Advanced Human Language Technologies
Exercises on Language Models and Estimation

Estimation and Smoothing

Exercise 1.

1. The probability $P(x)$ of an event x smoothed using Absolute Discount (AD) is:

$$P_{AD}(X) = \begin{cases} \frac{count(X)-\delta}{N} & if\ count(X) > 0 \\ \\ \frac{(B-N_0)\delta/N_0}{N} & otherwise \end{cases}$$

   Derive an Absolute Discount smoothing formula for the conditional probability of a trigram $P(z|xy)$, such that when no counts of order n are available, conditional probability of $(n-1)$-gram is recursively used, also with AD smoothing.

2. The probability $P(x)$ of an event x smoothed using Linear Discount (LD) is:

$$P_{LD}(X) = \begin{cases} \frac{(1-\alpha)count(X)}{N} & if\ count(X) > 0 \\ \\ \frac{\alpha}{N_0} & otherwise \end{cases}$$

   Derive a Linear Discount smoothing formula for the conditional probability of a trigram $P(z|xy)$, such that when no counts of order n are available, conditional probability of $(n-1)$-gram is recursively used, also with LD smoothing.

SOLUTION

1. Back-off using absolute discount:

$$P_{AD}(z|xy) = \begin{cases} \frac{count(xyz)-\delta}{count(xy)} & if\ count(xyz) > 0 \\ \\ \frac{(B-N_0)\delta}{count(xy)} P_{AD}(z|y) & otherwise \end{cases}$$

$$P_{AD}(z|y) = \begin{cases} \frac{count(yz)-\delta}{count(y)} & if\ count(yz) > 0 \\ \\ \frac{(B-N_0')\delta}{count(y)} P_{AD}(z) & otherwise \end{cases}$$

$$P_{AD}(z) = \begin{cases} \frac{count(z)-\delta}{N} & if\ count(z) > 0 \\ \\ \frac{(B-N_0'')\delta/N_0''}{N} & otherwise \end{cases}$$

   At each recursion level, we substract δ counts from each of the observed $(B-N_0)$ outcomes, freeing a total probability mass of $(B-N_0)\delta/count(a)$ (being a the conditioning context at each level). Thus, the lower-level distribution is multiplied by this value, shrinking it to sum up this probability mass. Note that the value of N_0 is not the same at each level, since the unobserved events vary depending on the considered conditioning context.

2. Back-off using linear discount:

$$P_{LD}(z|xy) = \begin{cases} (1-\alpha)\frac{count(xyz)}{count(xy)} & if\ count(xyz) > 0 \\ \\ \alpha P_{LD}(z|y) & otherwise \end{cases}$$

$$P_{LD}(z|y) = \begin{cases} (1-\alpha)\frac{count(yz)}{count(y)} & if\ count(yz) > 0 \\ \\ \alpha P_{LD}(z) & otherwise \end{cases}$$

$$P_{LD}(z) = \begin{cases} (1-\alpha)\frac{count(z)}{N} & if\ count(z) > 0 \\ \\ \frac{\alpha}{N_0} & otherwise \end{cases}$$

where α is the linear discount parameter, that in this case is straigthforwardly used to shrink the lower level distribution. N is the total number of observed events (i.e. length of the training sequence), and N_0 is the number of different potentially observable values that did not occurr in training data.

Exercise 2.

Certain named entity recognition system models each word in the input text as one symbol in the following alphabet Σ:

`A` Uppercase word (e.g. *IBM*)

`C` Capitalized word (e.g. *John*)

`f` Functional word (e.g. *the, and, a, an, in, of, by, ...*)

`a` Lowercase word (e.g. *will*)

`9` Number or code (e.g. *12*)

`p` Punctuation (e.g. *, . : ;*)

For instance, the sentence:
Tomorrow, John will be 12 years old. He likes music by Adam and the Ants.
would be encoded as
`C p C a a 9 a a p C a a f C f f C p`

The task is the following:

1. Use the given sample to estimate via MLE the probabilities $P(xy)$ and $P(xyz)$ for each observed bigram/trigram.
2. Compute the smoothing of the obtained probabilities using Laplace's Law. Explain the values selected for N and B. Give also the probability for unseen events.

$$P_{LAP}(x_1 \dots x_n) = \frac{count(x_1 \dots x_n) + 1}{N + B}$$

3. Compute the language model $P(z|xy) \quad \forall x, y, z \in \Sigma$ that would result from using *each* of the two previous estimations. Compare the results, discussing which option is more suitable to model these sequences.

