



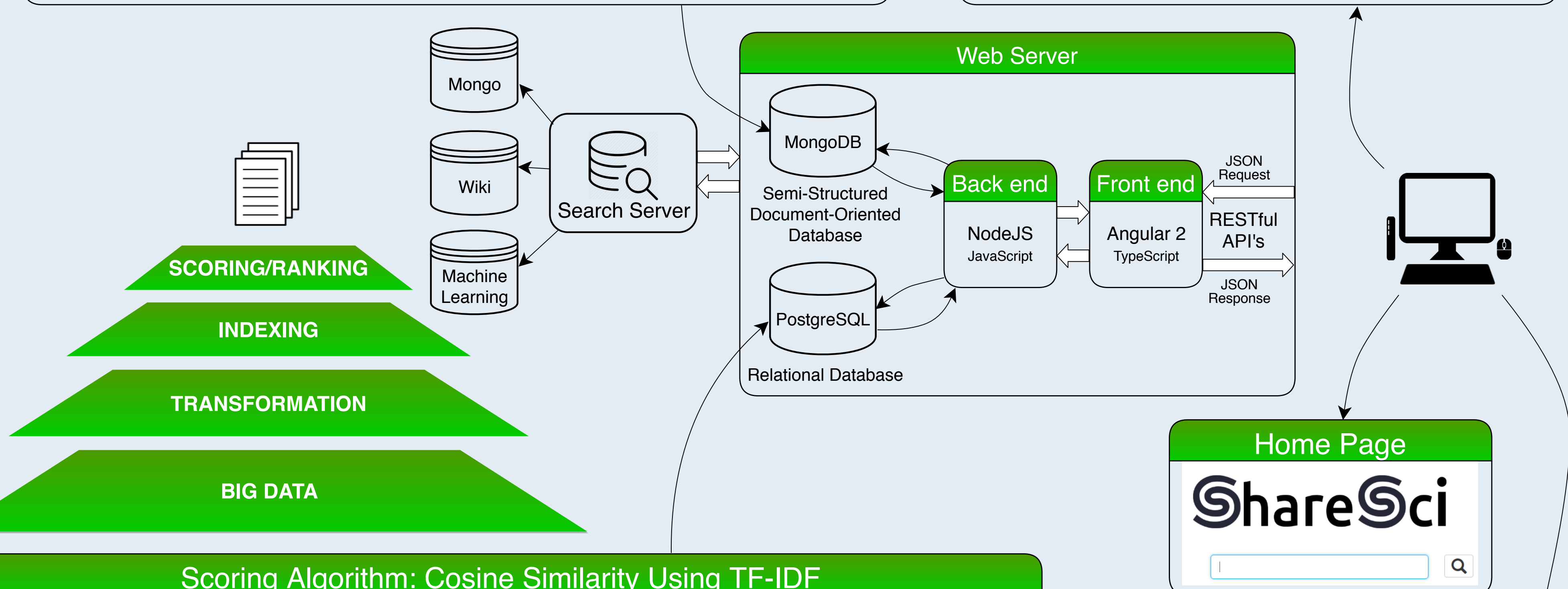
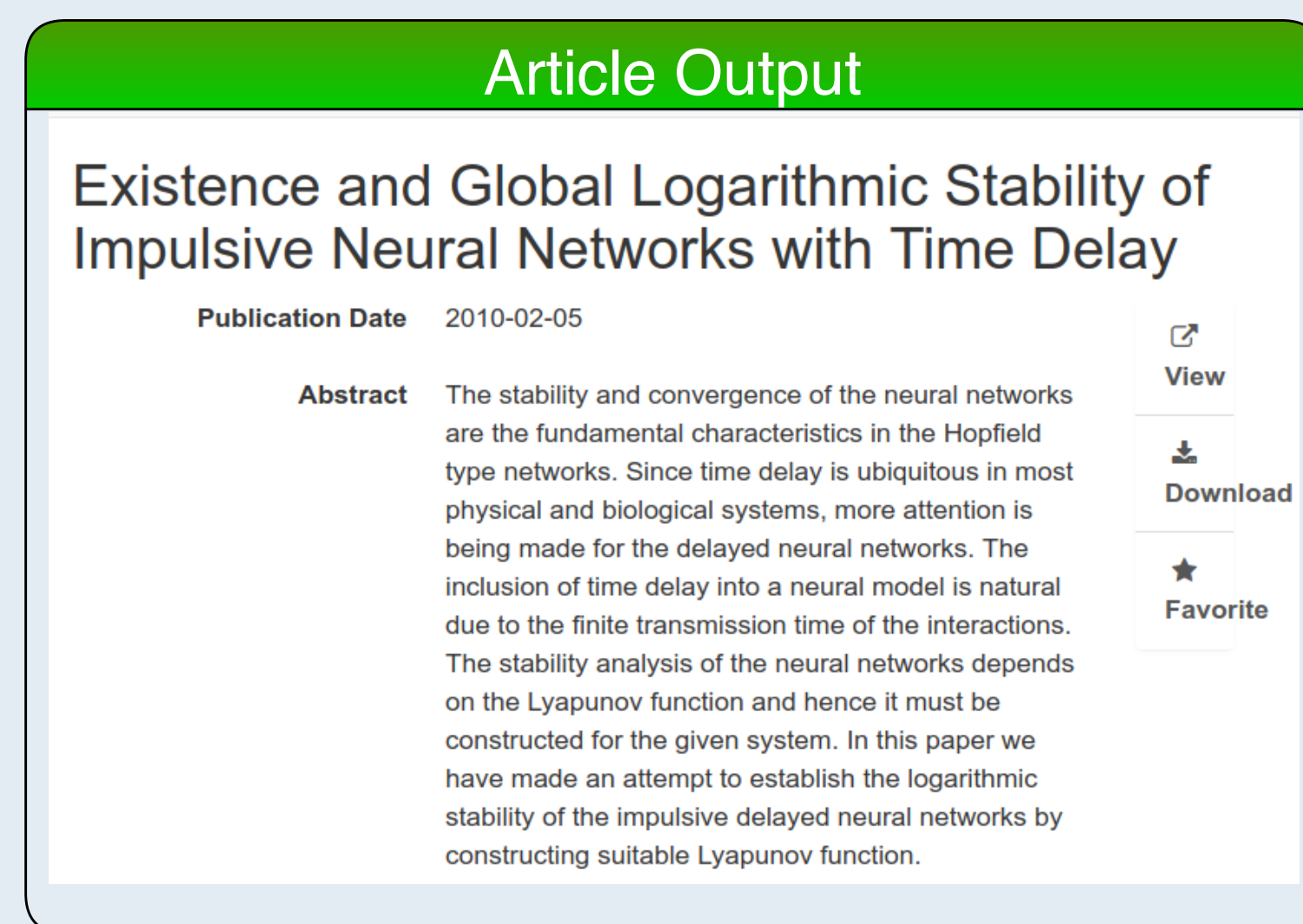
