

analysis-cross-entropy

January 31, 2021

1 Lambdas found by EM (with $\epsilon = 0.0001$)

- As expected, when we run EM on *train data*, we get $\lambda_3 = 1$.
- This is because this data has the same sequence of words, so the probabilities given by 3-gram is the highest (or at least it is newer lower than 2-gram)

	dataset	from	10	11	12	13
0	cz	train	0.0	0.0	0.000024	0.999976
2	en	train	0.0	0.0	0.000023	0.999977

- On the other hand, when we use the *holdout data*, the result is different. Uniform probability is not very probable, which makes sense.
- We can see that the Czech language has much more weight on p_1 . Most probably, because the Czech language has a much larger vocabulary and in the *train data*, there are not enough combinations to even correctly model the p_2 . In the English language, there are fewer words, so probably the p_2 gives the best probability.
- Both languages do not have λ_3 the highest. I would say that if we had much more good training data, then maybe the λ_3 would be a bit higher.

	dataset	from	10	11	12	13
1	cz	holdout	0.252835	0.442840	0.242932	0.061393
3	en	holdout	0.098392	0.264178	0.507790	0.129639

2 Cross-Entropy on the test data

- First, let's look at the default result (with unmodified λ_3).
- As stated above, Czech 2-grams and 3-grams are probably not very accurate for a lot of words, so 1-grams have the highest weight. Thus this leads to higher entropy than in the case of English.

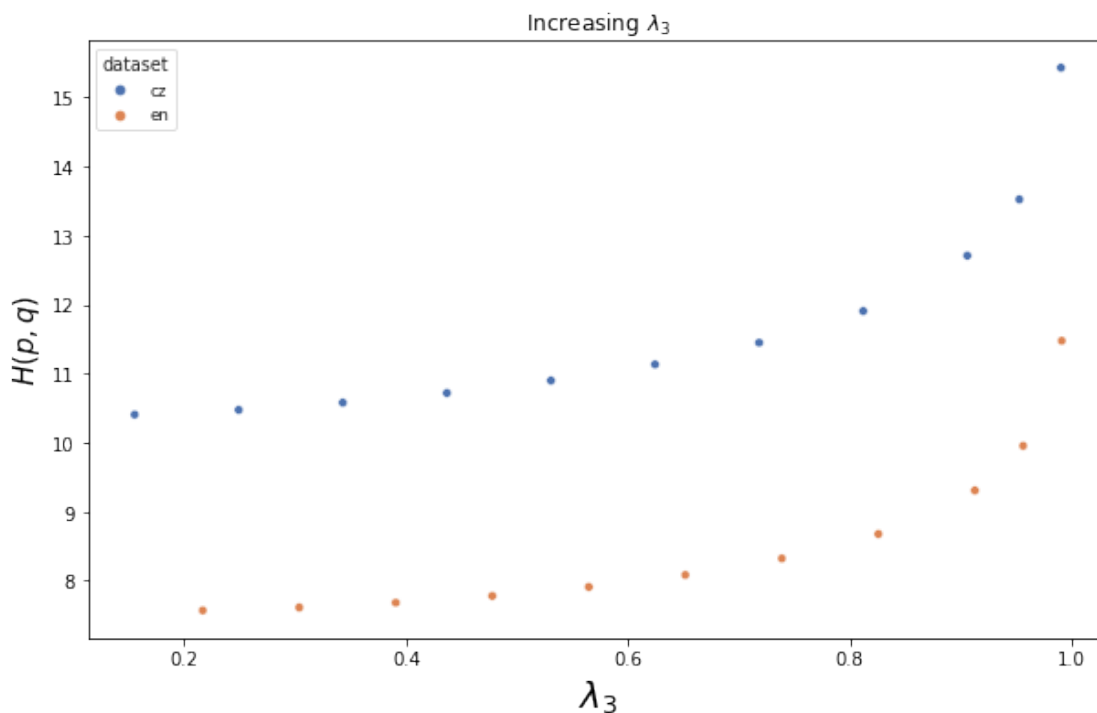
	dataset	modif	modif_val	L3_val	cross_entropy
0	cz	default	0.0	0.061393	10.393616
22	en	default	0.0	0.129639	7.560160

2.1 Increasing and decreasing the λ_3 and proportionally changing $\lambda_0, \lambda_1, \lambda_2$

- No we will increase and decrease the importance of the 3–gram and we will observe how the Cross-Entropy will change.

