Topic Analysis of RDA Activities

A preliminary report

RDA TAB and IGWG Chairs Meeting, Gaithersburg, MD January 2017
Motivation

• Supplement the “Six Words” exercise

• Learn what RDA groups are doing by analyzing the documents they produce

• Create a corpus by crawling the RDA web pages

• Apply well-established techniques of Probabilistic Latent Semantic Analysis
Methodology

- Crawl rd-alliance.org using Apache Nutch, put docs in Solr (N=6,102)
- Select all documents containing “case-statement” (N=196)
- Remove stop words, do lemmatization
- Use Gensim toolkit to do Latent Dirichlet Allocation (LDA)
Stopwords
Basic Concepts

- **Topic Modeling** identifies *topics* and their *distributions* across the *documents* in a *corpus*.

- **Generative probabilistic models** use statistical methods to discover hidden (i.e., "latent") themes (topics) in documents.

- "Generative" means we assume the source documents were generated from a mixture of topics, with each topic being a distribution over words belonging to that topic.

- A word can belong to multiple topics.

- Neither the order of the documents in the corpus, nor the order of the words in each document are taken as significant.
LDA results w/ 15 topics, 8 words

<table>
<thead>
<tr>
<th>Topic</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>plan</td>
<td>provenance</td>
<td>collection</td>
<td>link</td>
<td>broker</td>
<td>national</td>
<td>indigenous</td>
<td>cost</td>
<td>storage</td>
<td>legal</td>
<td>type</td>
<td>agriculture</td>
<td>brokering</td>
<td>metadata</td>
<td>rice</td>
</tr>
<tr>
<td>0</td>
<td>platform</td>
<td>disciplinary</td>
<td>privacy</td>
<td>literature</td>
<td>mediation</td>
<td>workflow</td>
<td>sov</td>
<td>database</td>
<td>qos</td>
<td>interoperability</td>
<td>creating</td>
<td>interoperability</td>
<td>governance</td>
<td>standard</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>dmp</td>
<td>discipline</td>
<td>pid</td>
<td>infrastructure</td>
<td>component</td>
<td>publishing</td>
<td>international</td>
<td>field</td>
<td>provider</td>
<td>law</td>
<td>past</td>
<td>agricultural</td>
<td>model</td>
<td>standard</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>fabric</td>
<td>professor</td>
<td>metadata</td>
<td>publishing</td>
<td>registry</td>
<td>tool</td>
<td>network</td>
<td>recovery</td>
<td>datac</td>
<td>domain</td>
<td>registry</td>
<td>semantic</td>
<td>infrastructure</td>
<td>middleware</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>requirement</td>
<td>reproducibility</td>
<td>quality</td>
<td>article</td>
<td>resource</td>
<td>publication</td>
<td>sovereignty</td>
<td>farmer</td>
<td>vocabulary</td>
<td>codata</td>
<td>standard</td>
<td>marine</td>
<td>interoperability</td>
<td>approach</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>course</td>
<td>illinois</td>
<td>digital</td>
<td>publisher</td>
<td>standard</td>
<td>preservation</td>
<td>security</td>
<td>centre</td>
<td>multiple</td>
<td>international</td>
<td>national</td>
<td>semantics</td>
<td>approach</td>
<td>support</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>management</td>
<td>collaboration</td>
<td>set</td>
<td>scholix</td>
<td>provide</td>
<td>policy</td>
<td>citation</td>
<td>network</td>
<td>document</td>
<td>ccm</td>
<td>record</td>
<td>vre</td>
<td>approach</td>
<td>support</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>dmps</td>
<td>need</td>
<td>related</td>
<td>hub</td>
<td>need</td>
<td>activity</td>
<td>trust</td>
<td>management</td>
<td>adoption</td>
<td>national</td>
<td>ccm</td>
<td>vre</td>
<td>vre</td>
<td>humanity</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Sample Topics

- Case Statement from “Agrisemantics” WG has high probability (P=0.999) that it contains Topic 11:
  - agriculture, interoperability, semantic, marine, semantics, infrastructure, vre
- Also “Virtual Research Environments” IG (P=0.997) and “Marine Data Management” WG (P=0.672)

- Statement for “On-Farm Data Sharing” WG contains Topic 7:
  - cost, database, field, recovery, farmer, centre, network, management

- Earlier LDA analysis had “Array Database” WG in Topic with:
  - array, database, big, domain, document, report, plan
- but in this analysis has P=0.995 that contains Topic 7
LDA results can be viewed at:

https://goo.gl/V6Yraj
What’s Next?
✧ Refine list of stop words
✧ Remove duplicate documents (last version only?)
✧ Further tune algorithms
✧ Aggregate topic probabilities by group
✧ Include other documents (e.g. outputs)
✧ Assign dates to document, show Topic trends
✧ Other?
Thanks for your attention!