CS6220 Unsupervised Data Mining

HW5 Topic Models, Summarization, Elastic Search

Make sure you check the syllabus for the due date. Please use the notations adopted in class, even if the problem is stated in the book using a different notation.

We are not looking for very long answers (if you find yourself writing more than one or two pages of typed text per problem, you are probably on the wrong track). Try to be concise; also keep in mind that good ideas and explanations matter more than exact details.

Submit all code files Dropbox (create folder HW1 or similar name). Results can be pdf or txt files, including plots/tabels if any.

"Paper" exercises: submit using Dropbox as pdf, either typed or scanned handwritten.

DATATSET **: 20 NewsGroups : news articles**

DATATSET **: DUC 2001 summarization dataset**

https://www-nlpir.nist.gov/projects/duc/guidelines/2001.html
(can be found in "DM resources")

Elastic Search

You will need to install ElasticSearch and corresponding plugins in order to manipulate and visualize text
server: https://www.elastic.co
visualization, control: https://www.elastic.co/downloads/kibana
API for Java: https://www.elastic.co/guide/en/elasticsearch/client/java-api/6.2/index.html
API for Python: https://elasticsearch-py.readthedocs.io/en/master/
API for Perl: http://search.cpan.org/dist/Search-Elasticsearch/lib/Search/Elasticsearch.pm

PROBLEM 1: Text Indexing

Index each dataset separetely in Elastic Search (one index for each dataset). First set up the indexes/types/fields in Kibana, then use an API to send all docs to the index. At the minimum you will need two fields: "doc_id", and "doc_text"; you can add other fields. For DUC dataset add a field "gold_summary".

PROBLEM 2: Topic Models

Obtain Topic Models (K=10, 20, 50) for both datasets by running LDA and NMF methods; you can call libraries for both methhods and dont have to use the ES index as source. For both LDA and NMF: print out for each topic the top 20 words (with probabilities)

The rest of of topic exercises and results are required only for the LDA topics:
- 20NG: how well the topics allign with the 20NG label classes? This is not asking for a measurement, but rather for a visual inspection to determine what topics match well with what classes. Does this cnage if one increases the topics from 20 to 50?
- ES: Add a type or inew index "topic" with fields "topic_id" and "top_words" to store for each topic the top 10 words with associated probabilities.
- ES: Add a field for documents "doc_topics" and update the index to store for each document the most important topics (up to 5) and doc-topic probabilities

PROBLEM 3: Extractive Summarization

Implement the KL-Sum summarization method for each dataset. Follow the ideas in this paper ; you are allowed to use libraries for text cleaning, segmentation into sentences, etc. Run it twice :
A) KL_summary based on words_PD; PD is a distribution proportional to counts of words in document
B) LDA_summary based on LDA topics_PD on obtained in PB2. The only difference is that PD, while still a distribution over words, is computed using topic modeling
- ES: Add two new fields to the document type, "KL_summary" and "LDA_summary" to store the obtained summaries.

For DUC dataset evaluate KL_summaries and LDA_summaries against human gold summaries with ROUGE. ROUGE Perl package

EXTRA CREDIT. KL Summarization: Can we make both PD and PS distributions over topics, instead of distributions over words? Would that help?

PROBLEM 4: Simple Sampling

You are not allowed to use sampling libraries/functions. But you can use rand() call to generate a pseudo-uniform value in [0,1]; you can also use a library that computes the pdf(x|params). make sure to recap first Rejection Sampling and Inverse Transform Sampling .
A. Implement simple sampling from continuous distributions: uniform (min, max, sample_size) and gaussian (mu, sigma, sample_size)
B. Implement sampling from a 2-dim Gaussian Distribution (2d mu, 2d sigma, sample_size)
C. Implement wihtout-replacement sampling from a discrete non-uniform distribution (given as input) following the Steven's method described in class (paper). Test it on desired sample sizes N significantly smaller than population size M (for example N=20 M=300)

PROBLEM 5: Conditional Sampling

Implement Gibbs Sampling for a multidim gaussian generative joint, by using the conditionals which are also gaussian distributions . The minimum requirement is for joint to have D=2 variables and for Gibbs to alternate between the two.

Extra Credit: Implement your own LDA using Gibbs Sampling, following this paper