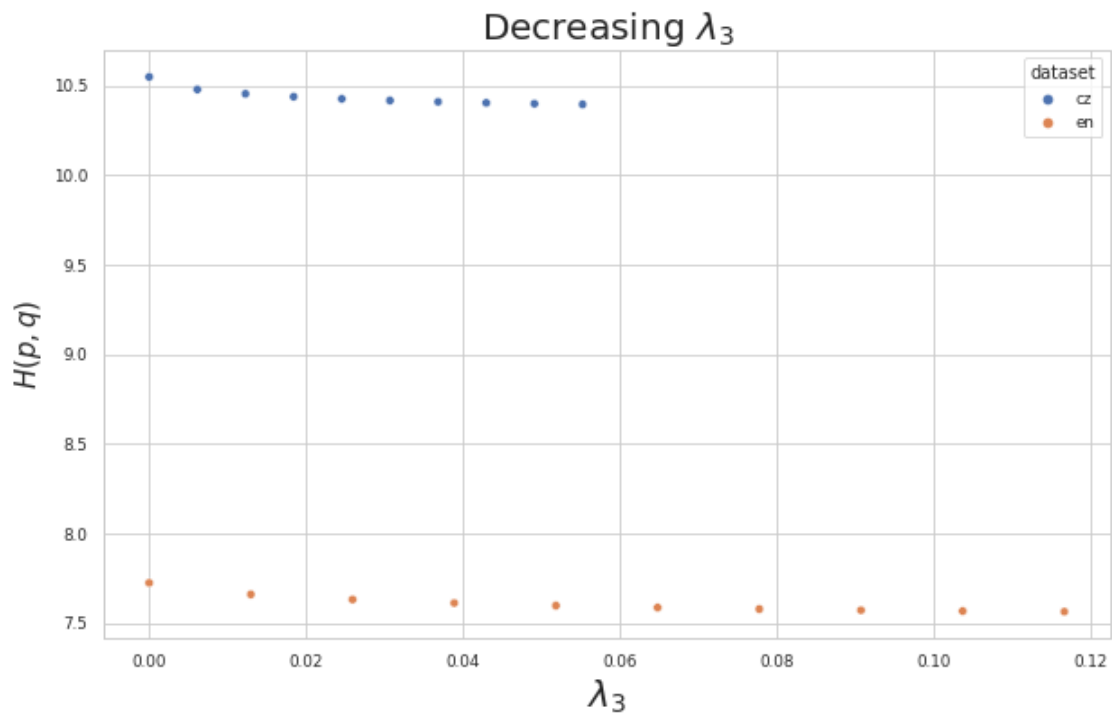
2.1.1 Increasing the λ_3

- In the figure below, we can see the results. We can see that for both languages, the increase of the λ_3 leads to higher entropy. This is probably caused because oftentimes the 3–gram does not know the answer producing $1/|V|$ and when most of the weight is on it, we are getting higher entropy. (But if we would perform the experiment on the *train data*, then the entropy would decrease to 0 ...)



2.1.2 Decreasing the λ_3

- On the figure below, we can see the results. We can see that for both languages, decrease of the λ_3 also leads to higher entropy.
- This means, that the value assigned to the 3–gram is probably quite good. We can see, that both increase and decrease of the λ_3 leads to increase of cross entropy on the *test data*.
- This would suggest that the EM (using the *holdout data*) found quite optimal values of λ_i and it produces lowest entropy on *test data* also (at least w.r.t. increase or decrease of λ_3)



2.1.3 Nicely plot of both decrease and increase together

- I Smoothed the results and plot it together to see the results better

