# CSE 556: Natural Language Processing Assignment 1

Sanskar Ranjan
2021096

Saumil Lakra
2021097

Jeremiah Malsawmkima Rokhum
2021533

Vishal Singh
2021575

## 1. Evaluation for Task 1

### 1.1. Explanation for Task 1:

**Helper functions:**

- pair_count_gen: generates the pairwise count of words from the corpus
- merge: merges the hyphen separated words
- word_hyphen: inserts hyphen between the words and $ at the end

**class Tokeniser:**

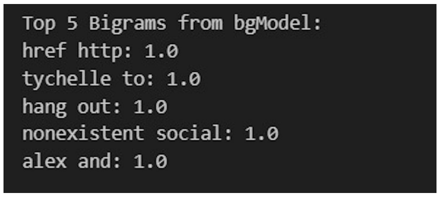
- word_dict_formation: creates a dictionary of word_hyphen and their count
- learn_vocabulary: generates vocabulary, merge rules and combines the words depending on the frequent characters.
- tokenise: tokenises the input string using the merge rules

Merge rules are writtes in merge_rules.txt and vocabulary is in tokens.txt

## 2. Evaluation for Task 2

### 2.1. Top 5 Bigrams:

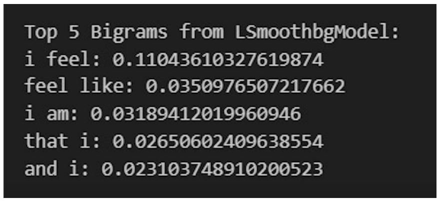
1. **Before Smoothing:**



```
Top 5 Bigrams from bgModel:
href http: 1.0
tychelle to: 1.0
hang out: 1.0
nonexistent social: 1.0
alex and: 1.0
```

Figure 1. Before Smoothing

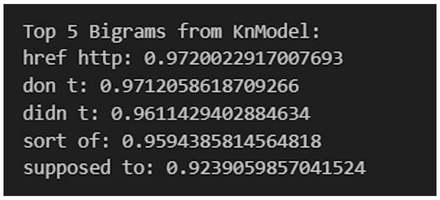
2. **After Laplace Smoothing:**



```
Top 5 Bigrams from LSmoothbgModel:
i feel: 0.11043610327619874
feel like: 0.0350976507217662
i am: 0.03189412019960946
that i: 0.02650602409638554
and i: 0.023103748910200523
```

Figure 2. After Laplace Smoothing

3. **After Kneser–Ney smoothing:**



```
Top 5 Bigrams from KnModel:
href http: 0.9720022917007693
don t: 0.9712058618709266
didn t: 0.9611429402884634
sort of: 0.9594385814564818
supposed to: 0.9239059857041524
```

Figure 3. After Kneser–Ney smoothing

### 2.2. Reasoning for method used for including emotion component

1. **Formula used for generating modified bigram probabilities:**

$$P(w_i|w_{i-1})_{emotion} = (count(w_i)/count(w_{i-1})) + \beta$$

$$\beta = emotion\_score[''emotion'']$$

2. **Explanation for the formula:** The Beta which corresponds to the emotional score of the bigram helps us generating emotion oriented samples . It helps us create 6 bigram models each corresponding to an emotion.

3. **Generation of the samples:** We save the bigram probability in such a way for each unigram word $w_i$ we can get all the

$$((count(w_i)/count(w_{i-1})) + \beta)$$

values from the dictionaries , we then normallise these values to calculate a probability space of words , from where we can choose the next word with random library and probability as the weights .

### 2.3. Two Generated Samples for each emotion

All of these samples have been taken from the gen_{emotion}.txt generated. It's present in the github link.

1. **Anger:** I smoke that were second chance to smother me up what i seems.

   I i is gone forever along those cracks by changing but seriously enough.

2. **Fear:** I sometimes it by being scared puff it scares me doubt that is.

   I the intensity of sharing my fears gotta stop caring in australia though.

3. **Joy:** I that keeps me feeling genuinely looking out or pleased but thank him.

   I the optimism of miles upon the optimism of bringing their creativity or.

4. **Love:** I that keeps me feeling genuinely looking out or pleased but thank him.

   I beautiful long outing yesterday that love hanging with great all the supporting.

5. **Sadness:** I have depression is damaged because it accelerated out books resonate with regret.

   I a tragic accident where going to hurt so unhappy with me morbid.

6. **Surprise:** I about saying im amazed seeing your suffering surely a portrayal of salt.

   I by the unexpected long and obstacles and curiosity is weird to witness.

### 2.4. Accuracy and macro F1 scores obtained from extrinsic evaluation

| Emotion | Accuracy | F1-Score |
|---|---|---|
| Joy | 0.46 | 0.63 |
| Sadness | 0.62 | 0.76 |
| Anger | 0.38 | 0.55 |
| Fear | 0.38 | 0.55 |
| Surprise | 0.72 | 0.83 |
| Overall | 0.51 | 0.53 |

Table 1. Accuracies and Macro F1 Scores

### 2.5. Reason for generation according to corresponding emotions

1. **Anger:** Instance: I smoke that were second chance to smother me up what i seems. **Reason:** There is a keyword ”smother”. This keyword makes the sentence more likely to contain an angry emotion.

2. **Fear:** Instance: I sometimes it by being scared puff it scares me doubt that is. **Reason**: There are keyword ”scared” and ”scares”. These keywords makes the sentence more likely to contain an emotion which depicts fear.

3. **Joy:** Instance: I the optimism of miles upon the optimism of bringing their creativity or. **Reason**: There are keywords ”optimism” and ”creativity”. These keywords makes the sentence more likely to contain a joyful emotion.

4. **Love:** Instance: I beautiful long outing yesterday that love hanging with great all the supporting. **Reason**: There are keywords ”beautiful”, ”love” and ”supporting”. These keywords makes the sentence more likely to contain an emotion of love.

5. **Sadness:** Instance: I have depression is damaged because it accelerated out books resonate with regret. **Reason**: There are keywords ”depression”, ”regret”. These keywords makes the sentence more likely to contain an emotion of sadness.

6. **Surprise:** Instance: I about saying im amazed seeing your suffering surely a portrayal of salt. **Reason**: There are keywords ”amazed”. This keyword makes the sentence more likely to contain emotions of surprise.

### 2.6. Credit Statement

Our contribution towards this assignment was fairly equal because, we discussed whenever each one of us faced issues and while proceeding with an idea.

1. **Saumil Lakra:** Task 1

2. **Jeremiah Rokhum:** Task 2: Q1 and Q2

3. **Sanskar Ranjan:** Task 2: Q3

4. **Vishal Singh:** Task 2: Q4