## Words: Encodings & Tokenization

CS 490A, Fall 2021

Applications of Natural Language Processing <u>https://people.cs.umass.edu/~brenocon/cs490a\_f21</u>

Brendan O'Connor & Laure Thompson

College of Information & Computer Sciences University of Massachusetts Amherst

#### Administrivia – Submissions

- Submit Exercise o if you haven't
- HWo due Thursday (9/9) at 11:59pm EST
- Future assignments will be due **before class**

#### Administrivia – Office Hours

- Akshay's OH: Wednesday 2-3:30pm , CS207
- Yuanguo's OH: **Thursday** 1-2:30pm, LGRT-T220
- Xiao's OH: Friday 10:15-11:45am, LGRT-T223

## Administrivia – Schedule

• Schedule on website:

people.cs.umass.edu/~brenocon/cs490a\_f21/schedule.html

- Readings and lectures will be posted here
- Complete **required** readings before class
  - "Reading" = required reading
  - "Optional" = supplemental



| Word      | Count |
|-----------|-------|
| article   | 100   |
| charming  | 32    |
| dog       | 0     |
| newspaper | 29    |
| poem      | 3     |
| :         | •     |

Topic Modeling Term-Covariate Ranking Supervised Learning

**Frequency Analysis** 

Working Representations





## What can we do with word counts?

## Analyze trends in *specific* language use



## Analyze trends in *specific* language use



## Compare *general* language use

|                                   |   |   | Foxy Brown<br>Juvenile<br>Master P<br>Salt-n-Pepa  | Run-D.M.C.<br>2Pac<br>Big L<br>Insane Clown<br>MC Lyte<br>Scarface<br>Three 6 Mafia<br>UGK<br>Dizzee Rascal      | Biz Markie<br>Ice T<br>Rakim<br>Brand Nubian<br>Geto Boys<br>Ice Cube<br>Jay-Z<br>Mobb Deep                                  | Beastie Boys<br>Big Daddy Kane<br>LL Cool J   |  | #<br>U   | ‡ of Un<br>Ised W<br>First 35                                     | ique W<br>ithin Ai<br>5,000 ly                 | ′ords<br>rtist′s<br>⁄rics        |
|-----------------------------------|---|---|--|--|--|---|--|--|---|--|----------------------------------|
|                                   |   |   | Snoop Dogg   | Jadakiss   | Outkast  | Busta Rhymes  |  |  |   | BY FRA <sup>1</sup>                            |                                  |
|                                   | DMX   | Bone Thugs-n<br>50 Cent<br>Juicy J<br>Drake<br>Future<br>Kid Cudi | Eve<br>Gucci Mane<br>Kanye West<br>Lil Wayne<br>Missy Elliot<br>Trick Daddy<br>Trina<br>Young Jeezy<br>Big Sean<br>BoB | Kano<br>Lil' Kim<br>Nelly<br>Rick Ross<br>T.I.<br>2 Chainz<br>A\$AP Ferg<br>Big KRIT<br>Brockhampton<br>Cupcakke | Public Enemy<br>Cam'ron<br>Eminem<br>The Game<br>Joe Budden<br>Kevin Gates<br>Royce da 5'9<br>Tech n9ne<br>Twista<br>Ab-Soul | Cypress Hill<br>De La Soul<br>Fat Joe<br>Gang Starr<br>KRS-One<br>Method Man<br>A Tribe Call<br>Atmosphere<br>Ludacris<br>Lupe Fiasco | Common<br>Das EFX                                      | 1  | .980s   199<br>Del the Funk<br>The Roots                          | 0s   2000s                                     | 2010s                            |
|                                   | 21 Savage   | Kid Ink<br>Kodak Black  | Childish Gam   | Hopsin<br>Jay Rock   | A\$AP Rocky  | Mos Det<br>Mure   | E-40<br>Goodie Mob                                     | Kool G Rap   | Blackalicious   |  |                                  |
| Lil Uzi Vert<br>NF                | Lil Baby<br>Lil Durk<br>Wiz Khalifa<br>YG<br>YoungBoy Nev | Lil Yachty<br>Logic<br>Migos<br>Travis Scott<br>Young Thug        | J Cole<br>Machine Gun<br>Meek Mill<br>Nicki Minaj<br>Russ  | Kendrick Lamar<br>Mac Miller<br>ScHoolboy Q<br>Tyga<br>Vince Staples   | Death Grips<br>Denzel Curry<br>\$uicideboy\$<br>Tyler the Cr<br>Wale   | Talib Kweli<br>Xzibit<br>Flatbush Zom<br>Joey BadA\$\$<br>Rittz   | Nas<br>Redman<br>Brother Ali<br>Action Bronson<br>KAAN | Kool Keith<br>Raekwon<br>CunninLynguists<br>Sage Francis<br>Watsky | Ghostface Ki<br>Immortal Tec<br>Jean Grae<br>Killah Priest<br>RZA | GZA<br>Wu-Tang Clan<br>Jedi Mind Tr<br>MF DOOM | Aesop Rock<br>Busdriver          |
| < <b>2,675</b><br>unique<br>words | <b>2,675-3,050</b><br>unique<br>words                     | <b>3,050-3,425</b><br>unique<br>words                             | <b>3,425-3,800</b><br>unique<br>words  | <b>3,800-4,175</b><br>unique<br>words  | <b>4,175-4,550</b><br>unique<br>words  | <b>4,550-4,925</b><br>unique<br>words   | <b>4,925-5,300</b><br>unique<br>words                  | <b>5,300-5,675</b><br>unique<br>words                              | <b>5,675-6,050</b><br>unique<br>words                             | <b>6,050-6,425</b><br>unique<br>words          | <b>6,425+</b><br>unique<br>words |



