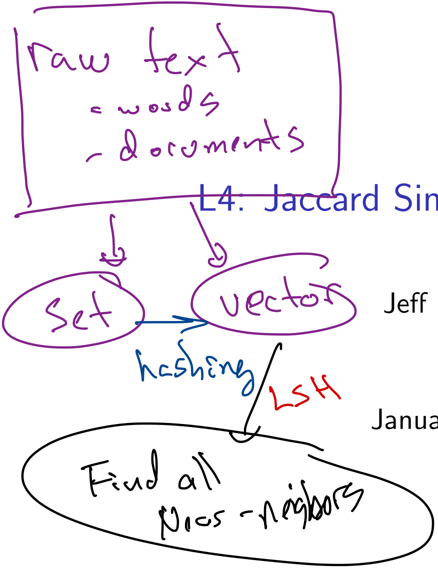
L4: Jaccard Similarity and *k*-Grams

Jeff M. Phillips

January 15, 2025



raw text
- words
- documents

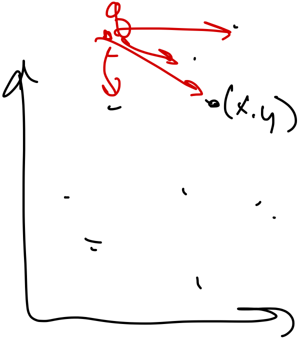
Set

vector

hashing

LSH

Find all Near-neighbors



Distances

Euclidean distance

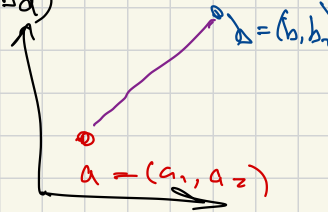
Data: $X \subset \mathbb{R}$ $a, b \in X$

$$a = (a_1, a_2, a_3, \ldots, a_d)$$

$$b = (b_1, b_2, \ldots\ b_d)$$

$$d_E(a,b) = \sqrt{\sum_{j=1}^{d} (b_j - a_j)^2}$$

$$= \|a - b\|$$



Distance

$d(a,b)$

if a,b close
then $d(a,b)$ small.

usually $a=b \Rightarrow d(a,b)=0$

$d \in [0,1]$ or $[0,\infty)$

given $s(a,b)$

$d(a,b) = 1 - s(a,b)$

or $= \sqrt{s(a,a) + s(b,b) - 2 \cdot s(a,b)}$

Similarity

$s(a,b)$

if a,b close
then $s(a,b)$ large

usually $a=b$
$\Rightarrow s(a,b)=1$

$s \in [0,1]$

Jaccard Similarity

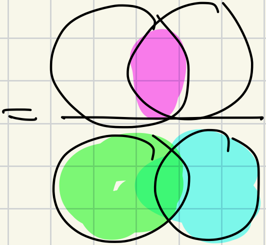
two sets A, B

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

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

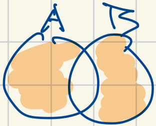
$$JS(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

$$= \frac{|\{0, 2, 5\}|}{|\{0, 1, 2, 3, 5, 6, 7, 9\}|} =$$



$$= \frac{3}{8} = 0.375$$

$A \Delta B$



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

if $x, y, z \geq 0$ $\quad z' \geq z$ $\quad$ "makes sense"

$JS = S_{1,0,0,1}$ → does not depend on domain, stuff not in A, B

$\text{Hamming} = S_{1,1,0,1}$

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

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

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) $\in \mathbb{R}^{11}$

$d = 11$

$d = 100,000$

Modeling Text

$D_1 =$ `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.`

Bag-of-Words:

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

$$
\begin{aligned}
v_1 &= (1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0) \\
v_2 &= (1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0) \\
v_3 &= (0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0) \\
v_4 &= (1, 0, 1, 0, 0, 0, 2, 1, 1, 1, 1, 0).
\end{aligned}
$$

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.
```

k-Grams with Words

$D_1 = \{$

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]}

k-Grams with Words

Shingling

D :=

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.
```

- no punctuation
- no whitespace
- no capitalization.
- wrap-around, continue next sentence.
- ~~k~~

Characters $k = 3$:
`{[iam], [ams], [msa], [sam], [ami], [mia]}`

Characters $k = 4$:
`{[iams], [amsa], [msam], [sams], [sami], [amia], [miam]}`

- characters vs. words vs. subword

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

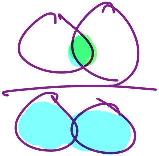
D_1 : {[I am], [am Sam]}

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

$$JS(D_1, D_2) = \frac{|D_1 \cap D_2|}{|D_2 \cup D_1|}$$

$|D_1 \cap D_2| = |\{[\text{I am}]\}| = 1$

$|D_2 \cup D_1| \leftarrow \{[\text{I am}], [\text{am Sam}], [\text{Sam I}]\} \Rightarrow 3$

$$= \frac{1}{3} = 0.333\ldots$$

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

$$\mathsf{JS}(D_1, D_2) = \quad 1/3 \quad \approx 0.333$$

$$JS(D_1, D_3) = 0 \qquad \text{since } D_1 \cap D_3 = \emptyset$$

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

$$\begin{aligned} \mathsf{JS}(D_1, D_2) &= \quad 1/3 \quad &\approx 0.333 \\ \mathsf{JS}(D_1, D_3) &= \quad 0 \quad &= 0.0 \end{aligned}$$

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

$$\begin{aligned} \mathsf{JS}(D_1, D_2) &= \quad 1/3 && \approx 0.333 \\ \mathsf{JS}(D_1, D_3) &= \quad 0 && = 0.0 \\ \mathsf{JS}(D_1, D_4) &= \quad 1/8 && = 0.125 \end{aligned}$$

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

$$
\begin{aligned}
\mathsf{JS}(D_1, D_2) &= && 1/3 && \approx 0.333 \\
\mathsf{JS}(D_1, D_3) &= && 0 && = 0.0 \\
\mathsf{JS}(D_1, D_4) &= && 1/8 && = 0.125 \\
\mathsf{JS}(D_2, D_3) &= && 0 && = 0.0 \\
\mathsf{JS}(D_2, D_4) &= && 2/7 && \approx 0.286 \\
\mathsf{JS}(D_3, D_4) &= && 3/11 && \approx 0.273
\end{aligned}
$$

Continuous Bag of Words

negation

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

$\{$ am, and, do $\} \in \mathbb{R}^{V}$

$(0 \quad 0, 1, 0, 1, 0, 1, 0, , 1, 0, 0)$