Shortcomings of lexical-based similarity measures

- Invariant to sequence order: randomly shuffling the order of the words in each document yields the same result.

- Doesn’t capture word synonymy, polysemy, or context

- Highly sensitive to text processing and tokenization scheme that is used

- Given the sparsity of the term document matrix, should we expect any distance-based similarity metric to be meaningful?

Low rank approximation of the term-document matrix

- We saw in the TF-IDF demo that documents can be characterized by its most heavily TF-IDF weighted words

- Idea: what if we were to project a document onto a set of basis' comprised of linear combinations of the constituent words such that the basis themselves capture semantic relationships between words?

- How would we learn such a transformation?
  - PCA?
  - SVD?

- Why is PCA not suitable for this problem?

- How many dimensions would we need for our sub manifold to capture most of the information in our sparse term-document matrix?