### Identify distinctive word use

Corpus: News articles from the late 1960s



Wasow 2020





#### Numerical



Frequency Analysis Topic Modeling Term-Covariate Ranking Supervised Learning

Vector Representations



## This is a simple sentence.

0101010001101000011010010111001100100000011010010111001100100000011000010010000001101101011000010110010101100000011001010111000001100101011100000110010101110000

# From Bits to Characters



## Bacon's cipher

Strategy: Assign each letter in the English alphabet a unique value

| $A \rightarrow 0$ | $X \rightarrow 23$ |
|-------------------|--------------------|
| $B \rightarrow 1$ | $Y \rightarrow 24$ |
| $C \rightarrow 2$ | $Z \rightarrow 25$ |

**Q:** How many bits do we need to uniquely encode these numbers?

## Bacon's cipher

Strategy: Assign each letter in the English alphabet a unique value

| $A \rightarrow 00000$         | $\mathrm{X}  ightarrow$ 01111 |
|-------------------------------|-------------------------------|
| ${ m B}  ightarrow$ 00001     | $Y \rightarrow$ 11000         |
| $\mathrm{C}  ightarrow$ 00010 | Z  ightarrow 11001            |

Ascii – 7 bits

USASCII code chart

| b7b6b | 5 -     |          |                |          |     | ° ° °      | °° , | 0<br> <br>0 | 0 | <sup>1</sup> 0 <sub>0</sub> | <sup>1</sup> о | 1<br>1<br>0 | 1      |
|-------|---------|----------|----------------|----------|-----|------------|------|-------------|---|-----------------------------|----------------|-------------|--------|
|       | Þ4<br>† | b 3<br>1 | b <sub>2</sub> | Ь ,<br>, | Row | 0          | I    | 2           | 3 | 4                           | 5              | 6           | 7      |
|       | 0       | 0        | 0              | 0        | 0   | NUL .      | DLE  | SP          | 0 | 0                           | Ρ              | Ň           | р      |
|       | 0       | 0        | 0              | 1        | 1   | SOH        | DC1  | !           | 1 | Α.                          | Q              | 0           | Q      |
|       | 0       | 0        | 1              | 0        | 2   | STX        | DC2  |             | 2 | 8                           | R              | . Þ         | r      |
|       | 0       | 0        | 1              | 1        | 3   | ETX        | DC 3 | #           | 3 | C                           | S              | С           | 5      |
|       | 0       | 1        | 0              | 0        | 4   | EOT        | DC4  | \$          | 4 | D                           | Т              | d           | t      |
|       | 0       | 1        | 0              | 1        | 5   | ENQ        | NAK  | %           | 5 | E                           | U              | e           | U      |
|       | 0       | 1        | 1              | 0        | 6   | ACK        | SYN  | 8           | 6 | F                           | V              | f           | v      |
|       | 0       | -        | 1              | 1        | 7   | 8EL        | ETB  | ,           | 7 | G                           | W              | g           | W      |
|       | 1       | 0        | 0              | 0        | 8   | BS         | CAN  | (           | 8 | н                           | X              | h           | ×      |
|       |         | 0        | 0              |          | 9   | нт         | EM   | )           | 9 | 1                           | Y              | i           | У      |
|       |         | 0        | 1              | 0        | 10  | LF         | SUB  | *           | : | J                           | Z              | j           | Z      |
|       | 1       | 0        | 1              | 1        |     | VT         | ESC  | +           | ; | К                           | C              | k           | (      |
|       | I       | 1        | 0              | 0        | 12  | FF         | FS   | •           | < | L                           | \              | l           | 1      |
|       | ł       | 1        | 0              | 1        | 13  | CR         | GS   | -           | = | м                           | כ              | m           | }      |
|       | 1       | 1        | 1              | 0        | 4   | SO         | RS   | •           | > | N                           | ^              | n           | $\sim$ |
|       | - Î     | 1        | 1              | 1        | 15  | <b>S</b> 1 | US   | 1           | ? | 0                           |                | 0           | DEL    |

