L3: Jaccard Similarity and *k*-Grams

Jeff M. Phillips

January 20, 2016

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

```
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]}

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

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

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

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

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

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

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

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