CSI4107 - Information Retrieval
WInter 2019

Search Engine Programming Project

Modules descriptions and requirements for FINAL system

VERSION 1.0
February 25th 2019

This document contains a short description for each module to be included in your Search Engine.

You might need additional modules as the "glue" to your system. That is fine. Just make sure to keep your design as modular as possible.

Some modules are "underspecified" in the descriptions below, which gives you freedom to implement them the way you want. Think flexible, modular and efficient.

Module 8a - Bigram Language Model

Purpose	Build a Bigram Language Model
Input	A collection of documents (Reuters)
Output	A bigram language model
Modules depending on this module.	Query Completion Module
Modules required by this module.	Corpus pre-processing.
YOUR choice	Perhaps you wish to build the language model with pre-processing of stopwords, stemming, etc.

Module 8b - Query Completion Module

Purpose	Provide suggestions to the user as a list of possible completions to their query.
Input	Bigram Language Model + Word(s) typed by user.
Output	A set of suggestions to the User.
Modules depending on this module.	Depends on your architecture, but nothing should depend on this, only the UI will have to make sure to send the selected choice to the retrieval model.
Modules required by this module.	Bigram Language Model.
YOUR choice	1. You can decide how you want to display the choices to the user, as well as how many choices to display. 2. You can also decide how this module gets activated, for example when the user types a space after a word.

Module 9a - Automatic Thesaurus Construction

Purpose	Build a thesaurus from the Reuters Collection.
Input	Reuters Collection. Dictionary of Terms.
Output	A structure that captures the similarity between all pairs of terms.
Modules depending on this module.	The Query Expansion module will need this thesaurus.
Modules required by this module.	Dictionary building module. You need the dictionary.
YOUR choice	1. You must decide on a similarity measure (either Jaccard, or Cosine). You must evaluate the similarity of all pairs of terms using this measure. 2. You must also decide on the "unit of comparison" (either document, sentence, paragraph).

Module 9b - Global Query Expansion (in VSM)

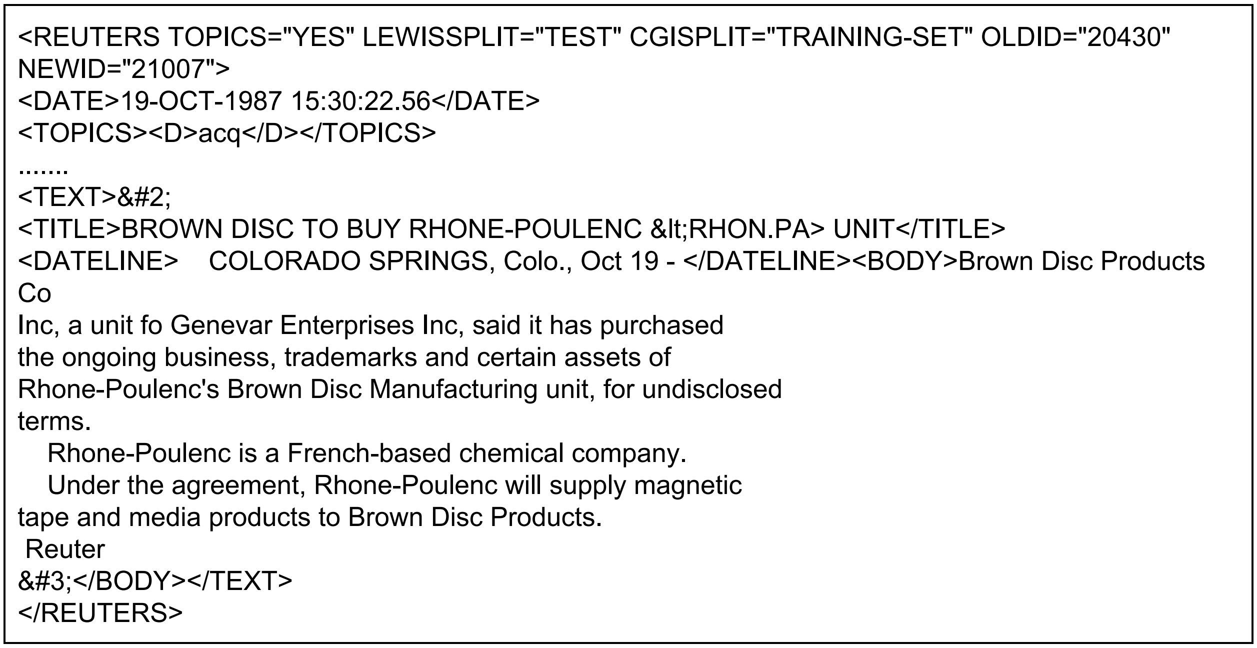
Purpose	Perform either explicit or implicit query expansion using the thesaurus.
Input	Thesaurus + Word typed by user.
Output	Expanded Query.
Modules depending on this module.	VSM will receive the expanded query.
Modules required by this module.	Thesaurus construction module. UI.
YOUR choice	1. How to perform the expansion. Explicit (showing the terms to the user and allowing user to choose) or Implicit (you perform the expansion without the user knowing). 2. If expansion is implicit, how to set the expansion limit (how many terms to expand) 3. If the user query contains more than one word, decide how you want to combine the similarities.

