CS6041 HW1, Noah Gardner, 000843905

1 Assignment 1

1. Please write regular expressions for the following. Points [20]

    (a) All binary strings.

    Examples: 1001, 1011, 1111

    Answer: `^[01]+$`

    (b) The email address contains only letters, @, and \. Symbols (both lower and upper cases).

    Examples: alice@gmail.com, bob@yahoo.com

    Answer: `^[a-zA-Z0-9]+[@][a-zA-Z0-9]+\.[a-zA-Z0-9]+$`

    (c) Valid integer numbers.

    Examples: 1, 12843, -89232, +1262

    Answer: `^[-+]?[0-9]+$`

    (d) Valid phone numbers that contain ten (10) digits.

    Formats: xxx-xxx-xxxx, (xxx) xxx-xxxx

    Examples: 453-126-4570, (453) 126-4560

    Answer: `^\([0-9]{3}\)[0-9]{3}-[0-9]{4}$|^[0-9]{3}-[0-9]{3}-[0-9]{4}$`

2. Determine the number of tokens and vocabulary, and types from the below text. Please list them in your answer too. [Points 5]

    Text: ”The quick brown fox jumps over the lazy dog.”

    Answer: Tokenization

    Tokens: `"The" "quick" "brown" "fox" "jumps" "over" "the" "lazy" "dog"`

    Number of tokens: 9

    Vocabulary: `"the" "quick" "brown" "fox" "jumps" "over" "lazy" "dog"`

    Types: 8

3. Write down all the steps of text normalization and give an example for each step. [Points 5]

    Answer: Steps of text normalization

    1. Segmenting or tokenizing the text.

       Example: ”The dog eats.” → `["The", "dog", "eats"]`

    2. Normalizing word formats.

       Example: Contractions are replaced with seperate tokens. **we’re** → **we**, **are**

    3. Segmenting sentences from the text.

Example: ”The dog eats. The cat sleeps.” → `[["The", "dog", "eats"] , ["The", "cat", "sleeps"]]`

4. We know how to compute similarity distance between two given strings using the edit distance algorithm. [Points 25]

   (a) Please write down the distance matrix for the following strings. Consider space ”” as a single character. [Points 15]

   String 1: Spokesman confirms

   String 2: Spokeswoman said

   Answer: See Table 1 in the appendix.

   (b) List down all the operations you need to perform. Please show backtracing matrix to validate your answer for the above example strings. [Points 10]

   Answer: See Tables 2 and 3 in the appendix.

   Edits

   0. `Spokes` matches, no move.

   1. Insert `w` (Cost 1)

   2. Insert `o` (Cost 1). `man(space)` matches, no move.

   4. Replace `c` with `s` (Cost 2).

   6. Replace `o` with `a` (Cost 2).

   8. Replace `i` with `n` (Cost 2).

   9. Delete `f` (Cost 1). `i` matches, no move.

   11. Replace `r` with `(space)` (Cost 2).

   13. Replace `m` with `(space)` (Cost 2).

   14. Delete `s` (Cost 1).

5. Please formulate your language model for the following text. Show the details of your LM formulation. [Points 25]

   Text: ”The day was grey and bitter cold, and the dogs would not take the scent. The big black bitch had taken one sniff at the bear tracks, backed off, and skulked back to the pack with her tail between her legs.”

   (a) Unigram model [Points 10]

   Answer: Unigram Model

$$\begin{aligned}
\text{p(and)} &= 3/41 \\
\text{p(at)} &= 1/41 \\
\text{p(back)} &= 1/41 \\
\text{p(backed)} &= 1/41 \\
\text{p(bear)} &= 1/41 \\
\text{p(between)} &= 1/41 \\
\text{p(big)} &= 1/41
\end{aligned}$$

$$\begin{aligned}
\text{p(bitch)} &= 1/41 \\
\text{p(bitter)} &= 1/41 \\
\text{p(black)} &= 1/41 \\
\text{p(cold)} &= 1/41 \\
\text{p(day)} &= 1/41 \\
\text{p(dogs)} &= 1/41 \\
\text{p(grey)} &= 1/41 \\
\text{p(had)} &= 1/41 \\
\text{p(her)} &= 2/41 \\
\text{p(legs)} &= 1/41 \\
\text{p(not)} &= 1/41 \\
\text{p(off)} &= 1/41 \\
\text{p(one)} &= 1/41 \\
\text{p(pack)} &= 1/41 \\
\text{p(scent)} &= 1/41 \\
\text{p(skulked)} &= 1/41 \\
\text{p(sniff)} &= 1/41 \\
\text{p(tail)} &= 1/41 \\
\text{p(take)} &= 1/41 \\
\text{p(taken)} &= 1/41 \\
\text{p(the)} &= 6/41 \\
\text{p(to)} &= 1/41 \\
\text{p(tracks)} &= 1/41 \\
\text{p(was)} &= 1/41 \\
\text{p(with)} &= 1/41 \\
\text{p(would)} &= 1/41
\end{aligned}$$

