Homework One: Natural Language Processing

Quin'darius Lyles-Woods

September 8, 2021

1 Regular Expressions

1.1 Binary Strings

```
(0|1)*
```

1.2 Email

```
^\S+@\S+.\S+$
```

1.3 Integers

```
\d
```

1.4 Phone Number

```
^(\+\d{1,2}\s)?\(?\d{3}\)?[\s.-]\d{3}[\s.-]\d{4}$
```

2 Tokens and Vocabulary

Text:

> "The quick brown fox jumps over the lazy dog."

Tokens	9
Types	8

3 Text Normalization

Steps of Normalization:

Tokenization Turning text into words.

Example: My name is quin and I live in Georgia.

This contains nine tokens or words.

Lemmatization Reducing a collection of words to their root meaning

Example: My name is quin and I am living in Georgia. I lived in Georgia for a while.

Living and lived will be reduced to live.

Stemming Close to Lemmatization, with Stemming you're taking away the suffixes.

Example: My name is quin and I am living in Georgia. I lived in Georgia for a while.

Living and lived will be reduced to liv.

Sentence Segmentation Breaking up the text with the delimiters of . , ! or ?

Example: My name is quin and I am living in Georgia. I lived in Georgia for a while. I like living here.

Sentence One My name is quin and I am living in Georgia.

Sentence Two I lived in Georgia for a while.

Sentence Three I like living here.

Spelling Correction Checking for spelling mistakes.

Example: My name is quin afd I live in Georgia.

The word afd should be and.

Non-Standerd Words Phone numbers, emails, dates, etc.

Example: My name is quin and my phone number is 704-470-7036, and email is qlyleswo@student.kennesaw.edu.

Phone Number 704-470-7036

Email qlyleswo@student.kennesaw.edu

4 Similarity Distance with Edit Distance Algorithm

String One: Spokesman confirms
String Two: Spokeswoman said

4.1 Steps for Distance Matrix

You must compare the strings and see what needs to be replaced to get them to equal each other.
The operations are:

Insert	+1
Delete	+1
Replace	+2

4.2 Backtracing Matrix Check

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

There are some messed up values in the matrix because the program I wrote to compute it was a little off with the matrix manipulations.

4.3 Distance Matrix

The total is 13 for the edit distance.

5 Language Model

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."

5.1 Unigram Model

5.2 Bigram Model

6 Unigram Perplexity

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?