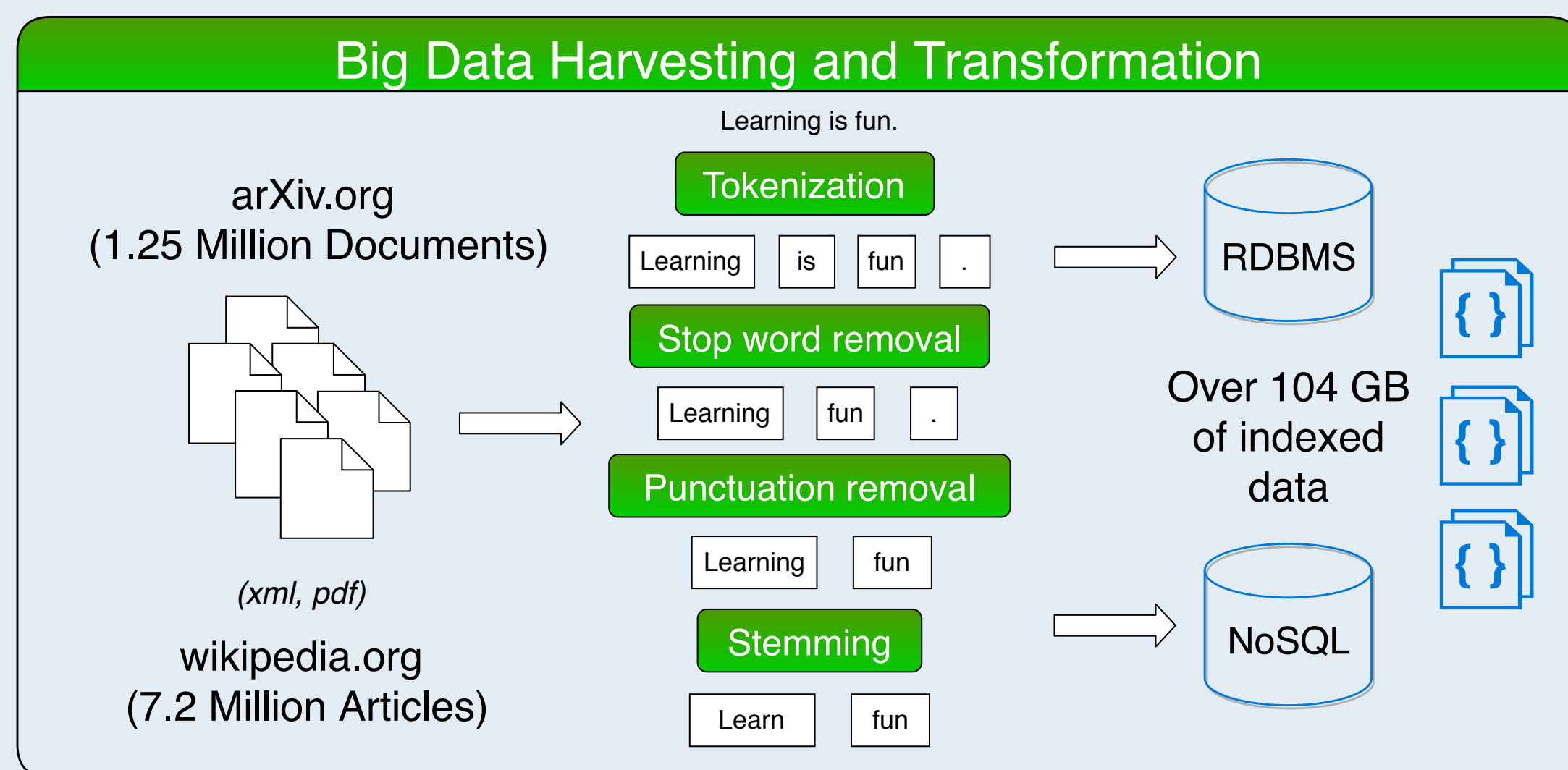
Document Search Engine for Collaborative Content Management