$$A \rightarrow 65$$
$$B \rightarrow 66$$
$$Z \rightarrow 90$$
$$a \rightarrow 97$$

$$b \rightarrow 98$$

$$z \rightarrow 123$$

Ascii – 7 bits

**USASCII code chart** 

| D7<br>D6<br>B | 5 -     |          |          |                |     | ° ° °      | °°,  | 0<br> <br>0 | 0<br> <br> | <sup>1</sup> 0 0 | י<br>ס | 1<br>1<br>0 | 1      |
|---------------|---------|----------|----------|----------------|-----|------------|------|-------------|------------|------------------|--------|-------------|--------|
|               | Þ4<br>† | b 3<br>1 | b 2<br>1 | Ь <sub>1</sub> | Row | 0          | ł    | 2           | 3          | 4                | 5      | 6           | 7      |
|               | 0       | 0        | 0        | 0              | 0   | NUL        | DLE  | SP          | 0          | 0                | Р      | ``          | Ρ      |
|               | 0       | 0        | 0        | 1              | 1   | SOH        | DC1  | !           | 1          | A                | Q      | 0           | q      |
|               | 0       | 0        | 1        | 0              | 2   | STX        | DC2  | 01          | 2          | B                | R      | . Þ         | r      |
|               | 0       | 0        | 1        |                | 3   | ΕΤΧ        | DC 3 | #           | 3          | C                | S      | С           | S      |
|               | 0       | 1        | 0        | 0              | 4   | EOT        | DC4  | \$          | 4          | D                | Т      | d           | t      |
|               | 0       |          | 0        | 1              | 5   | ENQ        | NAK  | %           | 5          | E                | U      | e           | U      |
|               | 0       | 1        | 1        | 0              | 6   | ACK        | SYN  | 8           | 6          | F                | V      | f           | v      |
|               | 0       | -        | 1        | 1              | 7   | 8EL        | ETB  | ,           | 7          | G                | W      | g           | W      |
|               | 1       | 0        | 0        | 0              | 8   | BS         | CAN  | (           | 8          | н                | X      | h           | ×      |
|               |         | 0        | 0        | 1              | 9   | нт         | EM   | )           | 9          | 1                | Y      | i           | У      |
|               |         | 0        | 1        | 0              | 10  | LF         | SUB  | *           | :          | J                | Z      | j           | Z      |
|               | 1       | 0        | 1        |                |     | VT         | ESC  | +           |            | к                | C      | k           | [ {    |
| i             | I       | 1        | 0        | 0              | 12  | FF         | FS   | •           | <          | L                | N      | 1           | 1      |
|               | ł       | 1        | 0        |                | 13  | CR         | GS   | -           | Ħ          | м                | )      | m           | }      |
|               | 1       | 1        | 1        | 0              | 4   | SO         | RS   |             | >          | N                | ^      | n           | $\sim$ |
|               | -1      | 1        | 1        | 1              | 15  | <b>S</b> 1 | US   | 1           | ?          | 0                |        | 0           | DEL    |

