Exercise 1  For your personal collection from the beginning of the course — or any other collection you might prefer by now — come up with the following:

1. a list of 3, 4, 5, or more natural-language queries that make sense on your collection;
2. for each such query, a reasonable set of keywords, as one would type them into a conventional search engine;
3. and for each query, a list of relevant documents, as exhaustive as possible (think of a way to get such a list without actually going through each document, especially if you have a large number of documents).

Your collection should contain a few hundreds or thousands of documents. The absolute minimum is 100; if your old collection is smaller than this, either add some documents, or come up with a new collection. Also, each query should have at least 10 relevant documents, and you should have at least one query with relatively few relevant documents, as well as one query with relatively many relevant documents.

Exercise 2  For each of your queries from Exercise 1, compute a ranking of the documents by taking as similarity scores simply the cosine of the angle between the respective query vector and each of the documents. Then compute your quality measure from the last exercise sheet — or any other quality measure you might prefer by now — for each of these rankings individually, as well as the average over all queries.

Exercise 3  Repeat Exercise 2, but this time using Latent Semantic Indexing (LSI) for computing similarity scores, that is, queries and documents are mapped to a space of some relatively small dimension $k$, and (cosine) similarities are then computed in this low-dimensional space. The choice of $k$ is up to you.

Exercise 4  Prepare a short (5-10 minutes) presentation on your results. Don’t forget to mention the main characteristics of your collection: where you got it from, about what it is, the number of documents, the number of terms, etc. Also comment on how reasonable you find the results you got.