SOLUTION

1) and 2): Compute P_{MLE} and P_{LAP} for each bigram and trigram in the training data:

For bigrams we use $N = 17$ and $B = 6 * 6 = 36$, thus $N + B = 53$

xy	$P_{MLE}(xy)$	$P_{LAP}(xy)$
aa	3/17	4/53
ap	1/17	2/53
9a	1/17	2/53
Ca	2/17	3/53
af	1/17	2/53
Cf	1/17	2/53
pC	2/17	3/53
fC	2/17	3/53
ff	1/17	2/53
a9	1/17	2/53
Cp	2/17	3/53
UNOBS	0	1/53

For trigrams, we use $N = 16$ and $B = 6 * 6 * 6 = 216$, thus $N + B = 232$

xyz	$P_{MLE}(xyz)$	$P_{LAP}(xyz)$
fCp	1/16	2/232
aaf	1/16	2/232
Cff	1/16	2/232
9aa	1/16	2/232
CpC	1/16	2/232
a9a	1/16	2/232
afC	1/16	2/232
aap	1/16	2/232
fCf	1/16	2/232
apC	1/16	2/232
Caa	2/16	3/232
pCa	2/16	3/232
aa9	1/16	2/232
ffC	1/16	2/232
UNOBS	0	1/232

3): Compute $P_{MLE}/(z|xy)$ and $P_{LAP}/(z|xy)$

In the Table below, we compute $P_{MLE}(z|xy) = count(xyz)/count(xy)$.

When applying Laplace's Law to a conditional proability, we just need to adapt our N and B. In this case, N is the number of observations we are considering (i.e. $count(xy)$) and B is the number of possible outcomes (which in this case is the number of possible different values for z, i.e. 6). Thus, the smoothed probability is computed as $P_{LAP}(z|xy) = (count(xyz) + 1)/(count(xy) + B)$

xyz	$P_{MLE}/(z\|xy)$	$P_{LAP}/(z\|xy)$
fCp	1/2	(1+1)/(2+6)
fCf	1/2	(1+1)/(2+6)
fC*	0	1/(2+6)
aaf	1/3	(1+1)/(3+6)
aap	1/3	(1+1)/(3+6)
aa9	1/3	(1+1)/(3+6)
aa*	0	1/(3+6)
Cff	1/1	(1+1)/(1+6)
Cf*	0	1/(1+6)
9aa	1/1	(1+1)/(1+6)
9a*	0	1/(1+6)
CpC	1/2	(1+1)/(2+6)
Cp*	0	1/(2+6)
a9a	1/1	(1+1)/(1+6)
a9*	0	1/(1+6)
afC	1/1	(1+1)/(1+6)
af*	0	1/(1+6)
apC	1/1	(1+1)/(1+6)
ap*	0	1/(1+6)
ffC	1/1	(1+1)/(1+6)
ff*	0	1/(1+6)
Caa	2/2	(2+1)/(2+6)
Ca*	0	1/(2+6)
pCa	2/2	(2+1)/(2+6)
pC*	0	1/(2+6)

(Asteriscs stand for any character unobserved after the first two)

Discussion: Using MLE we get probabilty zero for any sequence containing an unobserved trigram. Using Laplace we avoid this problem. Nevertheless, for this particular data, we can see that unobserved n-grams get values that are the same order of magnitude than seen events, indicating that we are deviating too much mass to unseen events.

Exercise 3.

Archeologists have found traces of ancient writings belonging to two different lost civilizations, the Thelmoth and the Uthlanga. Each civilization had its own language, different from the other, but both used the same runic alphabet (For simplicity, the runic alphabet has been mapped to uppercase latin letters here).

Archeologists need our help in deciding which of both cultures wrote the fragment:

`KFYP`

Some statistics have been gathered about each of the two languages, and we have the n-gram counts listed in the following table. N is the total number of observed unigram occurrences, and N_0 is the number of unobserved alphabet symbols. Any unlisted n-gram is assumed to have 0 observations. The symbol `$` is a placeholder for any phantom character needed at the beggining of the sequence.

Thelmoth		Uthlanga	
x	count(x)	x	count(x)
`$$K`	3	`$K`	5
`$$`	36	`$`	40
`KF`	70	`F`	93
`K`	110	`FY`	21
`KFY`	11	`FYP`	2
N	1092	N	1933
N_0	9	N_0	7

Assignment: Compute the probability of the mistery sequence for each language, and decide to which language it is most likely to belong.
Use the data in the table above, formulas given below, and $\alpha = 0.1$.
Develop your computations, providing just a single numeric result will not be accepted.

