

L3: Jaccard Similarity and k -Grams

Jeff M. Phillips

January 14, 2019

Messy input

• homework assignments h_1, h_2

• webpages

• emails

text

→ set object $S \in \Omega$

vector

in \mathbb{R}^d

$$v = (v_1, v_2, v_3, \dots, v_d) \in \mathbb{R}^d$$

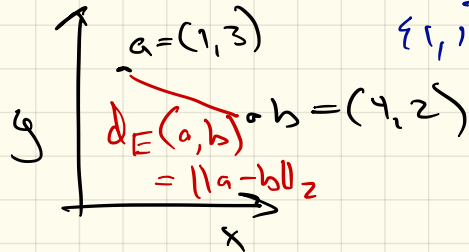
Similarity $s(h_1, h_2)$

is $s(h_1, h_2)$ large?

eg. $\Omega = [n]$

$$S = \{1, 7, 14\}$$

$$\{1, 14, 7, \cancel{7}\}$$



Similarity

$s(A, B)$

if small, \rightarrow A, B far

usually 1 \rightarrow same

$s \in [0, 1]$

A

Distance

$d(A, B)$

\rightarrow A, B close

0 \rightarrow same

$d \in [0, \infty)$

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

or

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

Jaccard Similarity

$$A = \{0, 1, 2, 5, 6\}$$

$$B = \{0, 2, 3, 5, 7, 9\}$$

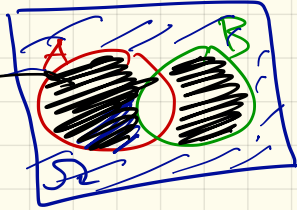
$$\begin{aligned} J_S(A, B) &= \frac{|A \cap B|}{|A \cup B|} \quad \leftarrow \begin{array}{l} \text{cardinality} \\ \text{size of set} \end{array} \\ &= \frac{|\{0, 2, 5\}|}{|\{0, 1, 2, 3, 5, 6, 7, 9\}|} = \frac{3}{8} = 0.375 \end{aligned}$$

$$D_S(A, B) = 1 - J_S(A, B)$$

Generalized Similarities

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

Sym. diff. \rightarrow



$$\begin{cases} x=1 & y=0 \\ z=0 & z'=1 \end{cases}$$

$$\text{Hamming}(A, B) = S_{1,1,0,1} = 1 - \frac{|A \Delta B|}{|S|}$$

$$\text{Andberg}(A, B) = S_{1,0,0,2} = \frac{|A \cap B|}{|A \cup B| + |A \Delta B|}$$

$$\text{Rogers-Tanimoto} = S_{1,1,0,2} = \frac{|S| - |A \Delta B|}{|S| + |A \Delta B|}$$

$$\text{Dice} = S_{2,0,0,1} = \frac{2|A \cap B|}{|A| + |B|}$$

Modeling Text

I am Sam.

Sam I am.

I do not like green eggs and ham.

I do not like them, Sam I am.

Modeling Text

I am Sam.

Sam I am.

I do not like green eggs and ham.

I do not like them, Sam I am.

Bag-of-Words:

(am, and, do, eggs, green, ham, I, like, not, Sam, them, zebra) = \mathbb{R}^{12}

Modeling Text

I am Sam.

Sam I am.

I do not like green eggs and ham.

I do not like them, Sam I am.

Bag-of-Words:

(am, and, do, eggs, green, ham, I, like, not, Sam, them, zebra)

$$v_1 = (1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0) \in \mathbb{R}^{12}$$

$$v_2 = (1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0)$$

$$v_3 = (0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0)$$

$$v_4 = (1, 0, 1, 0, 0, 0, 2, 1, 1, 1, 1, 0).$$

k-Grams with Words

$$k=3$$

I am Sam.

Sam I am.

I do not like green eggs and ham,
I do not like them, Sam I am.

k-Grams with Words

I am Sam.

Sam I am.

I do not like green eggs and ham.

I do not like them, Sam I am.

Single Document

Words $k = 1$:

{[I], [am], [Sam], [do], [not], [like], [green],
[eggs], [and], [ham], [them]}

k -Grams with Words

I am Sam.

Sam I am.

I do not like green eggs and ham.

I do not like them, Sam I am.

Words $k = 1$:


{[I], [am], [Sam], [do], [not], [like], [green],
[eggs], [and], [ham], [them]}

Words $k = 2$:

{[I am], [am Sam], [Sam Sam], [Sam I], [am I], [I
do], [do not], [not like], [like green], [green
eggs], [eggs and], [and ham], [ham I], [like them],
[them Sam]}

k -Grams with Characters

I am Sam.
Sam I am.



