HW for 2019-05-29

(due: 2019-06-04)

1: Latent semantic indexing in NIPS The UCI repository of data sets for machine learning includes several "bag of words" examples:

https://archive.ics.uci.edu/ml/datasets/Bag+of+Words

We are going to use latent semantic indexing to look for words in the NIPS data set (in the files docword.nips.txt.gz and vocab.nips.txt)

Using the code in the repository or your own code, apply latent semantic indexing to find the top five documents most relevant to the query term "circuit." To do this, you should

- Load the data into a matrix.
- Normalize the raw word counts into TF-IDF scores.
- Compute the truncated SVD (use 20 factors).
- Find the index in the vocabulary associated with the word "circuit."
- Compute the scores associated with the word "circuit."
- Sort the documents in ascending order by scores.
- Print the top words associated with the leading documents.

We have provided Octave codes for loading the data (load_docwords), computing TF-IDF scores (tf_idf), and showing the top words associated with a document (show_top_words). You will want to use the *sparse* SVD command (e.g. svds) to compute the dominant part of the singular value decomposition. Note that in the code provided, we have one document per row, unlike in the notes where we used one document per column.

2: Clustering by ID Consider a data set consisting of k clusters in n-dimensional space with $n > k$. These clusters might conventionally be recovered by k-means iteration, but if they are sufficiently well separated, they can also be computed via an interpolative decomposition; if $A \in \mathbb{R}^{n \times m}$ is the data set (consisting of m) points, then we have the approximate rank k decomposition

$$A \approx CT$$

where $C \in \mathbb{R}^k$ is a subset of representative columns of A, and the location of the largest element in each column of T indicates the cluster to which the corresponding column of A belongs. Complete the code in cluster_demo to illustrate that this works (note that computing via pivoted QR without later clean-up is fine).