L6: Jaccard Similarity, k-Grams, and Min Hashing

Jeff M. Phillips

September 8, 2025

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

Distance

- Small if pair A, B are close
- large if B is
- usually 0 if $A = B$
- $[0, \infty)$

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

Similarity

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

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

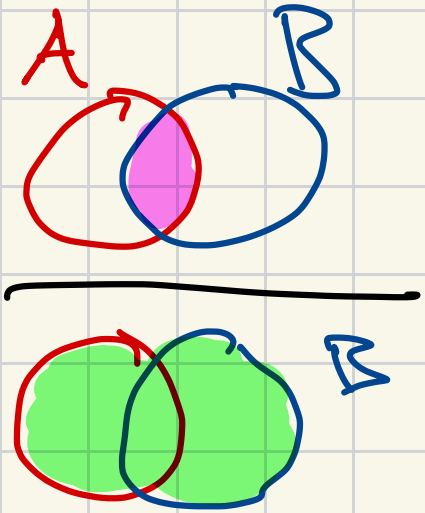
$$D(A,B) = \sqrt{S(A,A) + S(B,B) - 2S(A,B)}$$

Jaccard Similarity

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

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

sets



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

$$\in [0, 1]$$

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.

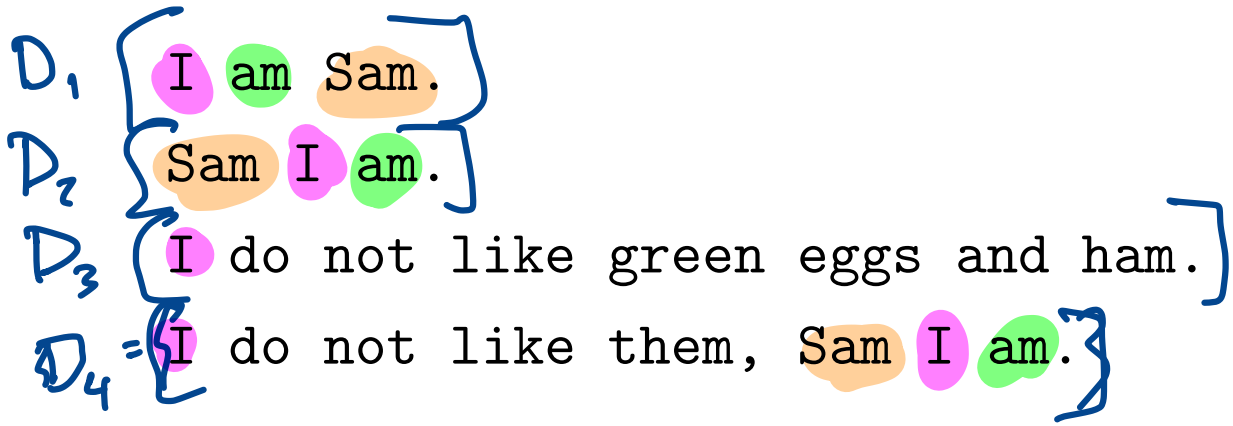
I do not like them, Sam I am.

Bag-of-Words:

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

(3, 1, 2, 1, 1 ... , 0)

Modeling Text



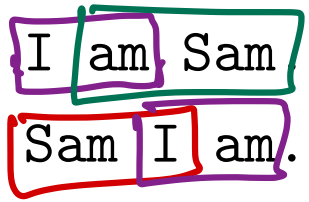
Bag-of-Words:

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

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

k-Grams with Words

↳ shingles



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.
```

Words $k = 1$:
{[I], [am], [Sam], [do], [not], [like], [green], [eggs], [and], [ham], [them]}

k-Grams with Words

Modelling – wrap around lines, sentences.

– consecutive

– no punctuation

– kept "stop words"

– choice of k.

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

$$JS(D_1, D_2) = \frac{|D_1 \cap D_2|}{|D_1 \cup D_2|} = \frac{1}{3} = \frac{1}{3}$$

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

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

$$JS(D_1, D_3) = \frac{|D_1 \cap D_3|}{|D_1 \cup D_3|} = \frac{0}{9} = 0$$

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

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

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

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

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 as k-gram sets.

Q1: on query q, find most similar.

Q2: find all close pairs.

Min hashing

$h_a \sim \mathcal{H}$

documents as k-grams

$h_a : 2^{\{k\text{-grams}\}} \to \mathbb{R}$

Another $h'_a \sim \mathcal{H}'$

$h'_a : \{k\text{-grams}\} \to \mathbb{R}$

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

Min Hashing $h(S)$ on set S.

```
Set  v = ∞
for  x ∈ S  do
     if (h'(x) < v)
          v ← h'(x)
return v
```

$v = h(S)$

Why $\Pr_{h \sim \mathcal{H}}[h(D_1) = h(D_2)] = J S(D_1, D_2)$

Let $v^* = h(D_1 \cup D_2)$

$v^* = x \in D_1 \cup D_2$ w.p. $\frac{1}{|D_1 \cup D_2|}$ (each)

When will $h(D_1) = h(D_2)$

iff $v^* \in D_1$ and $v^* \in D_2$

$\Longleftrightarrow$ $v^* \in D_1 \cap D_2$

↳ $|D_1 \cap D_2|$ options.

$\Pr[h(D_1) = h(D_2)] = \frac{|D_1 \cap D_2|}{|D_1 \cup D_2|} = J S(D_1, D_2)$