Natural Language Processing

Assignment: 1

Karan Gupta (2021258) | Raghav Sakhuja (2021274) | Shivesh Gulati (2021286) | Rahul Oberoi (2021555)

(Explanation for Task-2 Q2)

Laplace Smoothing provides an equal probability to every unseen bigram whereas Kneser Ney gives the probability proportional to its unigram probability for an unseen bigram.

We can see this in the CSVs generated where every unseen bigram has the same probability in Laplace while in Kneser Ney it is proportional to their unigram probability. Thus we feel that Laplace is a simpler but flawed smoothing technique compared to Kneser Ney which is much more complicated and resource-intensive and could give better results according to the requirements.

(1)

Top 5 bigrams **BEFORE** smoothing:

Word 1	Word 2	Probability
i	feel	0.03276
i	am	0.00943
feel	like	0.00795
i	was	0.00647
that	i	0.00557

Top 5 bigrams **AFTER** smoothing (Laplace):

Word 1	Word 2	Probability
i	feel	0.008993
i	am	0.002597
i	was	0.001784
i	have	0.001272
feel	like	0.001235

Top 5 bigrams **AFTER** smoothing (Kneser Ney):

Word 1	Word 2	Probability
i	feel	0.021986
i	am	0.006295
and	i	0.006122
feel	like	0.005417
that	i	0.004785

(2)

Reasoning for the method used for including emotion component:

For this, we have used the following formula:

$$P_{new}(w_i|w_{i-1})=\frac{1}{2}\left[\boldsymbol{\alpha}P(w_i|w_{i-1})+(1-\frac{\alpha}{2})\boldsymbol{\beta}_{uni}(w_i)+(1-\frac{\alpha}{2})\boldsymbol{\beta}_{bi}(w_{i-1},w_i)\right]$$

here ,we have normalized the emotion scores of ever unigram and bigram as $\boldsymbol{\beta}_{uni}$ and $\boldsymbol{\beta}_{bi}$. We use a hyperparameter $\boldsymbol{\alpha}$ which decides the weight that we will give to the emotion component and the learned component. The emotion scores were normalized to make sure that the distribution for unigram sums to 1, as well as the the sum of every bigram given its first word sums to 1.
We generate all the sentences by choosing words from the vocabulary according to distribution obtained from the above provided formula.

We have stored the emotion scores for unigrams and bigrams here:
https://drive.google.com/drive/folders/1VTeDDKB7I59rsqyEH856VDPOv5m8bO50?usp=sharing
As these CSVs are very large, hence they can be viewed from the above link however we were unable to add them to our submission folder due to size limitations.

For the value of $\boldsymbol{\alpha}$, we decided it to be 0.5. In this case, we are giving equal weightage to both unigram and bigram of 0.375 and 0.25 to the learned component.

(3)

2 generated samples for each of the emotion:

Sadness	alone dot crying cringe i emotional moved north i was ish starvation chuffed mourn for worse adn abandoning aided pray with sick harrass me	cinnamon disillusioned waste grave cya fade medication because rupture socially halloween dropping off december allen ungrateful pains weekends mean listless most returns garden working troubled and every pop broke midori
Joy	established indy convenience future satisfied jackets and members who was indonesia anyone i dazzling	independence lj entry rte prospects table i musicianship especially huge while gallery refresh memories building seen continues wildlife diverse americana kind of
Love	switch sympathetic ton of compassionate combined hive available for cherish sweetness soulmate and looking kindness coerce	adoring fans empathize with loves twitter ive officially baachan supportiveness comfort touched you lj toward warms favorite past sale christchurch longingly for tasty goodness hot stic supportive of
Anger	mock over want pissed off perceive that pools growing shirt little wiz khalifa bu causing spinning nerd beside prep hang university buy mother waving at my	i left happpy when there their horribly presence superman and overeat stamford bridge agreement i spoken
Fear	dried figs immersed jasons insecurities i paralysed pulse floating eh i feel funny matches tiggers convoluted awaken from unknown deal turkish color mark of iced capp im earlier this	i pms avoid status headphones lights partly dependent among intimidated subconscious at parallel waited theories i
Surprise	rapidly antique activities freakishly neurotic patriotism only solo overall i enthralled	called everyone i and flickr plunge called upon surprising amount of deciding cleaned hollow hypocrite periods resemblance to amazingness of enthralled harmonious i feel liquid

(4)

Accuracy and Macro F1 Scores obtained from extrinsic evaluation:

Smoothing Technique	Accuracy Scores	Macro F1 Scores
No Smoothing	0.63	0.6182748978204119
Laplace Smoothing	0.6766666666666666	0.6761520851034918
KneserNey Smoothing	0.65	0.6451522232775957

Explanation for the parameters obtained:
In the given question, both the tf-idf vectoriser and support vector machine model had learnable parameters, and hence, a pipeline was constructed with the vectoriser followed by the model. The vectorised input is then passed to the SVM model, and the optimal parameters are found.

The optimal parameters found are:
1) For the vectoriser:
i) n_gram:(1,2) which means that the vectoriser considers both unigrams as well as bigrams.
ii) stop_words:"english" this ensures that common English stop words are ignored

2) For the SVC()
i) C (Regularization Constant): 1.4
ii) Kernel: Linear

(5)

All the sentences and their respective details are in evaluation.txt. For generating the sentence, we have given equal weight to emotion as well as the bigram occurrence in the corpus. This made the sentences lose some of their structure and, in compensation, gain a better emotional aspect.

(6)

Credits:

1) **Karan Gupta (2021258)**

   i) Task-2 part-1: Creation of bigram model
   ii) Task-2 part-3: Generation of emotion scores for bigrams
   iii) Evaluation-part 5

2) **Raghav Sakhuja (2021274)**

   i) Task-1 : learn_vocab()
   ii) Taks-2 part-3: Incorporation of emotion scores into the probability formula
   iii) Task-2 part-3: Generation emotion scores for unigrams
   iv) Task-2 part-4 a): Sentence generation

3) **Shivesh Gulati (2021286)**

   i) Task-1: tokenize()
   ii) Task-2: Q4) part b) SVM training and extrinsic evaluation

4) **Rahul Oberoi (2021555)**

   i) Task-2 Q2) Laplace and Kneser Ney Smoothing
   ii) Task-2) Evaluation) part-1)
   Top 5 bigram generation for without and with the smoothing techniques.