Siddhartha Shrivastava CSCI - 5832

NLP- HW3 Part 2 Prof. James Martin

For the Part 2, I chose the part 1, which has the goal of reducing the perplexity of the trigram model.

In the Base model, I didn’t use any smoothing or regex function, just keeping the base trigram on basis of splitting the words and forming them in pair of 3 and then finding the sentence probability as well as the words probability. On basis of that, I found the perplexity using the formula provided in the reference textbook.

The resultant perplexity was for the first part (simple trigram model):

4282.980669

3867.169585

3618.568548

4133.708739

3540.948070

4057.944919

4048.521084

3610.525918

3973.893687

4148.329153

***Here we can see; the average perplexity is around 3950.***

***Using Option 1***

The goals of this part was to add smoothing and tokenization and reduce the resulting perplexity in comparison to the base model which I have achieved before.

Process:

In this model,

* Firstly, the character based splitting of the text is done, which included the special characters, as well as the whole string is converted to lower text format.
* Second, suffix-based tokenization is done which includes word but not limted to ‘ing’, ‘ous’, ’er’ etc.
* Each sentence has <s> added at the end as well at the starting of it.
* Add 1 Smoothing is added.

Using this model, I was able to bring down the model’s perplexity significantly with an average around 1000.

1 , Log probability=-262.646159

1 , perplexity=519.818636

----------------------------------

2 , Log probability=-550.837338

2 , perplexity=561.978536

----------------------------------

3 , Log probability=-189.346616

3 , perplexity=864.696974

----------------------------------

4 , Log probability=-647.674254

4 , perplexity=741.682279

----------------------------------

5 , Log probability=-184.953671

5 , perplexity=475.859373

----------------------------------

6 , Log probability=-447.618032

6 , perplexity=460.242654

----------------------------------

7 , Log probability=-397.456912

7 , perplexity=452.470121

----------------------------------

8 , Log probability=-794.652728

8 , perplexity=451.562276

----------------------------------

9 , Log probability=-190.294807

9 , perplexity=2021.898693

----------------------------------

10 , Log probability=-473.600783

10 , perplexity=788.737499

From above results, it can be concluded that the newer model has less perplexity than the base model.