Raw Text ① → Abstract Representation {Sets} ② → Vectors

Min Hashing

③ LSH ↓

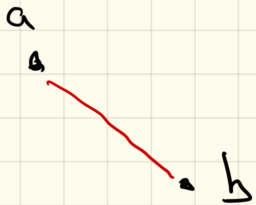
Fast Comparison

L3: Jaccard Similarity and k-Grams

Jeff M. Phillips

January 13, 2020

Distance



Euclidean distance

$a = (a_1, a_2, \ldots a_d) \in \mathbb{R}^d$

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

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

Inverse of Distance

↳ Similarity $s(a,b)$

Distance	Similarity
$d(a,b)$	$s(a,b)$
small if a, b close	large if a, b close
if large → a, b far	if small, not close
0 if same	1 if same
$d(a,b) \in [0,\infty)$	$s(a,b) \in [0,1]$

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

or

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

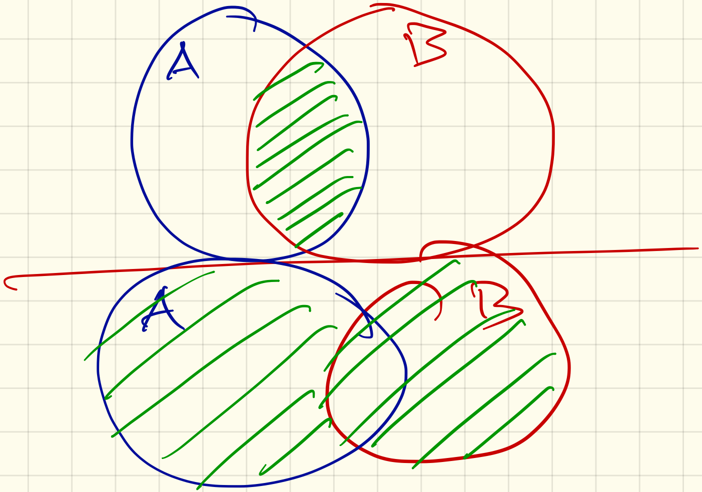
Jaccard Similarity

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

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

$$JS(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|\{0, 2, 5\}|}{|\{0,1,2,3,5,6,7,9\}|} = \frac{3}{8}$$

$$= 0.375$$



$d_J(A,B) = 1 - JS(A,B)$

Jaccard distance

Similarities between Sets

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

$x, y, z, z' \geq 0$ $\qquad z' \geq z$

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

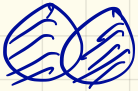
$$Ham(A,B) = S_{1,1,0,1}(A,B) = 1 - \frac{|A\Delta B|}{|[n]|}$$

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

$$RT(A,B) = S_{1,1,0,2}(A,B) = \frac{|[n]| - |A\Delta B|}{|[n]| + |A\Delta B|}$$

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

$A \Delta B$



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

Text → vector in $\mathbb{R}^d$

$d = 11$

```
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)

Modeling Text

count of each word
at ith coordinate

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)

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

k-Grams with Words

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

As single document (not 4)

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

Shingles

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]}

Jaccard ← sets

k-Grams with Characters

punctuation

`I am Sam.`

`Sam I am.`

capitalization

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

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

Modeling Choices

- words vs. characters
  ↳ more interpretable
- new lines
- k ? ← (size of gram)
- Capitalization
- punctuation
  ↳ highlight #

More complex rep.
↕
more data

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

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

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

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

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

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

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

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

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

Continuous Bag of Words

each word $\xrightarrow[\text{map}]{}$ vector $v_{word} \in \mathbb{R}^d$

bow

$(0, 0, 0, 1, 0, \dots, 0)$

#1

I am Sam Sam I am I $\boxed{\text{do not }\underline{\text{like}}\text{ green eggs}}$ and ham I

#2 $\boxed{\text{do not }\underline{\text{like}}\text{ them Sam}}$ I am

word = "like"

them do not green eggs sam

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

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

$$v_{like}^{*} = (0, 1/2, 1, 1, 0, \dots, 1/2, 1/2, 0, 0)$$