Homework One: Natural Language Processing

Quin'darius Lyles-Woods

September 6, 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

4 Similarity Distance with Edit Distance Algorithm

String One: Spokesman confirms
String Two: Spokeswoman said

4.1 Steps for Distance Matrix

4.2 Backtracing Matrix Check

4.3 Distance Matrix

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?