A keyphrase suggestion engine for semi-automated document characterization

Nicholas Taylor on behalf of James Powell, Dylan Johnson, Tim Mandzyuk, Daniel Waybright, and Alex Shocklee

Research Library

16 November 2023

LA-UR-23-32817
problem statement

• self-deposit users are predominantly task-driven
• users like more and higher-quality metadata, for discovery
• users are more ambivalent about providing supplemental metadata, when depositing

how can we streamline the generation of authoritative, supplemental metadata while leveraging the author's expertise?
solution?

• hire (many) more staff
• train them on breadth of LANL science
• train them on domain vocabularies
• have them read every submission
• have them supply keyphrases
• have authors validate quality
solution!

- apply ML and NLP tools
- have those “read” submissions
- then have them suggest keyphrases
- validate quality of metadata output
scope

• standalone web service
• input abstract full text
• output scored keyphrase suggestions
• model retraining optional
• performance on commodity hardware
• detailed logging

“checklist” by cylonfingers under CC BY-SA 2.0
solution

- Bidirectional Encoder Representations from Transformers (BERT)
- general-purpose NLP tool
- superior to dumb statistical techniques as it understands semantic similarity
- lighter-weight, performant models: KeyBERT, SciBERT, DistilBERT
- runs on standard VM (for now)
- JSON API
keyphrase app test interface

BERT API Test Page

Select BERTs to use [required]
- KeyBERT
  Similarity: 
  Confidence: 
  Keyphrase #: 
- SciBERT
  Similarity: 
  Confidence: 
  Keyphrase #: 
- DistILBERT
  Similarity: 
  Confidence: 
  Keyphrase #: 

Keybert Input [required]

Text to be keybert-ed...

(1000 character limit)
choose model(s)
specify similarity

- floating point number, 0-1
- maximum allowed similarity between generated keyphrases
- if similarity of any given pair exceeds threshold, lower-confidence keyphrase discarded
specify confidence

- floating point number, 0-1
- minimum confidence for a suggested keyphrase
- keyphrase not suggested if confidence succeeds threshold
specify number of keyphrases

- whole number
- maximum number of keyphrases to generate
supply full text

- 1,000 character limit
- designed to accommodate typical journal article abstract length
example values

• all models selected
• similarity: .9
• confidence: .35
• keyphrases: 5
• input: (abstract full text from first COVID-19 pre-print posted on arXiv)
example outputs

Mark the accepted keywords. [required]

<table>
<thead>
<tr>
<th>KeyBert</th>
<th>SciBert</th>
<th>DistilBert</th>
</tr>
</thead>
<tbody>
<tr>
<td>novel coronavirus (0.5918)</td>
<td>novel coronavirus (0.5555)</td>
<td>structural loop (0.9999413)</td>
</tr>
<tr>
<td>other coronaviruses (0.3677)</td>
<td>bat coronaviruses (0.4973)</td>
<td>structural modeling (0.99992114)</td>
</tr>
<tr>
<td>bat coronaviruses (0.5453)</td>
<td>other coronaviruses (0.493)</td>
<td>receptor binding module (0.9998897)</td>
</tr>
<tr>
<td>ncov spike glycoprotein (0.5177)</td>
<td>ncov spike glycoprotein (0.456)</td>
<td>coronavirus (0.99986994)</td>
</tr>
<tr>
<td>common receptor (0.4064)</td>
<td>china (0.3574)</td>
<td>fusion (0.9943869)</td>
</tr>
</tbody>
</table>

Add additional keywords here (separated by commas).

Transaction ID: 1b0e8961-eac8-45a2-8727-6791c4cedac1

SUBMIT
evaluating quality

• only informally, internally so far
• start with close partners
• (if sufficiently fast) deploy in production, leverage analytics to iterate
how we’ll use it

• integrate into self-deposit workflow to suggest candidate keyphrases
• separate work underway for automated extraction of document elements (including abstract) using GROBID
• potentially apply for digitization post-processing or previous submissions?
• prototype LLM-based fielded text extraction?
thank you!

generated image for prompt “friendly robot waving goodbye” by Craiyon