

CS769 Advanced Natural Language Processing
Homework 3

March 09, 2010

Answer 1.1

There are 9,723 word types in the vocabulary.

Answer 1.2

The perplexity of the unigram LM is 633.84 on test.text and 679.24 on training.text.

Answer 1.3

The perplexity of the bigram LM is 224.41 on test.text and 105.07 on training.text.

Answer 1.4

The perplexity of the trigram LM is 202.66 on test.text and 28.87 on training.text.

Answer 1.5

There are several things you can say -

Since the model was trained from the training text, you are overfitting if you measure the model against the same data; the model fits the data as well as possible (within the class of models selected,) but this perplexity does not tell you as much about the perplexity of new data in the future.

Also, the n-gram counts in the training data determine the model which best fits them, but other data will have slightly different counts, giving them higher perplexity.

Answer 1.6

You should have these numbers, or something close to them:

the	0.0508201
moive	0.00350727
mulan	5.6725e-05
album	6.91769e-06

Note that "the" occurs once in every 20 words; "movie" occurs once in about every 300.

Answer 2

The Maximum A Posteriori (MAP) estimate is

$$\begin{aligned}\theta^{MAP} &= \arg \max_{\theta} p(\theta|c_{1:V}) \\ &= \arg \max_{\theta} p(c_{1:V}|\theta)p(\theta)\end{aligned}$$

If we assume the prior as Dirichelet distribution with parameter $\epsilon + 1$,

$$p(\theta) = \frac{1}{B(\epsilon + 1)} \prod_{i=1}^V \theta_i^{\epsilon}$$

where $\frac{1}{B(\epsilon+1)}$ is a constant once the ϵ is given. Then, the MAP estimate is

$$\begin{aligned}\theta^{MAP} &= \arg \max_{\theta} p(c_{1:V}|\theta)p(\theta) \\ &= \arg \max_{\theta} \binom{|C|}{c_1 \cdots c_V} \prod_{i=1}^V \theta_i^{c_i} \frac{1}{B(\epsilon+1)} \prod_{i=1}^V \theta_i^{\epsilon} \\ &= \arg \max_{\theta} Z \prod_{i=1}^V \theta_i^{c_i+\epsilon}\end{aligned}$$

where Z is the constant regarding θ . So the objective function is

$$\theta^{MAP} = \arg \max_{\theta} \prod_{i=1}^V \theta_i^{c_i+\epsilon}$$

with the constraints $\theta_i \geq 0$ and $\sum_{i=1}^V \theta_i = 1$. Since the log function is monotonic, we apply log to the objective function and the Lagrangian is

$$L(\theta) = \sum_{i=1}^V (c_i + \epsilon) \log \theta_i - \beta \left(\sum_{i=1}^V \theta_i - 1 \right)$$

This is a concave function, and we set the gradient to zero:

$$\begin{aligned}\frac{\partial L}{\partial \theta_i} &= \frac{c_i + \epsilon}{\theta_i} - \beta = 0 \\ \frac{\partial L}{\partial \beta} &= \sum_{i=1}^V \theta_i - 1 = 0\end{aligned}$$

