Introduction

- Conventional topic models represent topics in a bag-of-words format that often requires "reading the tea leaves" to interpret; additionally, they offer a limited and inconvenient way to control topic semantics.
- We introduce TopicGPT, a framework that uses large language models (LLMs) to uncover latent topics in text corpora.

1. TopicGPT produces interpretable topics that consist of natural language labels and descriptions.

Wikipedia article

The Grant Park Music Festival (formerly Grant Park Concerts) is an annual ten-week classical music concert series held in Chicago, Illinois, USA. It features the Grant Park Symphony Orchestra and Grant Park Chorus along with featured guest performers and conductors...

Output topics

- TopicGPT: Music & Performing Art (Discuss creation, production, and performance of music, as well as related arts and cultural aspect).
- LDA: city, building, area, new, park.

2. TopicGPT is customizable to fit user needs through example topic guidance and semantic-based topic refinement.

1. Topic Generation

- Corpus

Example topics

- Trade
- Agriculture

Example topics (2-3 topics/dataset)

[Generated topics]

[Generated topics]

Instruction: Identify generalizable topics within the document.

2. Topic Assignment

Refinement Prompt

Remove infrequent topics

[Generated topics]

[Generated topics]

Instruction: Merge topic pairs that are near duplicates.

Assignment Prompt

[Refined topics]

[Refined topics]

Instruction: Assign generated topics to the provided document.

3. TopicGPT outperforms current state of the art in generating topics that are aligned with human-annotated topics.

- Baselines: LDA, BERTopic, SeededLDA
- Default setting: generator = GPT-4, assigner = GPT-3.5-turbo
- Dataset: Wikipedia articles (Wiki) and Congressional bills summaries (Bills)
- Alignment Metrics: Harmonic Mean Purity ($P_1$), Adjusted Rand Index (ARI), Normalized Mutual Information (NMI)

4. TopicGPT topics are semantically close to ground truth.

- Manual matching between ground truth, TopicGPT & LDA labels.
- Misaligned topics are categorized as (1) out-of-scope, (2) missing, or (3) repeated.

5. TopicGPT can also be extended to a hierarchical setting.

- Generated topics are treated as the top-level topics and LLMs are prompted to generate subtopics.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Setting</th>
<th>Out-of-scope</th>
<th>Missing</th>
<th>Repeated</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wiki</td>
<td>Default setting (k=31)</td>
<td>46.3</td>
<td>4.3</td>
<td>11.9</td>
<td>62.4</td>
</tr>
<tr>
<td></td>
<td>Refined (k=22)</td>
<td>30.3</td>
<td>0.0</td>
<td>0.0</td>
<td>30.3</td>
</tr>
<tr>
<td>Bills</td>
<td>Default setting (k=79)</td>
<td>56.1</td>
<td>2.1</td>
<td>22.0</td>
<td>80.2</td>
</tr>
<tr>
<td></td>
<td>Refined (k=24)</td>
<td>27.8</td>
<td>4.2</td>
<td>0.0</td>
<td>31.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>TopicGPT</th>
<th>LDA</th>
<th>BERTopic</th>
<th>SeededLDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wiki</td>
<td>P₁</td>
<td>ARI</td>
<td>NMI</td>
<td>P₁</td>
</tr>
<tr>
<td></td>
<td>0.73</td>
<td>0.58</td>
<td>0.71</td>
<td>0.59</td>
</tr>
<tr>
<td>Bills</td>
<td>P₁</td>
<td>ARI</td>
<td>NMI</td>
<td>P₁</td>
</tr>
<tr>
<td></td>
<td>0.57</td>
<td>0.40</td>
<td>0.52</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.33</td>
<td>0.49</td>
<td>0.33</td>
</tr>
</tbody>
</table>

TopicGPT: A Prompt-based Framework for Topic Modeling

Chau Minh Pham¹, Alexander Hoyle², Simeng Sun¹, Philip Resnik², Mohit Iyyer¹
¹University of Massachusetts Amherst, ²University of Maryland

Github: https://github.com/chtmp223/topicGPT