**Assignment 2**

After applying the EM algorithm over the training and heldout data, we derived the parameters .

We can now evalutate the proability of finding a succession of tokens in a real-life application with the following formula:

From now on, we are going to call this probability distribution as .

We have done this procedure both on the heldout data and training data. The operation done on the latter was done only for educational purposes, we get and .   
We expected this result because we are setting our parameters over the same set that we used to collect the original probabilities. Therefore, the 3-tuples () are always present when evaluating the probabilities and the trigram probabilities describe in the best way the problem we are modeling.   
Of course, we are not going to consider anymore this parameters as they clearly wrong for modeling our problem and not useful at all.

Once we have done we can go on evaluating the cross entropy:

Where is the cardinality of the training set.

The results are reported in the table on the next page.

**TEXTEN1.txt**

|  |  |
| --- | --- |
| **Entropy** | **Change in value** |
| 88137.6014 | Original value |
| 88493.4804 | +10% |
| 89330.2343 | +20% |
| 94608.4049 | +30% |
| 96049.5001 | +40% |
| 98089.6062 | +50% |
| 101279.7688 | +60% |
| 110580.7608 | +70% |
| 136952.2796 | +80% |
| 141195.9171 | +90% |
| 145576.2492 | +95% |
| 155506.0895 | +99% |

**TEXTCZ1.txt**

|  |  |
| --- | --- |
| **Entropy** | **Change in value** |
| 168211.0804 | Original value |
| 168522.0838 | +10% |
| 169156.3521 | +20% |
| 170124.8735 | +30% |
| 171542.9428 | +40% |
| 174281.4058 | +50% |
| 186397.9781 | +60% |
| 190395.7276 | +70% |
| 200424.4683 | +80% |
| 207673.3589 | +90% |
| 214720.3263 | +95% |
| 229606.0168 | +99% |

Let us consider the original value of , we can easily see that by increasing it the cross entropy gets higher.  
We remind that the cross entropy is:

where is the true probability distribution.

After this remark, we can say that what increases is actually the Kullback-Leiber distance , since is a constant.

This is totally expected because we are giving more importance to the trigram probability rather than any other n-gram probability. Since the probability of finding 3 tokens in the same exact succession is very low, the trigram probability is the one that assumes lower values and, therefore, that gives a huge contribution after taking the logarithm.   
By increasing the weight , the impact of this probability is even higher.

Now let us consider what happens after decreasing the value of the parameter :

**TEXTEN1.txt**

|  |  |
| --- | --- |
| **Entropy** | **Change in value** |
| 88137.60149173586 | Original value |
| 88112.01730237155 | -10% |
| 88108.07048049198 | -20% |
| 88127.71644700765 | -30% |
| 88173.90445271932 | -40% |
| 88251.08623425258 | -50% |
| 88366.22990362023 | -60% |
| 88531.08117940633 | -70% |
| 88768.25690477288 | -80% |
| 89134.53410179449 | -90% |
| 90053.43365921086 | -100% |

**TEXTCZ1.txt**

|  |  |
| --- | --- |
| **Entropy** | **Change in value** |
| 168211.0804 | Original value |
| 168190.9293 | -10% |
| 168193.0948 | -20% |
| 168220.2642 | -30% |
| 168276.2557 | -40% |
| 168366.7144 | -50% |
| 168500.5058 | -60% |
| 168692.9106 | -70% |
| 168974.629 | -80% |
| 169428.8586 | -90% |
| 171076.8593 | -100% |

In this case, instead, we can notice that something unusual happened. By decreasing the value of , we say a decrease in the value of the cross entropy, therefore the 2 probabilities distributions and are more similar to each other.

That is the result of the fact that we set our parameters over the heldout data which, of course, are different from the test data.  
Anyway we can observe that the difference from the cross entropy we get from our model and the one where is 20%/10% of its original value is not very big.   
Thus, we can say that the loss we obtained is totally acceptable.