IT550: Information Retrieval

Assignment 3: Document Representation (Part 2)

Instructor: Prasenjit Majumder

Learning Outcome: After this assignment you will learn how to represent document using LSA. You will also learn about matrix decomposition.

1 Problem description

TF-IDF is a discrete representation of term and document. This approach does not take the contextual information of a word into consideration while representing a term. So for this purpose we make use of Latent Semantic Analysis (LSA) which is a distributed representation technique for representation of a term. With this approach it is possible to represent a document or term in low dimensional vector space.

2 Latent Semantic Analysis (LSA)

LSA makes use of term-document matrix which is obtained using tf-idf. This term-document matrix is sparse. LSA applies low rank matrix decomposition technique (SVD) that projects the documents to a low dimensional vector space. Similar documents can be grouped together by calculating the cosine similarity between the documents. The term-document matrix (X) after using Truncated SVD is shown in equation 1

$$X_k = U_k \Sigma_k V_k^T \tag{1}$$

- Here U_k contain the term vectors. V_k contain the document vectors.
- The document vector for a new document can be calculated using equation 2

$$\hat{d_j} = \Sigma_k^{-1} U_k^T d_j \tag{2}$$

- The term vector for a new term can be calculated using equation 3

$$\hat{t_i} = t_i^T V_k \Sigma_k^{-1} \tag{3}$$

3 Implementation

3.1 Dataset

- For this assignment we will use Telegraph news articles, which is in XML format. It contains news on different categories for the year 2004 to 2007. You can download the dataset from this link: https://drive.google.com/open?id=1JuawXQmYVkjpfL3HOblqjDrqw8V1lHrC
- The Queries are in "en.topics.76-125.2010". The query is of the format shown in Figure 1. Use the sentences enclosed in desc tag for framing your query vector
- The "en.qrels.76-125.2010.txt" contains the documents that are relevant to a query. The format of a qrel is such: Query_No Q0 Document ID Relevance score.
- Relevance score is binary 0 or 1. 1 is for relevant, 0 is for otherwise.
- The documents in the dataset is in the format shown in Figure 2.

```
<top lang='en'>
<num>76</num>
<title>Clashes between the Gurjars and Meenas</title>
<desc>
Reasons behind the protests by Meena leaders against the
inclusion of Gurjars in the Scheduled Tribes.
</desc>
<narr>
The Gurjars are agitating in order to attain the status of a
Scheduled Tribe. Leaders belonging to the Meena sect have
been vigorously opposing this move. What are the main reasons
behind the Meenas' opposition? A relevant document should
mention the root cause(s) behind the conflict between these
two sects.
</narr>
</top>
```

Figure 1: Query Format

```
<DOC>
<DOCNO> </DOCNO>
<TEXT></TEXT>
</DOC>
```

Figure 2: News Format

3.2 Exercise

1. In this assignment you will perform LSA on all documents in the dataset.
2. Follow the steps in (http://intranet.daiict.ac.in/~daiict_nt01/Lecture/Prasenjit%20Majumder/IT550/LSI-Eg.pdf) for creating your LSA algorithm
3. Use the queries and all the documents and calculate the cosine similarity between the query and the document. Retrieve the top 5 similar documents.
4. List the documents that you have retrieved for a query and the actual relevant document.
5. How many relevant document did your system retrieve?

4 References

- An Introduction to Information Retrieval: Christopher D.Manning ,Prabhakar Raghavan, Hinrich Schütze
- https://en.wikipedia.org/wiki/Latent_semantic_analysis

5 Submission

- You have to submit your assignment in Jupyter notebook with proper comments and explanation of your approach.
- Your output should contain the query id, the top 5 document you retrieved and the relevant document associated with it.
- In the end you will have to specify how many queries fetched the relevant document.
- The submission deadline for this assignment in **10th Feb 2020 at 11 PM**