which gives the

$$\theta^{MAP} = \frac{c_i + \epsilon}{|C| + V\epsilon}$$

Answer 3.1

$P(English)$	$P(Spanish)$	$P(Japanese)$
0.333333333	0.333333333	0.333333333

$$\log(P(English))=\log(P(Spanish))=\log(P(Japanese))=-1.0986122886681098$$

	$P(c English)$	$P(c Spanish)$	$P(c Japanese)$
	0.179123201	0.168155771	0.123368009
a	0.060147894	0.104504282	0.131676325
b	0.011158062	0.008256824	0.010891573
c	0.021523835	0.037525417	0.005515604
d	0.021986003	0.039743669	0.017244991
e	0.105308332	0.113746996	0.060182923
f	0.018948897	0.008626533	0.003909795
g	0.017496369	0.007209317	0.014033373
h	0.047207183	0.004559739	0.031767088
i	0.055394163	0.049849036	0.09697689
j	0.001452529	0.006654754	0.002373804
k	0.00376337	3.08E-04	0.057390212
l	0.02898455	0.05292994	0.001466173
m	0.020533474	0.02581798	0.039796132
n	0.057903077	0.054162302	0.056692034
o	0.064439456	0.072462875	0.091112197
p	0.016770104	0.024277528	9.08E-04
q	5.94E-04	0.007702261	1.40E-04
r	0.053809587	0.059276604	0.042798296
s	0.066156081	0.065746503	0.042169936
t	0.080087152	0.035615257	0.056971305
u	0.026673709	0.033705096	0.070585771
v	0.009309389	0.005915337	2.79E-04
w	0.015515648	1.23E-04	0.01975843
x	0.001188433	0.002526342	6.98E-05
y	0.013865047	0.007887116	0.014173008
z	6.60E-04	0.002711196	0.007749773

Answer 3.2

$$P(\text{English}|\text{e10.txt}) = e^{-7842.884999}$$

$$P(\text{Spanish}|\text{e10.txt}) = e^{-8453.481807}$$

$$P(\text{Japanese}|\text{e10.txt}) = e^{-8760.617499}$$

If we normalize them, the $p(\text{English} | \text{e10.txt})$ is very close to one, and the other two are close to zero.

Answer 3.3

The confusion matrix for the classifier is

	English	Spanish	Japanese
English	10	0	0
Spanish	0	10	0
Japanese	0	0	10

Answer 3.4

$P(\text{English})$	$P(\text{Spanish})$	$P(\text{Japanese})$
0.333333333	0.333333333	0.333333333

	$P(c \text{English})$	$P(c \text{Spanish})$	$P(c \text{Japanese})$
a	0.166698167	0.158171521	0.146341463
b	0.093933094	0.091693635	0.110397946
c	0.009450009	0.00903452	0.009719421
d	0.023436023	0.025755124	0.017971759
e	0.022869023	0.031283711	0.026590867
f	0.098658099	0.098705502	0.08765817
g	0.00982801	0.011192017	0.008068953
h	0.010962011	0.008899676	0.013753897
i	0.02986203	0.024137001	0.025673941
j	0.065394065	0.064994606	0.07518797
k	0.003591004	0.003236246	0.005868329
l	0.013419013	0.016316073	0.029891803
m	0.034776035	0.030070119	0.023106547
n	0.025893026	0.028047465	0.03465982
o	0.054810055	0.058117584	0.050797726
p	0.071253071	0.076186624	0.078122135
q	0.013041013	0.016450917	0.014120668
r	0.003402003	0.004180151	0.002567394
s	0.053865054	0.053667745	0.053548505
t	0.06010206	0.06175836	0.051164497
u	0.061803062	0.05231931	0.050430955
v	0.039312039	0.041531823	0.049697414
w	0.005481005	0.006067961	0.002934165
x	0.012474012	0.010787487	0.011553273
y	0.001323001	0.001078749	0.001833853
z	0.012474012	0.012270766	0.0124702
	0.001890002	0.004045307	0.005868329

$$P(\text{English}|\text{e10.txt}) \propto e^{-7960.991462}$$

$$P(\text{Spanish}|\text{e10.txt}) \propto e^{-8009.752265}$$

$$P(\text{Japanese}|\text{e10.txt}) \propto e^{-8122.557663}$$

The confusion matrix for the classifier is

	English	Spanish	Japanese
English	10	0	0
Spanish	0	10	0
Japanese	0	0	10

Answer 4.1

LM2 is better on TEST when the size of TRAIN and TEST approaches infinity.

Answer 4.2

If the underlying LMg is a unigram and we can train and test the model with infinite data, LM1 is as good as LM2. (If you explicitly said that LM1 is better because of computational efficiency, I also give you full credits.)

Answer 4.3

If the underlying LMg is a unigram and the size of TRAIN is small, LM1 is expected to better than LM2 on TEST.