

Stav Cohen 316492776

Ron Kozitsa 312544240

Question 2:

1) Unigram weights

Given total tokens $N = 28521$ and vocabulary size $V = 2666$ (excluding <unk>).

a) No smoothing: $P(w) = c(w)/N$, weight = $\log_{10}(P(w))$. For <unk> , weight = $-\infty$.

Word	c(w)	P(w)	log10 P(w)
alice	386	0.01353389	-1.8686
cat	35	0.00122717	-2.9111
cheshire	7	0.00024543	-3.6101
hatter	55	0.00192840	-2.7148
mad	15	0.00052593	-3.2791
queen	68	0.00238421	-2.6227
red	15	0.00052593	-3.2791
the	1641	0.05753655	-1.2401
white	30	0.00105186	-2.9780
<unk>	0	0	$-\infty$

b) Laplace (add-1) smoothing ($\alpha=1$): treat <unk> as an extra type.

$$P(w) = (c(w)+1) / (N + (V+1)). \quad P(\text{<unk>}) = 1 / (N + (V+1)).$$

Word	P_Laplace(w)	log10 P_Laplace(w)
alice	0.0124086187	-1.9063

cat	0.0011542901	-2.9377
cheshire	0.0002565089	-3.5909
hatter	0.0017955624	-2.7458
mad	0.0005130178	-3.2899
queen	0.0022123894	-2.6551
red	0.0005130178	-3.2899
the	0.0526484545	-1.2786
white	0.0009939720	-3.0026
<unk>	0.0000320636	-4.4940

c) Allocate 5 occurrences to <unk>: set $c(<\text{unk}>) = 5$ and $N' = N + 5$.

$$P(w) = c(w)/N', \quad P(<\text{unk}>) = 5/N'.$$

Word	P_unk5(w)	log10 P_unk5(w)
alice	0.0135315151	-1.8687
cat	0.0012269509	-2.9112
cheshire	0.0002453902	-3.6101
hatter	0.0019280656	-2.7149
mad	0.0005258361	-3.2791
queen	0.0023837902	-2.6227
red	0.0005258361	-3.2791
the	0.0575264671	-1.2401
white	0.0010516722	-2.9781
<unk>	0.0001752787	-3.7563

2) Bigram weights for the given bigrams

Let $c(w_1, w_2)$ be the bigram count and $c(w_1)$ the unigram count of the history word.

a) No smoothing: $P(w_2|w_1) = c(w_1, w_2)/c(w_1)$.

b) Laplace ($\alpha=1$): $P(w_2|w_1) = (c(w_1, w_2) + 1)/(c(w_1) + V)$.

c) Witten–Bell: with $T(w_1) = \#$ unique continuations after w_1 (given as “number of prefixes” in the sheet),

For seen bigrams: $P(w_2|w_1) = c(w_1, w_2)/(c(w_1) + T(w_1))$.

Bigram (w_1, w_2)	$c(w_1, w_2)$	$c(w_1)$	$T(w_1)$	$\log_{10} P_{NS}$	$\log_{10} P_{Lap}$	$\log_{10} P_{WB}$	P_{Lap}	P_{WB}
(the, queen)	65	1641	449	-1.4022	-1.8146	-1.5072	0.01532389	0.03110048
(white, rabbit)	22	30	5	-0.1347	-2.0690	-0.2016	0.00853116	0.62857143
(cheshire, cat)	5	7	3	-0.1461	-2.6488	-0.3010	0.00224467	0.50000000
(the, hatter)	51	1641	449	-1.5075	-1.9182	-1.6126	0.01207337	0.02440191
(alice, </s>)	59	386	133	-0.8157	-1.7064	-0.9443	0.01965924	0.11368015

3) Bigram weights for unseen bigrams (0 occurrences)

Laplace: $P = (0+1)/(c(w_1)+V)$.

Witten–Bell (unseen): $P = (T(w_1)/(c(w_1)+T(w_1))) \cdot (1/(V - T(w_1)))$.

Unseen bigram (w_1, w_2)	$c(w_1)$	$T(w_1)$	P_{Lap}	$\log_{10} P_{Lap}$	P_{WB}	$\log_{10} P_{WB}$
(white, queen)	30	5	0.0003709199	-3.4307	0.0000536855	-4.2701
(cheshire, and)	7	3	0.0003741115	-3.4270	0.0001126549	-3.9482
(mad, hatter)	15	9	0.0003729952	-3.4283	0.0001411366	-3.8504