash – set

\[ J(A, B) = \frac{|A \cap B|}{|A \cup B|} \]
Min Hash – many hash functions
Min Hash – one hash with buckets
Min Hash – one hash with buckets + rotation
Min Hash – comparing hashes
Min Hash – comparing hashes – with banding
Agenda

1. Introduction
2. Document Fingerprints
3. Getting it into Lucene and SOLR
4. Vectors are interesting …
5. The journey
Similar Documents

• Lucene
  – MinHashFilter
  – [https://issues.apache.org/jira/browse/LUCENE-6968](https://issues.apache.org/jira/browse/LUCENE-6968)
  – 6 months

• SOLR
  – min_hash MinHashQParser
  – 3 days + a month to catch up on documentation
Similar Documents

• Analysed vs pre-analysed and stored
• Analysis chain
  – n-grams vs shingles etc
• Hashes, buckets, minimum set, rotation

• Similarity
Examples

• Wikipedia articles
• 5 - word shingles
• Pre-analysed and stored

• Aside
  – State in the index
  – Event sourcing/CQRS
# Oracle Corporation

<table>
<thead>
<tr>
<th>Page</th>
<th>Score</th>
<th>Normalised</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle Corporation</td>
<td>512</td>
<td>1.000</td>
</tr>
<tr>
<td>Oracle Cloud</td>
<td>9</td>
<td>0.018</td>
</tr>
<tr>
<td>Oracle Cloud Platform</td>
<td>5</td>
<td>0.010</td>
</tr>
<tr>
<td>Michelle K. Lee</td>
<td>5</td>
<td>0.010</td>
</tr>
<tr>
<td>Paul Grewal</td>
<td>4</td>
<td>0.008</td>
</tr>
<tr>
<td>Ultratech</td>
<td>4</td>
<td>0.008</td>
</tr>
</tbody>
</table>
## Oracle Cloud

<table>
<thead>
<tr>
<th>Page</th>
<th>Score</th>
<th>Normalised</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle Cloud</td>
<td>512</td>
<td>1.000</td>
</tr>
<tr>
<td>Oracle Cloud Platform</td>
<td>148</td>
<td>0.289</td>
</tr>
<tr>
<td>Oracle Corporation</td>
<td>17</td>
<td>0.033</td>
</tr>
<tr>
<td>Microsoft Azure</td>
<td>10</td>
<td>0.020</td>
</tr>
<tr>
<td>Recovery as a service</td>
<td>9</td>
<td>0.018</td>
</tr>
<tr>
<td>SHI International Corp</td>
<td>8</td>
<td>0.016</td>
</tr>
<tr>
<td>Cloud28+</td>
<td>8</td>
<td>0.016</td>
</tr>
<tr>
<td>Content as a service</td>
<td>6</td>
<td>0.012</td>
</tr>
</tbody>
</table>
## Brexit

<table>
<thead>
<tr>
<th>Topic</th>
<th>Score</th>
<th>Normalised</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brexit</td>
<td>512</td>
<td>1.000</td>
</tr>
<tr>
<td>Brexit negotiations</td>
<td>30</td>
<td>0.059</td>
</tr>
<tr>
<td>Brexit in popular culture</td>
<td>22</td>
<td>0.043</td>
</tr>
<tr>
<td>History of European Union–United Kingdom relations</td>
<td>19</td>
<td>0.037</td>
</tr>
<tr>
<td>Economic effects of Brexit</td>
<td>11</td>
<td>0.021</td>
</tr>
<tr>
<td>European Parliament election, 2019</td>
<td>8</td>
<td>0.016</td>
</tr>
<tr>
<td>Aftermath of the United Kingdom European Union membership referendum, 2016</td>
<td>7</td>
<td>0.014</td>
</tr>
<tr>
<td>United Kingdom invocation of Article 50 of the Treaty on European Union</td>
<td>7</td>
<td>0.014</td>
</tr>
</tbody>
</table>
## Scott Joplin

<table>
<thead>
<tr>
<th>Page</th>
<th>Score</th>
<th>Normalised</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scott Joplin</td>
<td>512</td>
<td>1.000</td>
</tr>
<tr>
<td>Treemonisha</td>
<td>38</td>
<td>0.074</td>
</tr>
<tr>
<td>List of compositions by Scott Joplin</td>
<td>15</td>
<td>0.030</td>
</tr>
<tr>
<td>The Entertainer (rag)</td>
<td>15</td>
<td>0.030</td>
</tr>
<tr>
<td>Scott Joplin House State Historic Site</td>
<td>13</td>
<td>0.025</td>
</tr>
<tr>
<td>Scott Joplin: Piano Rags</td>
<td>13</td>
<td>0.025</td>
</tr>
<tr>
<td>Joshua Rifkin</td>
<td>7</td>
<td>0.014</td>
</tr>
<tr>
<td>Bethena</td>
<td>5</td>
<td>0.010</td>
</tr>
</tbody>
</table>
Stuff still to do …

• New BKD type
• Normalised score
• Positions/Highlighting
  – Where was it similar?
• Hash size and collisions
• Rotation bug …
Agenda

1. Introduction
2. Document Fingerprints
3. Getting it into Lucene and SOLR
4. Vectors are interesting ...
5. The journey
Vectors are interesting ...

