Topic-modeling for Automatic Legal Documents

Tagging

COMP4801(2022-2023) Project Plan

Chan Wing Kit

Supervisor:
Prof. Benjamin Kao
1. **Introduction**
   
   1.1. **Searching Background**
   
   Given the rapid growth of digital contents available on the Internet, textual contents online are urging to be analyzed more effectively. The importance of searching algorithm and searching engine is undeniable since people need not to go through every document available on the site or over the whole Internet. We reply on keyphrase or keyword searching which is an idea to aim to filter out a small number of results that are most related to a very short query word set. The habit of human searching generates a problem, how to search more effectively? It is impossible that to force any arbitrary people make an accurate query. Indeed, keyphrase searching itself is not ideal at all, once the questioner made a misunderstand on their queried phrases or they may just realize what they screened is not what they want after processing numbers. In most of the searching cases, we need to refine our query often before achieving the result we want. However, time cost of screening is variant, it depends on the size of the document, also the preview of the document. In Google search, a hit result comes with an excerpt which are a two-line message captured with the hit highlighted keyphrase from the result document directly. It seems to be helping us to further filter out the result manually by just looking on the preview that where does the keyphrase appear, without clicking onto the URL. Nevertheless, this preview just becomes a fragmentation which cannot describe a document comprehensively, when the result is long and topical.

   1.2. **Tags**
   
   Tagging has been widely adopted as an application which assigning metadata with an object in Web 2.0. In general Web 2.0 application, the site allows user to tag and search the tag freely. In other words, tagging is a habit how people search and label digital contents, also the trend that how content provider manage their digital contents. A survey has concluded that tags perform well on improving searching, automatic classification and content recommendation. Moreover, the objective of tagging is to describe, summarize and organize the labeled contents [1].
In general, users are able to click on the tag for searching similar object with attached same tag or similar tag. Compared with automatic similar document matching, tags are visible and more interactive which can apply on previewing the search result. Tags guarantee the extensibility for further application, such as data visualization (tag clouds), trend analysis, classification...

1.3. **Natural Language Processing (Information Retrieval)**

Natural Language Processing (NLP) has become a mature research topic recently. Numbers of algorithms and models have been introduced to help computers understand and able to analysis human language to perform semantic analysis on long document. Term Frequency-Inverse Documentation Frequency (TD-IDF) has been adopted widely to extract keyword from document by frequency. It is described as lack of semantic explanation. [2] Since the extracted keywords are not capable to summarize the meaning of long documents, researcher introduced topic model based approaches which can able extract more semantic keywords. The motivation behind can be concluded into 2 points [3]:

1. Any Keyword/Keyphrase is fall into at least one topic.
2. The extracted Keyword/Keyphrase are comprehensive for representing its main topic.

Topic modeling seems to be a good approach that can be adopted in tagging.

1.4. **Legal domain specific**

Legal documents are usual long and full of domain specific terminology. The screening process of legal documents is costly, especially for law students and normal people. As one of the method of document summarization, automatic document tagging can provide a short conceptual briefing to people. People can quick understand which legal concepts are discussed in the document by looking on the tags assigned
to each search result. It saves time on reading sections of document. Also, a qualified tag should able to guide people refining their search query to make each query more accurate.

2. Objective
This project is going to apply NLP algorithm to build an automatic tagging system specific for legal documents. It aims to assign related legal concept tag with quality to each legal document. In practice, the application of these tags should able to improve user experience of legal documents searching. This project will mainly focus on exploiting Topic-modeling on solving keyphrase extraction tasks.

3. Methodology
3.1. Tag Pools
Since tagging is an information retrieval method, it must satisfy some requirements in Legal information retrieval systems. The requirements are follow [4]:

  Audience: The tags must be in term of short, clear and simple phrases that entry level user will not be confused by tags.

  Legal Terminology: A tag must represent a legal concept used in real world instead of any arbitrary vocab from document.

The tag pool size is fixed and each tag is normalized which means the tag pool should be pre-defined and referencing to real world legal concept. Also, the number of tags assigned to each document should have a maximum limit since this project is aiming to select most representative tags from document.

3.2. Latent Dirichlet Allocation
The most basic and effective topic modeling approach is Latent Dirichlet Allocation (LDA). By assuming there must be multiple topics hidden from document, which are a distribution over fixed word after training. LDA falls into a probabilistic model that computing conditional distribution of the topics by using joint distribution of training words. However, the
simple LDA model assumes the number of hidden topic is fixed which is greatly affected by human’s pertaining setting [5].

3.3. keyATM
Unsupervised topic modeling methods infers topics from text automatically [6], instead of fitting the topics to the existing human pre-defined themes which has been widely used and discussed across any specific study fields. keyATM is referred to Keyword Assisted Topic Models, an open-sourse software package introduced by Shusei Eshima, Kosuke Imai and Tomoya Sasaki. The model is a semi-supervised approach adopted the idea of topic modeling and demonstrated by providing a set of keywords to label topic first instead of finding topics without any prior knowledge, enhances the performance of topic models in practice. [7] keyATM seems to be one possible solution to adopted in automatic tagging documents since clustering by regulation is intuitively possible and make sense to audiences.

3.4. Topic modeling with LEGAL-BERT
LEGAL-BERT, a domain specific pre-trained language models based on Transformer extended from BERT. BERT itself has achieved impressive result in serval NLP task. LEAGAL-BERT is specific built for Legal domain analysis which has shown a significant performance in NLP tasks required legal domain-specific knowledges. [8] A previous research has demonstrated a method with LEGAL-BERT and Topic modeling combined together to perform clustering and generating relevant keywords from each cluster for labeling legal documents. [9]

3.5. Evaluation
LexGLUE provides a benchmark dataset for evaluating model performance on different Legal Language Understand tasks in English. [10] The EUR-Lex dataset included is specific for multi-label text classification that required to predict legal concept by given a textual document. [11] It is a potential evaluation method by comparing our model with the benchmark.
## 4. Schedule

<table>
<thead>
<tr>
<th>Time</th>
<th>Deliverables</th>
</tr>
</thead>
<tbody>
<tr>
<td>13 Oct, 2022</td>
<td>Detailed Project Pan</td>
</tr>
<tr>
<td></td>
<td>Project Website</td>
</tr>
<tr>
<td></td>
<td>Build the tag pool</td>
</tr>
<tr>
<td></td>
<td>Build and test 1 model with dataset</td>
</tr>
<tr>
<td>9-13 Jan 2023</td>
<td>First Presentation</td>
</tr>
<tr>
<td>22 Jan 2023</td>
<td>Preliminary implementation</td>
</tr>
<tr>
<td></td>
<td>Detailed interim report</td>
</tr>
<tr>
<td></td>
<td>Build and test all the possible models</td>
</tr>
<tr>
<td></td>
<td>Analyze the performances</td>
</tr>
<tr>
<td>18 Apr 2023</td>
<td>Finalized tested implementation</td>
</tr>
<tr>
<td></td>
<td>Final report</td>
</tr>
<tr>
<td>17-21 Apr 2023</td>
<td>Final presentation</td>
</tr>
<tr>
<td>3 May 2023</td>
<td>Project exhibition</td>
</tr>
</tbody>
</table>
5. References


