Vector space model

1. Biểu diễn

Các document và câu hỏi được lưu dưới dạng vector.

di,j= (wi,1, wi,2 , .. wi,t)

trong đó: wi,j: trọng số của từ j trong văn bản i

Sự giống nhau của môt văn bản và một câu hỏi là độ đo cosine giữa chúng.

1. Độ đo

* Cosine
* Term weighting
* Xếp nhóm tương tự (Clustering Analogy)
* Tần số thuật ngữ(Term frequency)
* Tần số văn bản ngược (Inverse Document Frequency)
* Tf-idf weighting
* Tính đơn điệu của tf-idf (tf-idf monotonicity)
* Tiêu chuẩn độ dài