Nick McCoy, Austin Balogh, John Wolf, Benjamin McKinsey

Advisor: Dr. Sunnie S. Chung

Department of Electrical Engineering and Computer Science

ShareSci



Scoring Algorithm: Cosine Similarity Using TF-IDF

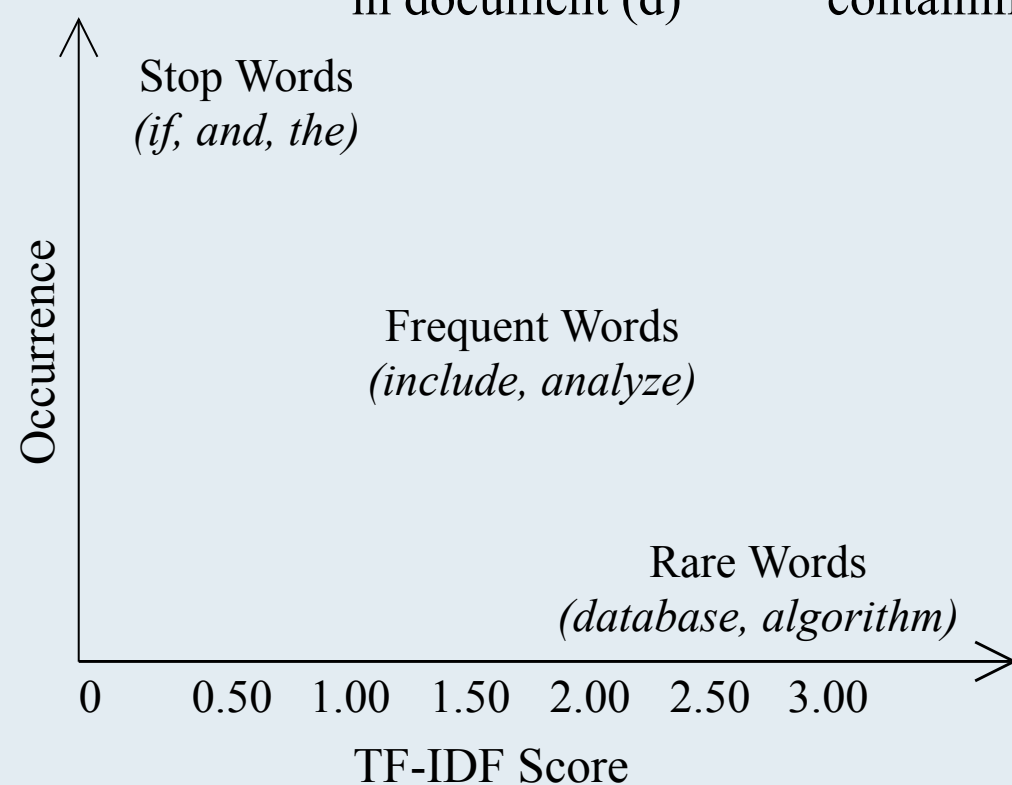
Term Frequency - Inverse Document Frequency