 $A \rightarrow 65 = 100 0001$  $B \rightarrow 66 = 100 0010$  $Z \rightarrow 90 = 101 1010$ 

$$a \rightarrow 97 = 110\ 0001$$
  
 $b \rightarrow 98 = 110\ 0010$   
 $z \rightarrow 123 = 111\ 1010$ 

#### Aside – End of Line Characters

USASCII code chart

| 07 <u>р</u><br>в | 5 -     |          |          |                |        | ° ° °      | °°,  | 0<br> <br>0 | 0 | <sup>1</sup> 0 0 | <sup>1</sup> 0 | 1<br>1<br>0 | 1      |   |
|------------------|---------|----------|----------|----------------|--------|------------|------|-------------|---|------------------|----------------|-------------|--------|---|
|                  | Þ4<br>† | b 3<br>1 | b 2<br>1 | Ь <sub> </sub> | Column | 0          | I    | 2           | 3 | 4                | 5              | 6           | 7      | $LF = Line Feed = \n$   |
|                  | 0       | 0        | 0        | 0              | 0      | NUL .      | DLE  | SP          | 0 | 0                | Р              | `           | Р      |   |
|                  | 0       | 0        | 0        | 1              | ł      | SOH        | DC1  | !           | 1 | Α.               | Q              | ٥           | Q      | CR - Carriage Return - r  |
|                  | 0       | 0        | 1        | 0              | 2      | STX        | DC2  | - 11        | 2 | B                | R              | b           | r      | $\int C \mathbf{K} - C a \Pi a g c \mathbf{K} c t u \Pi - \eta c$ |
|                  | 0       | 0        | 1        | I              | 3      | ETX        | DC 3 | #           | 3 | C                | S              | С           | S      |   |
|                  | 0       | 1        | 0        | 0              | 4      | EOT        | DC4  | •           | 4 | D                | Т              | d           | t      |   |
|                  | 0       | Ι        | 0        | 1              | 5      | ENQ        | NAK  | %           | 5 | E                | υ              | e           | U      |   |
|                  | 0       | 1        | 1        | 0              | 6      | ACK        | SYN  | 8           | 6 | F                | V              | f           | V      | End of Line Defaults  |
|                  | 0       | -        | 1        | 1              | 7      | BEL        | ETB  | '           | 7 | G                | W              | g           | w      |   |
|                  | 1       | 0        | 0        | 0              | 8      | BS         | CAN  | (           | 8 | н                | X              | h           | ×      | • Univ: \n  |
|                  | -       | 0        | 0        | 1              | 9      | нт         | EM   | )           | 9 | 1                | Y              | i           | У      |   |
|                  |         | 0        | 1        | 0              | 10     | LF         | SUB  | *           |   | J                | Z              | j           | z      | <b>XX7' 1</b> \ \   |
|                  | 1       | 0        |          |                |        | VT         | ESC  | +           | ; | к                | C              | k           | [ {    | • Windows: $r n$  |
|                  | ł       | 1        | 0        | 0              | 12     | FF         | FS   |             | < | L                | <u>۱</u>       | 1           | 1      |   |
|                  | ł       |          | 0        |                | 13     | CR         | GS   | -           | = | м                | 2              | m           | }      |   |
|                  | I       | 1        | 1        | 0              | 4      | SO         | RS   |             | > | N                | ~              | n           | $\sim$ |   |
|                  | -1      | 1        | I        | 1              | 15     | <b>S</b> 1 | US   | 1           | ? | 0                |                | 0           | DEL    |   |