Module 10a - Text categorization with kNN

Purpose	Assign one or more topics to the Reuters documents that are not classified.
Input	Collection of Reuters document, with a subset of documents containing topics. See Figure 1 below, in which the topic is assigned. This will be part of the TRAINING set.
Output	All of the Reuters collection with assigned topics. See Figure 2 below, with topic NOT assigned. With this module, the document in Figure 2 should also have an assigned topic.
Modules depending on this module.	The topic restriction module will need the set of topics assigned to a document.
Modules required by this module.	None.
YOUR choice	1. Decide on the k, in the kNN algorithm, and how you want to deal with the combination of retrieved topics. 2. Decide on the similarity measure used between the documents. 3. Decide on how to manage multiple topic assignment (a document can have different topics... how to deal with that... come up with a strategy).

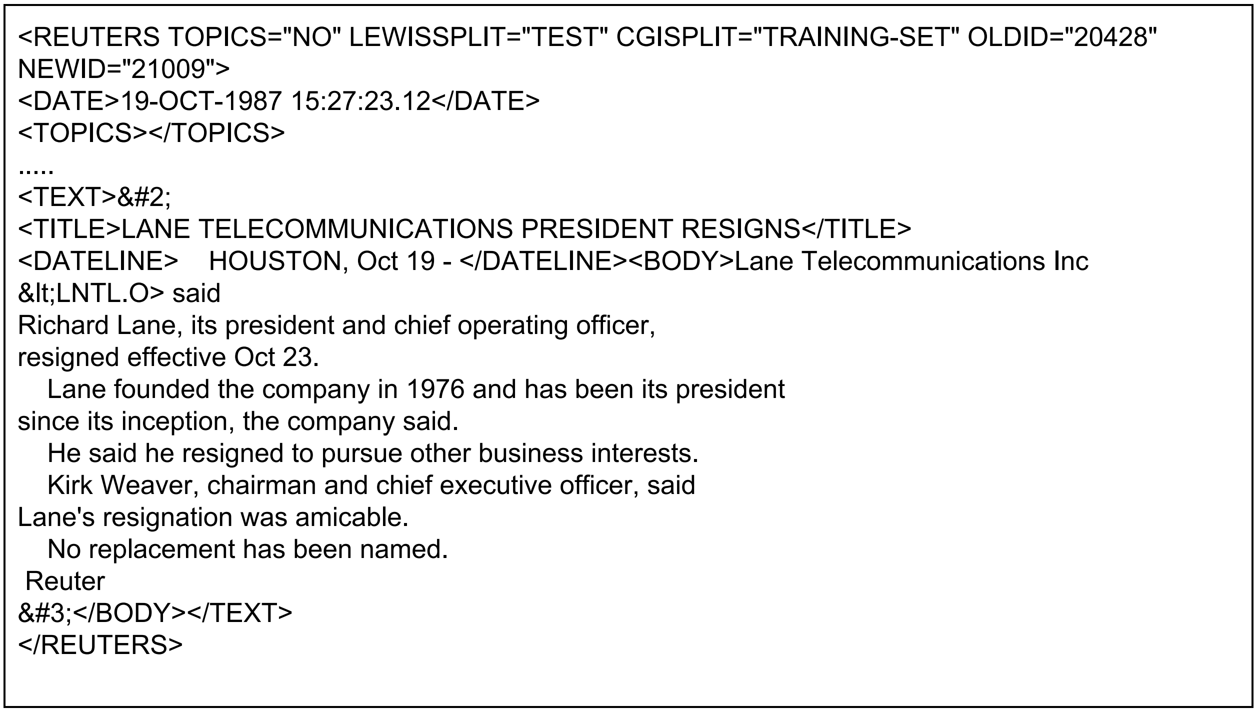
Module 10b - Topic Restriction

Purpose	Allow the user to only retrieve documents pertaining to certain topics.
Input	Choice of topic(s).
Output	Restriction for the retrieval models.
Modules depending on this module.	Depending on your architecture, the filtering of topic can affect the retrieval models directly, or it can be dealt with using a separate filtering module.
Modules required by this module.	Text categorization module, to make sure all reuters documents are assigned topics.
YOUR choice	1. Decide how to modify your UI to be able to perform the topic selection.



```
<REUTERS TOPICS="YES" LEWISSPLIT="TEST" CGISPLIT="TRAINING-SET" OLDID="20430"
NEWID="21007">
<DATE>19-OCT-1987 15:30:22.56</DATE>
<TOPICS><D>acq</D></TOPICS>
.......
<TEXT>&#2;
<TITLE>BROWN DISC TO BUY RHONE-POULENC &lt;RHON.PA> UNIT</TITLE>
<DATELINE>    COLORADO SPRINGS, Colo., Oct 19 - </DATELINE><BODY>Brown Disc Products
Co
Inc, a unit fo Genevar Enterprises Inc, said it has purchased
the ongoing business, trademarks and certain assets of
Rhone-Poulenc's Brown Disc Manufacturing unit, for undisclosed
terms.
    Rhone-Poulenc is a French-based chemical company.
    Under the agreement, Rhone-Poulenc will supply magnetic
tape and media products to Brown Disc Products.
 Reuter
&#3;</BODY></TEXT>
</REUTERS>
```

