WikiData for Curation Technologies: Topic Detection with Knowledge Graph Embeddings

Background

In machine learning and natural language processing, many techniques are successful in the research world yet cannot be applied directly to real life problems. For example, well-established methods for topic modeling can find what topic a given document belongs to. However, these methods provide topics only as numeric IDs that are not human-readable and not suitable for organizing your document collection. We incorporate the knowledge from WikiData into existing topic modeling approaches such that WikiData items represent topics. Our work demonstrates how WikiData is a valuable resource for emerging curation technologies.

QURATOR

The QURATOR project is a co-operation between DFKI, Wikimedia and eight other partners. QURATOR contributes to making curating activities and the generation of digital content in various industry contexts more efficient and cost-effective by automating high-quality processes, and to supporting knowledge workers and editors in curating digital content.

Knowledge Graph Embeddings

To apply machine learning on any data, you first need to represent the data as numerical vectors, the so-called embeddings. The goal is that semantically similar data, for example, words or graph nodes, are assigned to vectors which are close to each other [2].

From Topic IDs to Labels

LDA topics are not human-readable (a topic is only an ID). In the QURATOR project we solve this and also make the topics “linked-data ready” by connecting them with Wikidata. Our approach: Find Wikidata items that correspond to the LDA topic terms by computing the weighted mean over vectors of topic terms and finding the nearest Wikidata item:

\[ v_{\text{topic}} = \frac{1}{k}(w_1 * v_{\text{term}_1} + w_2 * v_{\text{term}_2} + \cdots + w_k * v_{\text{term}_k}) \]

Contact

Malte Ostendorff
Email: malte.ostendorff@dfki.de
Twitter: @XYOU
Web: www.dfki.de

References


Topic Detection with Latent Dirichlet Allocation

The method Latent Dirichlet Allocation (LDA) is a popular technique to model topics within text documents [1]. Underlying idea is as follows:

- Documents with similar topics will use similar groups of words.
- Documents are probability distributions over latent topics
- Topics are probability distributions over words

Finding Topics in BBC News Articles

When we apply Latent Dirichlet Allocation on a corpus on BBC articles the following topics are detected:

<table>
<thead>
<tr>
<th>Term</th>
<th>Topic #1</th>
<th>Topic #2</th>
<th>Topic #3</th>
<th>Topic #4</th>
<th>Topic #5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Music</td>
<td>Government</td>
<td>People</td>
<td>Player</td>
<td>Award</td>
<td></td>
</tr>
<tr>
<td>2 World</td>
<td>Party</td>
<td>Market</td>
<td>England</td>
<td>People</td>
<td></td>
</tr>
<tr>
<td>3 Number</td>
<td>Labour</td>
<td>Company</td>
<td>Match</td>
<td>Computer</td>
<td></td>
</tr>
<tr>
<td>4 Record</td>
<td>Election</td>
<td>Mobile</td>
<td>Wales</td>
<td>Includes</td>
<td></td>
</tr>
<tr>
<td>5 Include</td>
<td>People</td>
<td>Phone</td>
<td>Ireland</td>
<td>Music</td>
<td></td>
</tr>
</tbody>
</table>

The original topics are Sport (T4), Business (T5), Politics (T2), Tech (T3) and Entertainment (T1).