Hashing

- SHA-1 hashing (recommended)

$$\text{hash}(\underbrace{\text{concat}(\text{salt}, x)}_{\text{string}}) \to [m]$$

- Multiplicative Hashing

$$h_a(x) = \lfloor m \cdot \text{frac}(x \times a) \rfloor$$

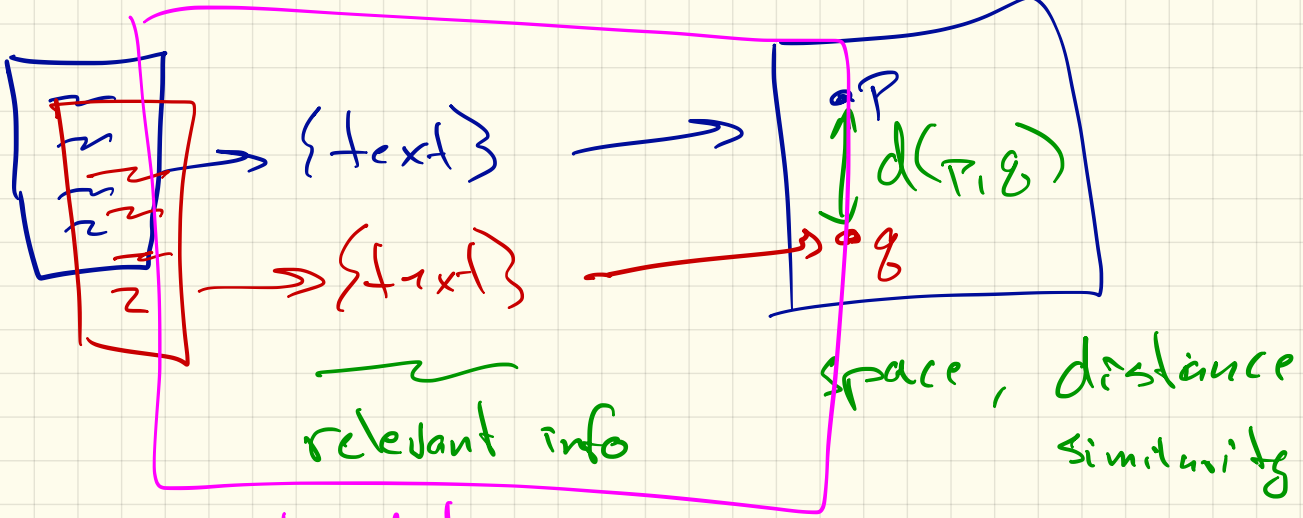
a: deterministic

- ~~Modular $h(x) = x \pmod m$~~ don't

Distances between document (text)

Jaccard Distance, k-Grams

- Give 2 homeworks: did the plagerize close to copies?
- Given keyword in Google, which webpages are similar? Are 2 pages duplicates.
- emails? is it spam?



Today

p, q are sets

Jaccard Distance

Alta-Vista

Common

$p, q \in \mathbb{R}^d$

$p = (p_1, p_2, \ldots, p_d)$

Sets

$\{a, b, c\} = \{a, a, b, c\}$

$= \{c, b, a\}$

$\neq \{a, b\}$

$\neq \{a, b, c, d\}$

Distance

$d(A, B)$

small $\iff$ A, B close

large $\iff$ far

$0 \iff A \simeq B$

$[0, \infty)$

Similarity

$s(A, B)$

small $\iff$ A, B far

large $\iff$ close

$1 \iff A \simeq B$

$[0, 1]$

$d(A, B) = 1 - s(A, B)$ (Jaccard)

$d(A, B) = \sqrt{s(A, A) + s(B, B) - 2 s(A, B)}$

Jaccard Distance / Similarity

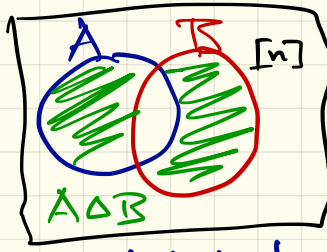
$A = \{0, 1, 2, 5, 6\} \cup \{3\}$

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

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

$$JD(A,B) = 1 - JS(A,B) = \frac{3}{8} = 0.375$$

Generalized Set Distances



- Hamming Sim

$$Ham(A,B) = \frac{|A \cap B| + |\overline{A \cup B}|}{|A \cap B| + |\overline{A \cup B}| + |A \Delta B|} = 1 - \frac{|A \Delta B|}{|[n]|}$$

- Andberg Sim

$$Andb(A,B) = \frac{|A \cap B|}{|A \cup B| + |A \Delta B|}$$

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

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

$$JS(A,B) = S_{1,0,0,1}(A,B)$$

L3: Jaccard Similarity and k-Grams

Jeff M. Phillips

January 17, 2018

k-Grams with Words

k-gram

2-word gram

I am Sam.

Sam I am.

I do not like green eggs and ham.

I do not like them, Sam I am.

{[I am], [am Sam], [Sam Sam],

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], [ham I], [like them], [them Sam]}

k-Grams with Characters

$k = 3$

Modeling

- word vs. char
- value of k
- punctuation
- white space
- Capitalization

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

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

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

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

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

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

Bag-of-Words Model

[Sam I am] ← document D1

$D1 \rightarrow p \in \mathbb{R}^m$

$p = (p_1, p_2, p_3, \ldots p_m)$

"1" = Sam "2" = I "3" = am

p_i = # occurences of word "i"

$p = (0, 1, 1, 0, 1, 0, 0, \ldots 0)$

↓

0, 2, 1, 0, 1, 0, 0

+Sam

Continuous Bag of Words

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