Concept decompositions for large sparse test data using clustering

Idea of this project is to cluster collection of text data using k-means algorithm on vector space model. Specifically, we interpret n text documents as d-dimensional vectors of certain weighted word frequencies which we cluster. Using obtained results we approximate corresponding word-by-document matrix.

We distribute our project goals over 3 main parts: Vector space model creation, Cluster construction, Decomposition analysis.

Vector space model creation

There are 2 main subtasks in order to create a vector space model:

[1] Data extraction

[2] Model creation

[1] We used unordered\_map container from C++ std package to build the hash-table of depth 2 for storing words and vectors of their occurrence (matrix row). This container shows the best performance over the number of words and search queries we use to count them all.

There are 5 main steps in our data extraction, given bellow in respective order:

1. Extracting all unique words from a set of documents

(ignoring case, ignoring non-alphabetic terms)

2. Eliminate non-content-bearing ‘stop-words’

(list of stop-words we used can be found in Frakes and Baeza-Yates, 1992, Chapter 7)

3. Stemming words to their root

(we used a Porter stemmer found in Frakes and Baeza-Yates, 1992, Chapter 8)

4. For each document, count the number of occurrences of given word

(full benefits of unordered\_map were exploited)

5. Eliminate high/low-frequency words

(this can be chosen in user interface)

[2] After Data Extraction, only d words remain and we continue by creating a d\*n word-by-document matrix. After performing txn (normalized term frequency) algorithm on the previously mentioned matrix, we have normalized document vectors which we use in k-means algorithm that follows.

Cluster construction

Decomposition analysis

Using the centroids and clusters, we create a new matrix which approximates matrix from the beginning. This process is called - concept decomposition – computing the least-squares approximation onto linear subspace spanned by the concept vectors.

We calculated another approximation by k-truncated SVDs to compare which decomposition approximates the matrix better. We measure the error in approximating using the Frobenius norm of the difference matrix.

RESULTS:

We test our work on a set of text-books and fantastic novels to show how they are correctly clustered…..

Furthermore, matrix decomposition obtained by above clustering show error of only \_\_\_ compared to SVD error of \_\_\_