## Latin-1 (ISO-8859-1): Ascii + an extra bit

| 8_<br>128 |              |           |           |           |           |           |               |           |          |           |           |           |      |             |                             |           |
|-----------|--------------|-----------|-----------|-----------|-----------|-----------|---------------|-----------|----------|-----------|-----------|-----------|------|-------------|-----------------------------|-----------|
| 9_<br>144 |              |           |           |           |           |           |               |           |          |           |           |           |      |             |                             |           |
| A_<br>160 | NBSP<br>00A0 | i<br>00A1 | ¢<br>00A2 | £<br>00A3 | ¤<br>00A4 | ¥<br>00A5 | <br> <br>00A6 | §<br>00A7 | <br>00A8 | ©<br>00A9 | a<br>00AA | «<br>00AB |      | SHY<br>00AD | (R)<br>00AE                 | —<br>00AF |
| В_        | 0            | ±         | 2         | 3         | ,         | μ         | ¶             |           | ,        | 1         | 0         | »>        | 1⁄4  | 1/2         | <sup>3</sup> / <sub>4</sub> | ز         |
| 176       | 00B0         | 00B1      | 00B2      | 00B3      | 00B4      | 0085      | 00B6          | 00B7      | 00B8     | 00B9      | 00BA      | 00BB      | 00BC | 00BD        | 00BE                        | 00BF      |
| C_        | À            | Á         | Â         | Ã         | Ä         | Å         | Æ             | Ç         | È        | É         | Ê         | Ё         | Ì    | Í           | Î                           | Ї         |
| 192       | 00C0         | 00C1      | 00C2      | 00С3      | 00C4      | 00C5      | 00C6          | 00C7      | 00C8     | 00С9      | 00ca      | 00св      | 00CC | 00CD        | 00ce                        | 00СF      |
| D_        | Ð            | Ñ         | Ò         | Ó         | Ô         | Õ         | Ö             | ×         | Ø        | Ù         | Ú         | Û         | Ü    | Ý           | Þ                           | ß         |
| 208       | 00D0         | 00D1      | 00D2      | 00D3      | 00D4      | 00D5      | 00D6          | 00D7      | 00D8     | 00D9      | 00da      | 00db      | 00DC | 00dd        | 00DE                        | øødf      |
| E_        | à            | á         | â         | ã         | ä         | å         | æ             | Ç         | è        | é         | ê         | ё         | Ì    | í           | Î                           | ї         |
| 224       | 00E0         | 00E1      | 00E2      | 00E3      | 00E4      | 00E5      | 00E6          | 00E7      | 00E8     | 00E9      | 00ea      | 00ЕВ      | 00ЕС | 00ED        | 00EE                        | 00еғ      |
| F_        | ð            | ñ         | ò         | ó         | â         | õ         | ö             | ÷         | ø        | ù         | ú         | û         | ü    | ý           | b                           | ÿ         |

#### ISO-8859-7: A different ASCII extension

| 8_<br>128              |                                |                                |                   |                                |                                |                                |                                |                                |                                |                                |                                |                                |                                |                   |                                |                          |
|------------------------|--------------------------------|--------------------------------|-------------------|--------------------------------|--------------------------------|--------------------------------|--------------------------------|--------------------------------|--------------------------------|--------------------------------|--------------------------------|--------------------------------|--------------------------------|-------------------|--------------------------------|--------------------------|
| 9_<br>144              |                                |                                |                   |                                |                                |                                |                                |                                |                                |                                |                                |                                |                                |                   |                                |                          |
| A_<br>160              | NBSP<br>00A0                   | ,<br>2018                      | ,<br>2019         | £<br>00A3                      | €<br>20AC                      | <b>Д</b> р<br>20АF             | <br> <br>00A6                  | §<br>00A7                      | <br>00A8                       | ©<br>00A9                      | 037A                           | <b>«</b><br>00Ab               |                                | SHY<br>00AD       |                                |                          |
| B_<br>176              | 0<br>00B0                      | ±<br>00B1                      | 2<br>00B2         | 3<br>00B3                      | ,<br>0384                      | •^<br>0385                     | ́А<br>0386                     | 00B7                           | Έ<br>0388                      | Н<br>0389                      | ́Т<br>038А                     | »»<br>00bb                     | Ю<br>038С                      | 1/2<br>00BD       | ′Ү<br>038Е                     | Ώ<br>038F                |
| C_<br>192              | ΰ                              | А                              | В                 | Г                              | Δ                              | Е                              | Z                              | Н                              | Θ                              | Ι                              | K                              | Λ                              | М                              | N                 | Ξ                              | Ο                        |
|                        | 0390                           | 0391                           | 0392              | 0393                           | 0394                           | 0395                           | 0396                           | 0397                           | 0398                           | 0399                           | 039A                           | 039B                           | 039C                           | 039D              | 039E                           | 039F                     |
| D_<br>208              | 0390<br>П<br>03A0              | 0391<br>P<br>03A1              | 0392              | 0393<br><u>Σ</u><br>03A3       | 0394<br>Т<br>03А4              | 0395<br>Y<br>03A5              | 0396<br>Ф<br>03А6              | 0397<br>X<br>03A7              | 0398<br>Ψ<br>03A8              | 0399<br>Ω<br>03A9              | 039А<br>Ї<br>03АА              | 039В<br><u>Ÿ</u><br>03АВ       | 039C<br>ά<br>03AC              | 039D<br>É<br>03AD | 039E<br>ή<br>03AE              | 039F<br><u>í</u><br>03AF |
| D_<br>208<br>E_<br>224 | 0390<br>П<br>03A0<br>Ü<br>03B0 | 0391<br>P<br>03A1<br>α<br>03B1 | 0392<br>β<br>03B2 | 0393<br>Σ<br>03A3<br>γ<br>03B3 | 0394<br>Τ<br>03A4<br>δ<br>03B4 | 0395<br>Y<br>03A5<br>E<br>03B5 | 0396<br>Φ<br>03A6<br>ζ<br>03B6 | 0397<br>X<br>03Α7<br>η<br>03Β7 | 0398<br>Ψ<br>03A8<br>θ<br>03B8 | 0399<br>Ω<br>03A9<br>1<br>03B9 | 039A<br>Ï<br>03AA<br>K<br>03BA | 039B<br>ϔ<br>03AB<br>λ<br>03BB | 039C<br>ά<br>03AC<br>μ<br>03BC | 039D              | 039E<br>ή<br>03AE<br>ξ<br>03BE | 039F                     |