Characters $k = 3$:

{[iam], [ams], [msa], [sam], [ami], [mia]}

k -Grams with Characters

I am Sam.

Sam I am.

Characters $k = 3$:

{[iam], [ams], [msa], [sam], [ami], [mia]}

Characters $k = 4$:

{[iams], [amsa], [msam], [sams], [sami], [amia],
[miam]}

Modeling Choices

- choice of k
- word vs. character
- character: spaces?
- capitalization
- set vs. vector (dictionary)
- punctuation.

k -Grams and Jaccard

D_1 : I am Sam.

D_2 : Sam I am.

D_3 : I do not like green eggs and ham.

D_4 : I do not like them, Sam I am.

Words $k = 2$:

{[I am], [am Sam], [Sam Sam], [Sam I], [am I], [I do], [do not], [not like], [like green], [green eggs], [eggs and], [and ham], [like them], [them Sam]}

k-Grams and Jaccard

D_1 : [I am], [am Sam]

D_2 : [Sam I], [I am]

D_3 : [I do], [do not], [not like], [like green]
[green eggs], [eggs and], [and ham]

D_4 : [I do], [do not], [not like], [like them], [them Sam]
[Sam I], [I am]

k-Grams and Jaccard

D_1 : [I am], [am Sam]

D_2 : [Sam I], [I am]

D_3 : [I do], [do not], [not like], [like green]
[green eggs], [eggs and], [and ham]

D_4 : [I do], [do not], [not like], [like them], [them Sam]
[Sam I], [I am]

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

k-Grams and Jaccard

D_1 : [I am], [am Sam]

D_2 : [Sam I], [I am]

D_3 : [I do], [do not], [not like], [like green]
[green eggs], [eggs and], [and ham]

D_4 : [I do], [do not], [not like], [like them], [them Sam]
[Sam I], [I am]

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

$$JS(D_1, D_2) = 1/3 \approx 0.333$$

k-Grams and Jaccard

D_1 : [I am], [am Sam]

D_2 : [Sam I], [I am]

D_3 : [I do], [do not], [not like], [like green]
[green eggs], [eggs and], [and ham]

D_4 : [I do], [do not], [not like], [like them], [them Sam]
[Sam I], [I am]

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

$$JS(D_1, D_2) = \frac{1}{3} \approx 0.333$$

$$JS(D_1, D_3) = 0 = 0.0$$

k-Grams and Jaccard

D_1 : [I am], [am Sam]

D_2 : [Sam I], [I am]

D_3 : [I do], [do not], [not like], [like green]
[green eggs], [eggs and], [and ham]

D_4 : [I do], [do not], [not like], [like them], [them Sam]
[Sam I], [I am]

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

$$JS(D_1, D_2) = 1/3 \approx 0.333$$

$$JS(D_1, D_3) = 0 = 0.0$$

$$JS(D_1, D_4) = 1/8 = 0.125$$

k-Grams and Jaccard

D_1 : [I am], [am Sam]

D_2 : [Sam I], [I am]

D_3 : [I do], [do not], [not like], [like green]
[green eggs], [eggs and], [and ham]

D_4 : [I do], [do not], [not like], [like them], [them Sam]
[Sam I], [I am]

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

$$JS(D_1, D_2) = 1/3 \approx 0.333$$

$$JS(D_1, D_3) = 0 = 0.0$$

$$JS(D_1, D_4) = 1/8 = 0.125$$

$$JS(D_2, D_3) = 0 = 0.0$$

$$JS(D_2, D_4) = 2/7 \approx 0.286$$

$$JS(D_3, D_4) = 3/11 \approx 0.273$$

Continuous Bag of Words

word vector

$$v_{\text{king}} - v_{\text{queen}} \\ \rightarrow v_{\text{man}} = v_{\text{woman}}$$

↑ green
 $v_{\text{green}} \in \mathbb{R}^{300}$

I am Sam Sam I am I do not like green eggs and ham I
do not like them Sam I am

$$v_{\text{like 1}} = (0, 0, \dots, 1, \dots, 1, 0, 0)$$

$$v_{\text{like 2}} = (0, 0, \dots, 1, \dots, 1, 0, 0)$$

$$v_{\text{like 3}} = (0, 1/2, \dots, 1, \dots, 1, 0, 0, \dots)$$

