December 4, 2013

# CSCI 2830
# Solutions to Application 2, part 2: Refining the Search Engine

## The Assignment

Using your revised search engine, test the following queries for ranks 1, 2, 3, 4:

$q^{(1)}$ Baking bread.

$q^{(2)}$ Baking.

$q^{(3)}$ What is the difference between pastry and bread?

$q^{(4)}$ Show me all recipes for pastry.

$q^{(5)}$ A query of your devising. (Specify.)

Summarize your results in a clear and well-designed table.

What matrix rank and cosine cutoff do you recommend for a search engine based on the vector space model using the QR decomposition for rank reduction? Your search engine will likely not operate perfectly, so you'll need to identify the best compromise between irrelevant documents returned and relevant documents missed.

Your choices should be defended with about half a page of clear, consistent, and well-reasoned argument. Perfect spelling and reasonable grammar are expected. It is ok to repeat things you wrote in the first part of the assignment to support your argument.

**Solution:** For full credit on this problem, you need to have experimented with all 5 queries and provided a reasonable explanation of your results. You can recommend any rank for the rank reduction as long as you argue convincingly that your choice makes sense.

A rank of 3 works as well as any other for the provided queries. It clears out some of the irrelevant documents that are observed for rank 4, and does not lead to as many nonsensical results as does rank 2. The query you invented may be the deciding factor. Rank 1 results in a matrix with unfortunately placed zeros that may lead to NaNs in the course of the computation.

In practice, a factorization known as the singular value decomposition is used in place of QR. A rank of 100 is used, even for enormous data sets. This method is famous as *lexical semantic analysis (LSA)* or *lexical semantic indexing*, but I am not convinced that it really works all that well! Jim Martin and I wrote one more scientific assessment of LSA that indicates that it is not worth its level of fame.