• Dense Vectors
  – Embeddings
  – Post processing

• Approximate nearest neighbours
  – Locality Sensitive Hashing (LSH – SimHash, spectral hashing, ...)
  – K-Means tree
  – Randomized KD forest
  – Vector to text encoding
  – Brute force
NLP

• NVIDIA DGX-1 Deep Learning System with 8x 32GB Tesla V100 Volta GPUs, 12nm, HBM2, 1 petaFLOP FP16 Performance
NLP

- NVIDIA DGX-1 Deep Learning System with 8x 32GB Tesla V100 Volta GPUs, 12nm, HBM2, 1 petaFLOP FP16 Performance

£138,258.49

$18/hour
Vectors are interesting ...

• NLP and transfer learning
  – Sentence representations
  – NLP’s image net moment?
  – Transfer learning needs less data ....

• Text vs images
Agenda

1. Introduction
2. Document Fingerprints
3. Getting it into Lucene and SOLR
4. Vectors are interesting ...
5. The journey
The Journey

• Papers ...
• Math vs application
• Implementation
• Scalability
Papers ...

• Performance Comparison of Learning to Rank Algorithms for Information Retrieval

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>NDCG@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF*IDF</td>
<td>0.7051</td>
</tr>
<tr>
<td>BM25</td>
<td>0.7800</td>
</tr>
<tr>
<td>RankSVM</td>
<td>0.8087</td>
</tr>
<tr>
<td>LambdaMART</td>
<td>0.8092</td>
</tr>
<tr>
<td>AdditiveGroves</td>
<td>0.8165</td>
</tr>
</tbody>
</table>

• https://pdfs.semanticscholar.org/cd12/e191d2c2790e5ed60e5186462e6f8027db1f.pdf
Papers ...


<table>
<thead>
<tr>
<th>Model</th>
<th>MQ2008</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NDCG@1</td>
</tr>
<tr>
<td>BM25-TITLE</td>
<td>0.344</td>
</tr>
<tr>
<td>RankSVM</td>
<td>0.375</td>
</tr>
<tr>
<td>RankBoost</td>
<td>0.381</td>
</tr>
<tr>
<td>AdaRank</td>
<td>0.360</td>
</tr>
<tr>
<td>LambdaMART</td>
<td>0.378</td>
</tr>
<tr>
<td>DSSM</td>
<td>0.286</td>
</tr>
<tr>
<td>CDSSM</td>
<td>0.283</td>
</tr>
<tr>
<td>Arc-I</td>
<td>0.295</td>
</tr>
<tr>
<td>SQA-noFeat</td>
<td>0.291</td>
</tr>
<tr>
<td>DRMM</td>
<td>0.368</td>
</tr>
<tr>
<td>Arc-II</td>
<td>0.299</td>
</tr>
<tr>
<td>MatchPyramid</td>
<td>0.351</td>
</tr>
<tr>
<td>Match-SRNN</td>
<td>0.369</td>
</tr>
<tr>
<td>DeepRank-2DGRU</td>
<td>0.391</td>
</tr>
<tr>
<td>DeepRank-CNN</td>
<td><strong>0.406</strong></td>
</tr>
<tr>
<td>SQA</td>
<td>0.402</td>
</tr>
<tr>
<td>DeepRank-CNN-Feat</td>
<td>0.418</td>
</tr>
</tbody>
</table>
A Dual Embedding Space Model for Document Ranking

<table>
<thead>
<tr>
<th>Model</th>
<th>Explicitly Judged Test Set</th>
<th>Implicit Feedback based Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NDCG@1</td>
<td>NDCG@3</td>
</tr>
<tr>
<td>BM25</td>
<td>21.44</td>
<td>26.09</td>
</tr>
<tr>
<td>LSA</td>
<td>04.61*</td>
<td>04.63*</td>
</tr>
<tr>
<td>DESM (IN-IN, trained on body text)</td>
<td>06.69*</td>
<td>06.80*</td>
</tr>
<tr>
<td>DESM (IN-IN, trained on queries)</td>
<td>05.56*</td>
<td>05.59*</td>
</tr>
<tr>
<td>DESM (IN-OUT, trained on body text)</td>
<td>01.01*</td>
<td>01.16*</td>
</tr>
<tr>
<td>DESM (IN-OUT, trained on queries)</td>
<td>00.62*</td>
<td>00.58*</td>
</tr>
<tr>
<td>BM25 + DESM (IN-IN, trained on body text)</td>
<td>21.53</td>
<td>26.16</td>
</tr>
<tr>
<td>BM25 + DESM (IN-IN, trained on queries)</td>
<td><strong>21.58</strong></td>
<td>26.20</td>
</tr>
<tr>
<td>BM25 + DESM (IN-OUT, trained on body text)</td>
<td>21.47</td>
<td>26.18</td>
</tr>
<tr>
<td>BM25 + DESM (IN-OUT, trained on queries)</td>
<td>21.54</td>
<td><strong>26.42</strong></td>
</tr>
</tbody>
</table>

Papers ...

- Learning a Deep Listwise Context Model for Ranking Refinement

<table>
<thead>
<tr>
<th>Model</th>
<th>Loss Function</th>
<th>Microsoft Letor Dataset 30K</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>nDCG@1</td>
</tr>
<tr>
<td>Initial List</td>
<td>DNN</td>
<td>LambdaMART</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ListMLE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SoftRank</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AttRank</td>
</tr>
<tr>
<td>LambdaMART</td>
<td>LIDNN</td>
<td>ListMLE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SoftRank</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AttRank</td>
</tr>
<tr>
<td></td>
<td>DLCM</td>
<td>ListMLE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SoftRank</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AttRank</td>
</tr>
</tbody>
</table>

• Balancing Speed and Quality in Online Learning to Rank for Information Retrieval [1711.09446.pdf]
Thank you

We’re hiring!