

# Assignment 1

FULL\_GITHUB\_REPOSITORY\_URL

Group 22

Matthew Evans  
mre042000

Ruochen Meng  
rxm220120

Saul Garay  
SAG180009

## 1 Implementation Details

```
import numpy
print("this is a piece of code")
```

Listing 1: Example of a Code Piece.

### 1.1 Unigram and bigram probability computation

### 1.2 Smoothing

### 1.3 Unknown word handling

1. Sort tokens in descending order by occurrence count.
2. Iterate over token counts, accumulating mass,  $M$ , i.e.,

$$M_{\text{cum}} = \sum_{t_i}^{t_n} C(t_i)$$

If  $(M_{\text{cum}}/M_{\text{total}}) < M_{\text{coverage}}$ , assign the token's mass to the <UNK> token.

### 1.4 Implementation of perplexity

## 2 Eval, Analysis and Findings

## 3 Others