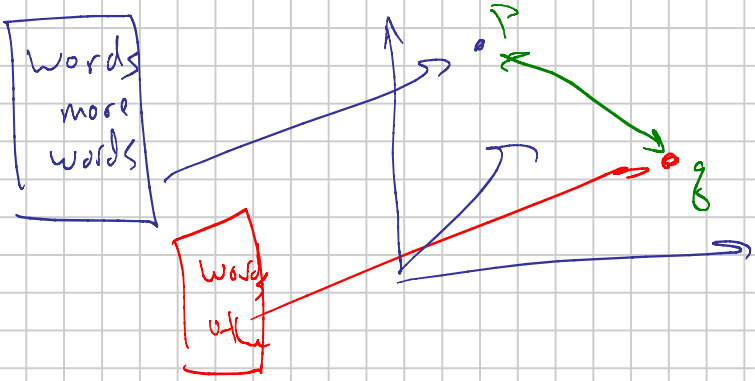
Jaccard Similarity + k-Grams

Note Title 1/20/2016

Which documents are similar?

- homework → cheating
- webpages → Google top 10 pages are distinct
- email → spam



$$\text{Distance}(p, q) = \sqrt{\sum_{i=1}^{d} (p_i - q_i)^2}$$

$p \in \mathbb{R}^d$

$p = (p_1, p_2, \ldots, p_d)$

$\{(p_2, p_1, \ldots, p_d)\}$ $\neq$ $s = 1$

Sets $\{a, b, c\} = \{a, c, b\} = \{a, a, b, c\}$

Distance

$d(A, B)$

small if A, B close

$d \to [0, \infty)$

$A = B \Rightarrow d \to 0$

Similarity

$s(A, B)$

large if A, B close

$s \to [0, 1]$

$A = B$ $s \to 1$

$$d(A, B) = 1 - s(A, B)$$

$$d(A, B) = \sqrt{s(A, A) + s(B, B) - 2s(A, B)}$$

Jaccard Similarity

$A = \{1, 2, 3\}$ $B = \{2, 3, 4, 6\}$

$$JS(A,B) = \frac{|A \cap B|}{|A \cup B|} =$$

$$= \frac{|\{2,3\}|}{|\{1,2,3,4,6\}|} = \frac{2}{5}$$

---

symmetric difference $A \Delta B = (A \cup B) \setminus (A \cap B)$

compliment $\bar{A} = [n] \setminus A$

$$S_{x,y,z,z'}(A,B) = \frac{x|A \cap B| + y|\overline{A \cup B}| + z|A \Delta B|}{x|A \cap B| + y|\overline{A \cup B}| + z'|A \Delta B|}$$

$$JS(A,B) = S_{1,0,0,1}(A,B) = \frac{|A \cap B|}{|A \cap B| + |A \Delta B|}$$

$$Ham(A,B) = 1 - \frac{|A \Delta B|}{|[n]|} = S_{1,1,0,1}(A,B)$$

Andberg Sim

$$Andb(A,B) = \frac{|A \cap B|}{|A \cup B| + |A \Delta B|}$$

Rogers-Tanimoto Sim

$$RT(A,B) = \frac{|[n]| - |A \Delta B|}{|[n]| + |A \Delta B|}$$

Dice Sim

$$Dice(A,B) = \frac{2|A \cap B|}{2|A \cup B| + |A \Delta B|} = \frac{2|A \cap B|}{|A| + |B|}$$

$S_{x,y,z,z'}$ $\longrightarrow$ $D(A,B) = 1 - S(A,B)$

metric

$\longrightarrow$ LSH scheme

Document $\rightarrow$ Set

Bag of Words $\rightarrow$ vector

$(v_1, v_2, \ldots, v_d)$

v_i = # of word i in doc

d = # of words

vs.

k-Grams

"Have a nice day"

{[Have a], k=2

[a nice],

[nice day]}

Modeling choices

- words vs. characters
  ← short texts
- choice of k
  word : k = {2, 3}
  char : k = {3, 4}
- Capitalization [Sa] ≟ [sa]
- Punctuation Ph.D. Mr.
- Stop Word : {a, the, to, and, that, it, is}
  ↳ remove stop words
  ↳ [the pizza oven]