The tf-idf weight of a term is the product of its tf weight and its idf weight

$$W_{t,d} = \log(1 + tf_{t,d}) \times \log_{10}(N/df_t)$$

$$\text{TF-IDF}(t,d) = \text{TF}(t,d) \times \text{IDF}(t)$$

Value of term (t) in document (d) Frequency of term (t) in document (d) Relevance of term (t) in set of documents containing term

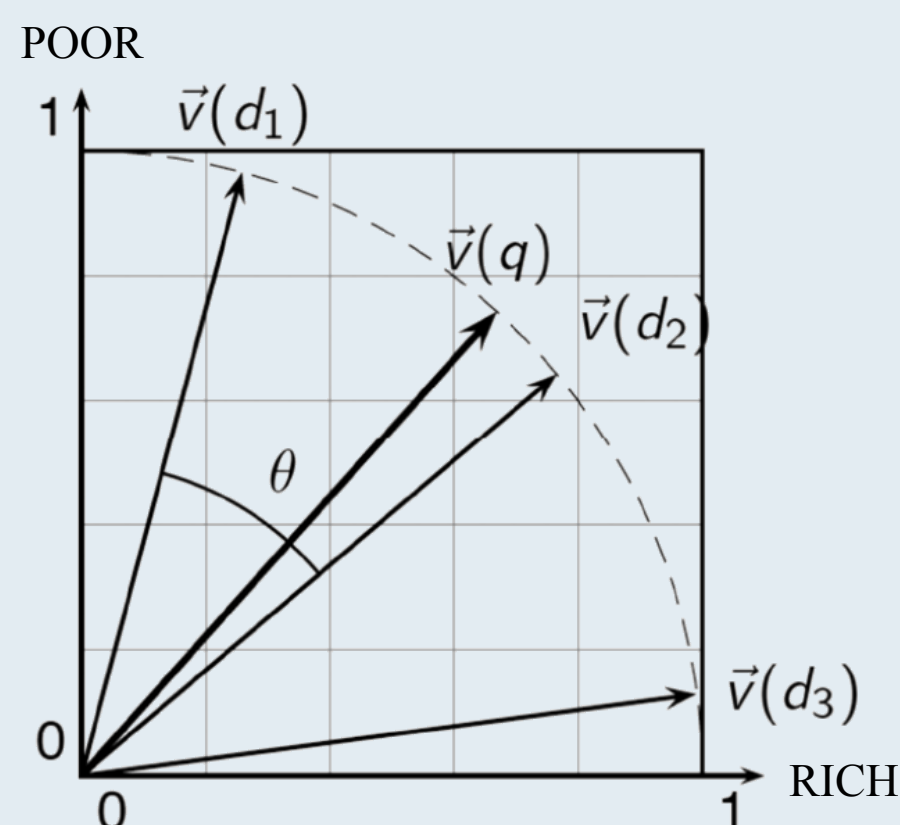


Cosine Similarity

Describes the similarity between documents

$$\cos(\vec{q}, \vec{d}) = \frac{\vec{q} \cdot \vec{d}}{|\vec{q}| |\vec{d}|} = \frac{\vec{q}}{|\vec{q}|} \cdot \frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2} \sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

q_i is the tf-idf weight of term i in the query
 d_i is the tf-idf weight of term i in the document



$$\text{Score}(q, d) = \sum_{t \in q \cap d} tf \cdot idf_{t,d}$$

