L3: Jaccard Similarity and $k$-Grams

Jeff M. Phillips

January 14, 2019
Messy input

- homework assignments
- web pages
- emails

Text $\rightarrow$ set object $S \subseteq \mathbb{R}^2$

Set $S$ of points $a, b$ in $\mathbb{R}^2$:

$$a = (1, 3) \quad \Rightarrow \quad d_E(a, b) = \sqrt{(1 - 4)^2 + (3 - 2)^2} = 11 \text{ and } b = (4, 2)$$
\[
\text{Similarity: } S(A, B) \\
\text{Distance: } d(A, B) \\
\text{if small, } A, B \text{ far} \\
\text{usually } 1 \Rightarrow \text{ same} \\
se [0, 1] \\
\text{or} \\
d(A, B) = 1 - S(A, B) \\
d(A, B) = \sqrt{S(A, A) + S(B, B) - 2S(A, B)} \\
d \in [0, \infty)
\]
\[ Jaccard \ Similarity \]

\[
A = \{0, 1, 2, 5, 63\} \quad \text{and} \quad B = \{0, 2, 3, 5, 7, 9\}
\]

\[
JSC(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{1\{0, 2, 53\}}{1\{0, 1, 2, 3, 5, 6, 7, 93\}} = \frac{3}{8} = 0.375
\]

\[
D_5(A, B) = 1 - JSC(A, B)
\]
Generalized Similarities

\[ S_{x,y,z,z'}(AB) = x |A \cap B| + y |A \cup B| + z |A \Delta B| \]

\[ x |A \cap B| + y |A \cup B| + z |A \Delta B| \]

Hamming \( (A, B) = S_{1,1,1,0,1} = 1 - \frac{|A \Delta B|}{|2|} \)

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

Rogers-Tanimoto \( = S_{1,1,0,2} = \frac{|A \cap B|}{|A \cup B| + |A \Delta B|} \)

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

\[ \begin{align*}
\nu_1 &= (1,0,0,0,0,0,1,0,0,1,0,0) \\
\nu_2 &= (1,0,0,0,0,0,1,0,0,1,0,0) \\
\nu_3 &= (0,1,1,1,1,1,1,1,1,1,0,0) \\
\nu_4 &= (1,0,1,0,0,0,2,1,1,1,1,0).
\end{align*} \]
$k$-Grams with Words

$I$ am $S$am.
$S$am $I$ am.

$I$ do not like green eggs and ham.

$I$ do not like them, $S$am $I$ am.
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**

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 12
- word vs. character
- character vs. spaces?
- capitalization
- set vs. vector (dictionary)
- punctuation
$k$-Grams and Jaccard

\[ D_1: \text{I am Sam}. \]
\[ D_2: \text{Sam I am}. \]
\[ D_3: \text{I do not like green eggs and ham}. \]
\[ D_4: \text{I do not like them, Sam I am}. \]

Words $k = 2$:
\[ \{[\text{I am}], [\text{am Sam}], [\text{Sam Sam}], [\text{Sam I}], [\text{am I}], [\text{I do}], [\text{do not}], [\text{not like}], [\text{like green}], [\text{green eggs}], [\text{eggs and}], [\text{and ham}], [\text{like them}], [\text{them Sam}] \} \]
$k$-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]

Jaccard Similarity:

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

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

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

$JS(D_1, D_4) = \frac{1}{8} \approx 0.125$

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

$JS(D_2, D_4) = \frac{2}{7} \approx 0.286$

$JS(D_3, D_4) = \frac{3}{11} \approx 0.273$
**k-Grams and Jaccard**

\[ D_1: [I\ am],\ [am\ Sam] \]
\[ D_2: [Sam\ I],\ [I\ am] \]
\[ D_3: [I\ do],\ [do\ not],\ [not\ like],\ [like\ green]\]
\[\quad [green\ eggs],\ [eggs\ and],\ [and\ ham] \]
\[ D_4: [I\ do],\ [do\ not],\ [not\ like],\ [like\ them],\ [them\ Sam] \]
\[\quad [Sam\ I],\ [I\ am] \]

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

$JS(D_1, D_2) = \frac{1}{3} \approx 0.333$
$k$-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: $JS(A, B) = \frac{|A \cap B|}{|A \cup B|}$

$JS(D_1, D_2) = \frac{1}{3} \approx 0.333$
$JS(D_1, D_3) = 0 = 0.0$
$JS(D_1, D_4) = \frac{1}{8} = 0.125$
**k-Grams and Jaccard**

\[ D_1: [I\ am], [am\ Sam] \]
\[ D_2: [Sam\ I], [I\ am] \]
\[ D_3: [I\ do], [do\ not], [not\ like], [like\ green] \]
\[ \quad [green\ eggs], [eggs\ and], [and\ ham] \]
\[ D_4: [I\ do], [do\ not], [not\ like], [like\ them], [them\ Sam] \]
\[ \quad [Sam\ I], [I\ am] \]

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

\[
JS(D_1, D_2) = \frac{1}{3} \approx 0.333
\]
\[
JS(D_1, D_3) = 0 = 0.0
\]
\[
JS(D_1, D_4) = \frac{1}{8} = 0.125
\]
\[
JS(D_2, D_3) = 0 = 0.0
\]
\[
JS(D_2, D_4) = \frac{2}{7} \approx 0.286
\]
\[
JS(D_3, D_4) = \frac{3}{11} \approx 0.273
\]
Continuous Bag of Words

*I am Sam*

\[ u_{\text{zing}} = u_{\text{green}} \]

\[ v_{\text{man}} = v_{\text{woman}} \]

\[ v_{\text{green}} \in \mathbb{R}^{300} \]

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

\[ u_{\text{illeg}} = (0, 0, \ldots, 1, \ldots, 1, 0, 0) \]

\[ u_{\text{line 2}} = (0, 0, \ldots, 1, \ldots, 1, 0, 0) \]

\[ u_{\text{line 3}} = (0, 0, \ldots, 1, \ldots, 1, 0, 0, \ldots) \]