Information Retreival

* Mikolov used fixed-length paragraphs from first 10 results from a million popular queries
* Two passages from the same query, a third froma different query
* Use paragraph vectors to compute distances between the paragraphs
* Genism has similar\_vector function that does this using cosine similarity.
* Two similar documents should have a closer cosine similarity
* DM was beter than DBOW, but better with both.
* Concatenation in DM is better than sum – model loses ordering information
* Window size 5-12 error rate to fluctuate .7%
* Dir