Vector. Representation of Docs.

Cosine Distance

Latent Semantic Analysis. (LSA) ⇒
- Discover Latent topics
- Dimension Reduction.
- Represent words in terms of topics

Vector Notation.

vec(d) of d.: V dimensional Vector $\qquad W_{i(M),d} = f_{i(M)} \times IDF(i(M))$

Document Vector: $(0, \cdots 0, w_{i(1)d}, 0 \cdots 0, w_{i(2)d}, \cdots\cdots w_{i(M),d}, 0, \cdots 0)$

- $D = \{d_1, \cdots d_N\}$ set of docs contains N documents
- No. of different words in. D: V (vocab size).
- one document d: M terms $t_{i(1)}, \cdots t_{i(M)}$

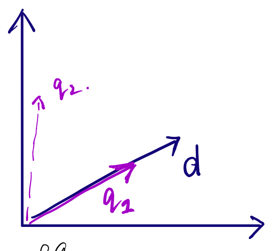
  ↳ $f_{i(M)}$ is the frequency

if d_1, d_2 are docs

- $vec(d_1) = vec(d_2) \Leftrightarrow d_1 = d_2$?
- $vec(d_1) = \underline{\underline{\lambda}}\, vec(d_2)$

  same proportion.
  same words.

Document Length $\quad Len = \sqrt{\Sigma w^2}$.



eg. $d = \lambda q_1$

q_1. d: Same words. same proportion.

q_2. d: different words

⇒ greater angle: vec(d) and vec(q)

less similar bextween q and d.

Cosine Similarity

- Doc d and query q.

dimension = N

$$\mathrm{CSim}(q, d) = \cos\theta = \frac{\mathrm{vec}(q) \cdot \mathrm{vec}(d)}{\|q\| \cdot \|d\|} = \frac{\sum_{t \in q \cap d} w_{tq} \cdot w_{td}}{\|q\| \cdot \|d\|}$$

$$= \mathrm{Sim}(q, d)$$

- Doc d_1 and d_2

$$\mathrm{CSim}(d_1, d_2) = \cos\theta$$