(b) Bigram model [Points 15]

Answer: Bigram Model (token pairs with 0.0 and 1.0 probability omitted)

$$\begin{aligned}
\text{p(bear|the)} &= 1/6 \\
\text{p(big|the)} &= 1/6 \\
\text{p(bitter|and)} &= 1/3 \\
\text{p(day|the)} &= 1/6 \\
\text{p(dogs|the)} &= 1/6 \\
\text{p(legs|her)} &= 1/2 \\
\text{p(pack|the)} &= 1/6 \\
\text{p(scent|the)} &= 1/6 \\
\text{p(skulked|and)} &= 1/3 \\
\text{p(tail|her)} &= 1/2 \\
\text{p(the|and)} &= 1/3
\end{aligned}$$

6. You are given a training set of 30 numbers that consists of 21 zeros and 1 each of the other digits 1-9. Now we see the following test set: 0 0 0 0 0 3 0 0 0 0. What is the unigram perplexity? [Points 20]

   Answer:

   P(0) = 21/30

   P(3) = 1/30

Perplexity

$$PP(W) = (P(0)^9 * P(3)^1)^{-1/10} = ((21/30)^9 * (1/30))^{-1/10} = 1.93 \qquad (1)$$

2 Appendix

Table 1: 4a. Distance Matrix

	#	S	p	o	k	e	s	w	o	m	a	n		s	a	i	d		
#	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
S	1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
p	2	1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
o	3	2	1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
k	4	3	2	1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
e	5	4	3	2	1	0	1	2	3	4	5	6	7	8	9	10	11	12	13
s	6	5	4	3	2	1	0	1	2	3	4	5	6	7	8	9	10	11	12
m	7	6	5	4	3	2	1	2	3	2	3	4	5	6	7	8	9	10	11
a	8	7	6	5	4	3	2	3	4	3	2	3	4	5	6	7	8	9	10
n	9	8	7	6	5	4	3	4	5	4	3	2	3	4	5	6	7	8	9
	10	9	8	7	6	5	4	5	6	5	4	3	2	3	4	5	6	7	8
c	11	10	9	8	7	6	5	6	7	6	5	4	3	4	5	6	7	8	9
o	12	11	10	9	8	7	6	7	6	7	6	5	4	5	6	7	8	9	10
n	13	12	11	10	9	8	7	8	7	8	7	6	5	6	7	8	9	10	11
f	14	13	12	11	10	9	8	9	8	9	8	7	6	7	8	9	10	11	12
i	15	14	13	12	11	10	9	10	9	10	9	8	7	8	9	8	9	10	11
r	16	15	14	13	12	11	10	11	10	11	10	9	8	9	10	9	10	11	12
m	17	16	15	14	13	12	11	12	11	10	11	10	9	10	11	10	11	12	13
s	18	17	16	15	14	13	12	13	12	11	12	11	10	9	10	11	12	13	14

Table 2: 4b. Backtracing Distance Matrix (Part 1)

