Further Improvement of VSM?

- Improved instantiation of **dimension?**
  - stemmed words, stop word removal, phrases, latent semantic indexing (word clusters), character n-grams, …
  - bag-of-words with phrases is often sufficient in practice
  - Language-specific and domain-specific tokenization is important to ensure “normalization of terms”
- Improved instantiation of **similarity function?**
  - cosine of angle between two vectors?
  - Euclidean?
  - dot product seems still the best (sufficiently general especially with appropriate term weighting)