## Unicode: A single mapping to rule them all



Standard that maps characters to *codepoints* 

Unicode 14.0 is releasing next week!

- Overall: 144,697 characters, 159 scripts
- New: 838 characters, 5 scripts, 37 emojis



## Unicode Encodings: UTF-8

- Variable length encoding: Represents each *codepoint* as **1-4 bytes**
- Compatible with ASCII
- Python 3's default text encoding

## Unicode Encodings: UTF-16

- Variable length encoding: Represents each *codepoint* as **2 or 4** bytes
- Not compatible with ASCII

**Q:** When might we prefer UTF-16 over UTF-8?

### UTF-8 vs. UTF-16

#### <u>UTF-8 > UTF-16</u>

- Need ASCII compatibility
- More efficient for languages that can be represented by 1-2 bytes in UTF-8 (e.g. English, Spanish, Greek)

#### <u>UTF-16 > UTF-8</u>

• Majority of characters would be represented by 3 bytes in UTF-8, these are represented by 2 bytes in UFT-16 (e.g. Chinese, Japanese, Korean, Vietnamese)



# Regular Expressions

## **Regular Expressions**

- A formal language for specifying specific text strings
- How can we search for any of these?

woodchucks woodchucks Woodchuck Woodchucks



## Regular Expressions: Character Classes

• Disjunction of listed characters

| Pattern      | Matches              |
|--------------|----------------------|
| [wW]oodchuck | Woodchuck, woodchuck |
| [1234567890] | Any digit            |

• Ranges

| Pattern | Matches              | Example                         |
|---------|----------------------|---------------------------------|
| [A-Z]   | An upper case letter | <u>D</u> renched Blossoms       |
| [a-z]   | A lower case letter  | <u>m</u> y beans were impatient |
| [0-9]   | A single digit       | Chapter 1: Down the Rabbit Hole |

## Regular Expressions: Negation

• Negation of class: [^Ss] (^ only means negation in first position)

| Pattern | Matches                  | Example                            |
|---------|--------------------------|------------------------------------|
| [^A-Z]  | Not an upper case letter | O <u>y</u> fn pripetchik           |
| [^Ss]   | Neither S nor s          | <u>I</u> have no exquisite reason" |
| [^e^]   | Neither e nor ^          | <u>L</u> ook here                  |
| a^b     | The pattern a^b          | Look up <u>a^b</u> now             |

## Regular Expressions: More disjunction

• Woodchuck is another name for groundhog

| Pattern                   | Matches       |
|---------------------------|---------------|
| groundhog woodchuck       |               |
| yours mine                | yours<br>mine |
| a b c                     | = [abc]       |
| [gG]roundhog [wW]oodchuck |               |



## Regular Expressions: ? \* + .

| Pattern | Matches                    | Examples  |
|---------|----------------------------|---|
| colou?r | Optional previous char     | <u>color</u> <u>colour</u>                        |
| oo*h!   | 0 or more of previous char | <u>oh!</u> <u>ooh!</u> <u>oooh!</u>               |
| o+h!    | 1 or more of previous char | <u>oh!</u> <u>ooh!</u> <u>oooh!</u> <u>ooooh!</u> |
| baa+    |                            | <u>baa baaa baaaa baaaaa</u>                      |
| beg.n   |                            | <u>begin begun began beg1n</u>                    |

# Demo / Activity