	*	S	p	o	k	e	s	w	o	m	a
*	0	1	2	3	4	5	6	7	8	9	10
S	1	↖0	↖←1	↖←2	↖←3	↖←4	↖←5	↖←6	↖←7	↖←8	↖←9
p	2	↖↑1	↖0	↖←1	↖←2	↖←3	↖←4	↖←5	↖←6	↖←7	↖←8
o	3	↖↑2	↖↑1	↖0	↖←1	↖←2	↖←3	↖←4	↖←5	↖←6	↖←7
k	4	↖↑3	↖↑2	↖↑1	↖0	↖←1	↖←2	↖←3	↖←4	↖←5	↖←6
e	5	↖↑4	↖↑3	↖↑2	↖↑1	↖0	↖←1	↖←2	↖←3	↖←4	↖←5
s	6	↖↑5	↖↑4	↖↑3	↖↑2	↖↑1	↖0	↖←1	↖←2	↖←3	↖←4
m	7	↖↑6	↖↑5	↖↑4	↖↑3	↖↑2	↖↑1	↖↑←2	↖↑←3	↖2	↖←3
a	8	↖↑7	↖↑6	↖↑5	↖↑4	↖↑3	↖↑2	↖↑←3	↖↑←4	↖↑3	↖2
n	9	↖↑8	↖↑7	↖↑6	↖↑5	↖↑4	↖↑3	↖↑←4	↖↑←5	↖↑4	↖↑3
	10	↖↑9	↖↑8	↖↑7	↖↑6	↖↑5	↖↑4	↖↑←5	↖↑←6	↖↑5	↖↑4
c	11	↖↑10	↖↑9	↖↑8	↖↑7	↖↑6	↖↑5	↖↑←6	↖↑←7	↖↑6	↖↑5
o	12	↖↑11	↖↑10	↖↑9	↖↑8	↖↑7	↖↑6	↖↑←7	↖6	↖↑←7	↖↑6
n	13	↖↑12	↖↑11	↖↑10	↖↑9	↖↑8	↖↑7	↖↑←8	↖↑7	↖↑←8	↖↑7
f	14	↖↑13	↖↑12	↖↑11	↖↑10	↖↑9	↖↑8	↖↑←9	↖↑8	↖↑←9	↖↑8
i	15	↖↑14	↖↑13	↖↑12	↖↑11	↖↑10	↖↑9	↖↑←10	↖↑9	↖↑←10	↖↑9
r	16	↖↑15	↖↑14	↖↑13	↖↑12	↖↑11	↖↑10	↖↑←11	↖↑10	↖↑←11	↖↑10
m	17	↖↑16	↖↑15	↖↑14	↖↑13	↖↑12	↖↑11	↖↑←12	↖↑11	↖10	↖↑←11
s	18	↖↑17	↖↑16	↖↑15	↖↑14	↖↑13	↖↑12	↖↑←13	↖↑12	↖↑11	↖↑←12

Table 3: 4b. Backtracing Distance Matrix (Part 2)

	n		s	a	i	d		
*	11	12	13	14	15	16	17	18
S	↖←10	↖←11	↖←12	↖←13	↖←14	↖←15	↖←16	↖←17
p	↖←9	↖←10	↖←11	↖←12	↖←13	↖←14	↖←15	↖←16
o	↖←8	↖←9	↖←10	↖←11	↖←12	↖←13	↖←14	↖←15
k	↖←7	↖←8	↖←9	↖←10	↖←11	↖←12	↖←13	↖←14
e	↖←6	↖←7	↖←8	↖←9	↖←10	↖←11	↖←12	↖←13
s	↖←5	↖←6	↖←7	↖←8	↖←9	↖←10	↖←11	↖←12
m	↖←4	↖←5	↖←6	↖←7	↖←8	↖←9	↖←10	↖←11
a	↖←3	↖←4	↖←5	↖←6	↖←7	↖←8	↖←9	↖←10
n	↖2	↖←3	↖←4	↖←5	↖←6	↖←7	↖←8	↖←9
	↖↑3	↖2	↖←3	↖←4	↖←5	↖←6	↖←7	↖←8
c	↖↑4	↖↑3	↖↑←4	↖↑←5	↖↑←6	↖↑←7	↖↑←8	↖↑←9
o	↖↑5	↖↑4	↖↑←5	↖↑←6	↖↑←7	↖↑←8	↖↑←9	↖↑←10
n	↖↑6	↖↑5	↖↑←6	↖↑←7	↖↑←8	↖↑←9	↖↑←10	↖↑←11
f	↖↑7	↖↑6	↖↑←7	↖↑←8	↖↑←9	↖↑←10	↖↑←11	↖↑←12
i	↖↑8	↖↑7	↖↑←8	↖↑←9	↖8	↖←9	↖←10	↖←11
r	↖↑9	↖↑8	↖↑←9	↖↑←10	↖↑9	↖↑←10	↖↑←11	↖↑←12
m	↖↑10	↖↑9	↖↑←10	↖↑←11	↖↑10	↖↑←11	↖↑←12	↖↑←13
s	↖↑11	↖↑10	↖9	↖←10	↖↑←11	↖↑←12	↖↑←13	↖↑←14