Precise Document Retrieval by Minimizing Kolmogorov Distance with Document Generation

CS898 Course Project
Xueguang Ma
Overview

• Text Retrieval
  • Lexical Retrieval
  • Dense Retrieval

• Kolmogorov Distance
  • Query-Documents Distance

• Minimizing Kolmogorov Distance by Document Generation
Text Retrieval

• Lexical Retrieval
  • BM25
  • TF-IDF

```
{"tok3": 1, "tok100": 2, "tok20002": 1}
{"tok64": 8.3, "tok100": 10.2, "tok20002": 53.5}
```

```
dot product of dim >30k
```
Text Retrieval

• Dense Retrieval
  • DPR

BERT

score

dot product of
dim 768

question

passage
Information Distance

\[ E(x, y) = \max\{K(x|y), K(y|x)\} = K(xy) - \min\{K(x), K(y)\} \]

\[ NCD(x, y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}} \]

The similarity metric. Ming Li et.al.
import gzip

def compute_kolmogorov_distance(query, text):
    C_query = len(gzip.compress(query.encode()))
    C_text = len(gzip.compress(text.encode()))
    C_query_text = len(gzip.compress((query + ' ' + text).encode()))
    ncd = (C_query_text - min(C_query, C_text)) / max(C_query, C_text)
    return ncd

Less is More: Parameter-Free Text Classification with Gzip
Zhiying Jiang et.al.
Dataset

• TREC DL19, DL20
  • Web Search Dataset
  • 8.8 Million Document Corpus
  • 0-3 Relevancy Score Judged
    • We consider score >=2 as relevant
Experiments

• Observe the accuracy that positive pairs has lower Kolmogorov distance than negative pairs $\text{NCD(positive)} < \text{NCD(negative)}$

• i.e. check how information distance align with query—document relevancy judgment.

• Q, P1, P2, ..., N1, N2 ...
## Query – Document Distance

<table>
<thead>
<tr>
<th></th>
<th>DL19</th>
<th>DL20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (min)</td>
<td>11.62%</td>
<td>12.96%</td>
</tr>
<tr>
<td>Accuracy (avg)</td>
<td>23.25%</td>
<td>33.33%</td>
</tr>
</tbody>
</table>
Why?

• Query – Document similarity is not symmetric
• There is gap in query/document distribution, makes its hard to measure information distance.

• Solution:
  • Hypothetical Document Generation
  • Casting query-document distance measure to document—document distance.
Hypothetical Document Generation

InstructGPT, Capture the user intent and generate a Hypothetical document.

1. Capture the query intent
2. Map query into document distribution
Hypothetical Document Generation

Example:

**Question**: where was Michael klim born? (ncd = 0.85)

**Generated Doc**: ...Michael Klim was born on August 12, 1976 in Melbourne, Australia. He is the son of Polish immigrants and was raised in the city's western suburbs. ....(ncd = 0.71)

**Relevant Doc**: Michael Klim was born in 1977 in Poland. He is married to Lindy Rama. They have one child...
• $Q, P_1, P_2, ..., N_1, N_2 ...$

• $Q \rightarrow D, P_1, P_2, ..., N_1, N_2, ...$
Pseudo Document– Document Distance

<table>
<thead>
<tr>
<th>Query--Document</th>
<th>DL19</th>
<th>DL20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (min)</td>
<td>11.62%</td>
<td>12.96%</td>
</tr>
<tr>
<td>Accuracy (avg)</td>
<td>23.25%</td>
<td>33.33%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (min) (GPT3.5)</td>
<td>65.12%</td>
<td>40.74%</td>
</tr>
<tr>
<td>Accuracy (min) (Curie)</td>
<td>51.16%</td>
<td>37.04%</td>
</tr>
<tr>
<td>Accuracy (avg) (GPT3.5)</td>
<td>100%</td>
<td>94.44%</td>
</tr>
<tr>
<td>Accuracy (avg) (Curie)</td>
<td>100%</td>
<td>96.29%</td>
</tr>
</tbody>
</table>
Retrieval with Hypothetical Document Generation

- write a passage to answer the question
  how long does it take to remove wisdom tooth

- write a scientific paper passage to answer the question
  How has the COVID-19 pandemic impacted mental health?

- write a passage in Korean to answer the question in detail
  인간은 언제 뿌를 사용했는가?

- It usually takes between 30 minutes and two hours to remove a wisdom tooth...

- depression and anxiety had increased by 20% since the start of the pandemic...

- How wisdom teeth are removed...
  Some... a few minutes, whereas others can take 20 minutes or longer....

- two studies investigating COVID-19 patients... significantly higher level of depressive ...

- 볼을 처음 사용한 시기는 호모 에렉투스가 살았던 142만 년 전으로 거슬러간다...
## Retrieval with Hypothetical Document Generation

<table>
<thead>
<tr>
<th></th>
<th>DL19</th>
<th></th>
<th>DL20</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>map</td>
<td>Ndcg@10</td>
<td>Recall@1k</td>
<td>map</td>
</tr>
<tr>
<td>BM25</td>
<td>30.1</td>
<td>50.6</td>
<td>75.0</td>
<td>28.6</td>
</tr>
<tr>
<td>Contriever</td>
<td>24.0</td>
<td>44.5</td>
<td>88.0</td>
<td>24.0</td>
</tr>
<tr>
<td>HyDE (BM25)</td>
<td>39.7</td>
<td>59.4</td>
<td>85.6</td>
<td>38.4</td>
</tr>
<tr>
<td>HyDE (Contriever)</td>
<td>41.8</td>
<td>61.3</td>
<td>88.0</td>
<td>38.2</td>
</tr>
</tbody>
</table>
Thank You

[1] Less is More: Parameter-Free Text Classification with Gzip
Zhiying Jiang, Matthew Y.R. Yang, Mikhail Tsirlin, Raphael Tang, and Jimmy Lin

[2] The similarity metric
Ming Li, Xin Chen, Xin Li, Bin Ma, and Paul MB Vitányi.

[3] Precise Zero-Shot Dense Retrieval without Relevance Labels
Luyu Gao*, Xueguang Ma*, Jimmy Lin, and Jamie Callan