Figure 1 - Example of Reuters document with topic assigned (TRAINING)



```
<REUTERS TOPICS="NO" LEWISSPLIT="TEST" CGISPLIT="TRAINING-SET" OLDID="20428"
NEWID="21009">
<DATE>19-OCT-1987 15:27:23.12</DATE>
<TOPICS></TOPICS>
.....
<TEXT>&#2;
<TITLE>LANE TELECOMMUNICATIONS PRESIDENT RESIGNS</TITLE>
<DATELINE>    HOUSTON, Oct 19 - </DATELINE><BODY>Lane Telecommunications Inc
&lt;LNTL.O> said
Richard Lane, its president and chief operating officer,
resigned effective Oct 23.
    Lane founded the company in 1976 and has been its president
since its inception, the company said.
    He said he resigned to pursue other business interests.
    Kirk Weaver, chairman and chief executive officer, said
Lane's resignation was amicable.
    No replacement has been named.
 Reuter
&#3;</BODY></TEXT>
</REUTERS>
```

Figure 2 - Example of Reuters document with topic NOT assigned. Should be assigned by your Text Categorization Module.

Optional Module - Query Expansion with WordNet

Purpose	Use an external thesaurus (WordNet) to perform Query Expansion.
Input	WordNet + Word typed by user.
Output	Expanded Query.
Modules depending on this module.	VSM and Boolean Retrieval Model will receive the expanded query. This will not work for probabilistic model.
Modules required by this module.	UI to obtain the query word(s).
YOUR choices.	1. How to perform the expansion. Explicit (showing the terms to the user and allowing user to choose) or Implicit (you perform the expansion without the user knowing). 2. Decide when to apply this expansion. If a term has multiple definitions... might not be appropriate. 3. Decide if you want to include only synonyms or hypernyms as well. 4. If expansion is implicit, decide how to set the expansion limit (how many terms to expand), and how to weigh (in VSM) the added terms. 5. If the user query contains more than one word, decide how you want to perform the expansion.

Optional Module - Visualization of automatic thesaurus

Purpose	Provide the user with a visual map of related terms.
Input	Automatically built thesaurus (module 9a) + User Query word.
Output	In the UI, a graphical view of the thesaurus.
Modules depending on this module.	UI
Modules required by this module.	Automatically built thesaurus (module 9a)
YOUR choice	Make sure you present the information so that the similarity between terms is captured in your graphical representation. As a reference, you can look at the semantic maps that I showed in the slides on Query Expansion (use case of CV search).

Optional Module - Local Query Expansion in Vector Space Model

Purpose	Perform implicit query expansion using the relevance provided, within the VSM, using the Rocchio Algorithm.
Input	Relevance + Word typed by user.
Output	Expanded Query
Modules depending on this module.	VSM.
Modules required by this module.	Relevance Gathering module (if included in Vanilla System), otherwise perform Pseudo-Relevance assuming the top N documents are relevant.
	1. The use of the expansion could be programmed as a parameter perhaps to the VSM. 2. Since the expansion is implicit, how to set the expansion limit (how many terms to expand) 2. If the user query contains more than one word, decide how you want to combine the similarities.

Optional Module - Dynamic document Clustering with adapted UI

(Counts for 2 modules)

Purpose	Cluster the documents and present the clustered documents to the user
Input	Document collection. + User Query.
Output	Clusters among the retrieved documents.
Modules depending on this module.	The UI.
Modules required by this module.	
YOUR choices	This counts for 2 modules as there is a lot to do, but you also have a lot of freedom. Provided a Query Q, the retrieved set of documents D must be partitioned (clustered) among N clusters. You can then show the clusters to the user, or use the clusters to ask "More like this?" providing the centroid of each cluster (or even an automatically generated label for each cluster).

Optional Module - Text categorization with Naive Bayes

Purpose	Assign one or more topics to the Reuters documents that are not classified using a Naive Bayes classifier trained on the Reuters documents which are assigned topics.
Input	Collection of Reuters document, with a subset of documents containing topics (to be used as training set).
Output	All of the Reuters collection with assigned topics.
Modules depending on this module.	The topic restriction module will need the set of topics assigned to a document.
Modules required by this module.	None.
YOUR choices	If you implement this module, the topic restriction (module 10b) can then be based on this categorization or the kNN categorization. Provide the choice as a parameter in your UI.