L6: Similarities

Jaccard + MinHashing

Data Mining
Sep 10, 2026
Jeff M. Phillips

Distance

$D : X \times X \to \mathbb{R}_{\geq 0}$

- small if pair a, b are close
- large if far
- usually 0 if (and only if) $a = b$
- $[0, \infty)$

Similarity

$S : X \times X \to \mathbb{R}_{\geq 0}$

- large if pair A, B are close
- small if far
- usually 1 if $A = B$
- $[0, 1]$

$$D(A, B) = 1 - S(A, B)$$

$$D(A, B) = \sqrt{S(A, A) + S(B, B) - 2S(A, B)}$$

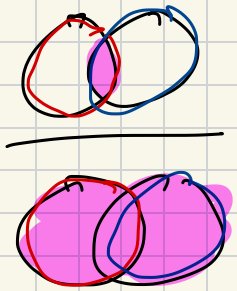
Jaccard Similarity

$S_J : [Set] \times [Set] \rightarrow [0,1]$

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

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

$$JS(A, B) = \frac{|A \cap B|}{|A \cup B|} \in [0, 1]$$



example $= \frac{|\{0, 2, 5\}|}{|\{0, 1, 2, 3, 5, 6, 7, 9\}|} = \frac{3}{8}$

$= 0.375$

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)

(3, 1, 2, 1, 1, 1, 5, 2, 2, 3, 1, 0)

weighted → BM25

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)

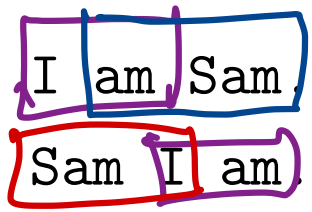
$$v_1 = (1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0)$$
$$v_2 = (1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0)$$
$$v_3 = (0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0)$$
$$v_4 = (1, 0, 1, 0, 0, 0, 2, 1, 1, 1, 1, 0).$$

$v_1 = v_2$

k-Grams with Words

↳ Shingle

$k = 2$



I do not like green eggs and ham.

I do not like them, Sam I am.

$k = 5$

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]}

Set

[I] occurs once

k-Grams with Words

Modeling Choices

- parts of speech
- choice of k
- existing (not)
- remove "stop" words. the of.
- punctuation
- across sentence

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]`}

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

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

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

$$= \frac{|[\text{I am}]|}{|\{[\text{I am}], [\text{am Sam}], [\text{Sam I}]\}|}$$

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

$$\text{JS}(D_1, D_2) = 1/3 \approx 0.333$$
$$\text{JS}(D_1, D_3) = 0 = 0.0$$

$\frac{0}{9}$

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

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

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

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

I have 1 billion documents → Set k-grams

Q1. on a query q, retrieve all similar documents

Q2. find all pairs of similar documents.

Min Hashing

$h_a \sim \mathcal{H}$ ← family hash functions

↳ draw at random

salt ↑

$h_a : \{k\text{-grams}\} \to \mathbb{R}$

$h'_a : (\text{set of kgrams}) \to \mathbb{R}$

$$\Pr_{h'_a \sim \mathcal{H}}\left[h'_a(D_1) = h'_a(D_2)\right] = JS(D_1, D_2)$$

MinHashing

$h'(T)$ T set $= \{t_1, t_2 \ldots t_n\}$

Set $v = \infty$

for $t_i \in T$ do

if $(h_a(t_i) < v)$

$v \leftarrow h_a(t_i)$

return $v \Rightarrow h'(T)$

$h_a : \{k\text{-grams}\} \rightarrow \mathbb{R}$

$h_a : (\{\text{sets of } k\text{grams}\}) \rightarrow \mathbb{R}$