L3: Jaccard Similarity and $k$-grams

Jeff M. Phillips

January 17, 2017
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

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], [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]}
\textit{k-Grams and Jaccard}

\begin{align*}
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.}
\end{align*}

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]

Jaccard Similarity:

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

$\text{JS}(D_1, D_3) = 0 = 0.0$

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

$\text{JS}(D_2, D_3) = 0 = 0.0$

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

$\text{JS}(D_3, D_4) = 3/11 \approx 0.273$
$k$-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: $\text{JS}(A, B) = \frac{|A \cap B|}{|A \cup B|}$
$k$-Grams and Jaccard

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

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**

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

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

\[
\begin{align*}
JS(D_1, D_2) &= \frac{1}{3} \approx 0.333 \\
JS(D_1, D_3) &= 0 \quad = 0.0 \\
JS(D_1, D_4) &= \frac{1}{8} \quad = 0.125 \\
JS(D_2, D_3) &= 0 \quad = 0.0 \\
JS(D_2, D_4) &= \frac{2}{7} \quad \approx 0.286 \\
JS(D_3, D_4) &= \frac{3}{11} \quad \approx 0.273
\end{align*}
\]
I am Sam Sam I am I do not like green eggs and ham I do not like them Sam I am