Hints:

1. The probability of a sequence $x_1, \ldots, x_n$ according to a trigram model is computed as:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P_{LD}(x_i \mid x_{i-2}\, x_{i-1})$$

   (for simpliciy, we assume that x_0 and x_{-1} are defined as an arbitrary phantom element `$`)

2. The trigram transition probability $P_{LD}(z \mid xy)$ smoothed using Linear Discount (LD) can be computed with the following recursion:

$$P_{LD}(z \mid xy) = \begin{cases} (1-\alpha)\frac{count(xyz)}{count(xy)} & if\ count(xyz) > 0 \\ \alpha P_{LD}(z \mid y) & otherwise \end{cases}$$

$$P_{LD}(z \mid y) = \begin{cases} (1-\alpha)\frac{count(yz)}{count(y)} & if\ count(yz) > 0 \\ \alpha P_{LD}(z) & otherwise \end{cases}$$

$$P_{LD}(z) = \begin{cases} (1-\alpha)\frac{count(z)}{N} & if\ count(z) > 0 \\ \alpha \frac{1}{N_0} & otherwise \end{cases}$$

SOLUTION

We need to compute the probability of the sequence `KFYP` for each language, and see which one is higher. The probabilty is

$$P(\texttt{KFYP}) = P_{LD}(\texttt{K}|\texttt{\$\$}) \cdot P_{LD}(\texttt{F}|\texttt{\$K}) \cdot P_{LD}(\texttt{Y}|\texttt{KF}) \cdot P_{LD}(\texttt{P}|\texttt{FY})$$

We compute each smoothed ngram probability using counts from **Thelmoth** table:

- $P_{LD}(\texttt{K}|\texttt{\$\$}) = (1-\alpha)\dfrac{count(\texttt{\$\$K})}{count(\texttt{\$K})} = (1-\alpha)\dfrac{3}{36} = 0.075$
- $P_{LD}(\texttt{F}|\texttt{\$K}) = \alpha P_{LD}(\texttt{F}|\texttt{K})$ (because $count(\texttt{\$KF}) = 0$)
  $= \alpha(1-\alpha)\dfrac{count(\texttt{KF})}{count(\texttt{K})} = \alpha(1-\alpha)\dfrac{70}{110} = 0.057272727$

- $P_{LD}(\texttt{Y}|\texttt{KF}) = (1-\alpha)\dfrac{count(\texttt{KFY})}{count(\texttt{KF})} = (1-\alpha)\dfrac{11}{70} = 0.141428571$
- $P_{LD}(\texttt{P}|\texttt{FY}) = \alpha P_{LD}(\texttt{P}|\texttt{Y})$ (because $count(\texttt{FYP}) = 0$)
  $= \alpha\alpha P_{LD}(\texttt{P})$ (because $count(\texttt{YP}) = 0$)
  $= \alpha\alpha\alpha\dfrac{1}{N_0}$ (because $count(\texttt{P}) = 0$)
  $= \alpha\alpha\alpha\dfrac{1}{9} = 0.000111111$

Thus, the probability of the sequence according to Thelmoth model is:

$$P(\texttt{KFYP}) = 0.075 \cdot 0.057272727 \cdot 0.141428571 \cdot 0.000111111 = 0.67 \cdot 10^{-7}$$

We compute each smoothed ngram probability using counts from **Uthlanga** table:

- $P_{LD}(\texttt{K}|\texttt{\$\$}) = \alpha P_{LD}(\texttt{K}|\texttt{\$})$ (because $count(\texttt{\$\$K} = 0)$
  $= \alpha(1-\alpha)\dfrac{count(\texttt{\$K})}{count(\texttt{\$})} = \alpha(1-\alpha)\dfrac{5}{40} = 0.01125$
- $P_{LD}(\texttt{F}|\texttt{\$K}) = \alpha P_{LD}(\texttt{F}|\texttt{K})$ (because $count(\texttt{\$KF}) = 0$)
  $= \alpha\alpha P_{LD}(\texttt{F})$ (because $count(\texttt{KF}) = 0$)
  $= \alpha\alpha(1-\alpha)\dfrac{count(\texttt{F})}{N} = \alpha\alpha(1-\alpha)\dfrac{93}{1933} = 0.000433006$
- $P_{LD}(\texttt{Y}|\texttt{KF}) = \alpha P_{LD}(\texttt{Y}|\texttt{F})$ (because $count(\texttt{KFY}) = 0$)
  $= \alpha(1-\alpha)\dfrac{count(\texttt{FY})}{count(\texttt{F})} = \alpha(1-\alpha)\dfrac{21}{93} = 0.020322581$
- $P_{LD}(\texttt{P}|\texttt{FY}) = (1-\alpha)\dfrac{count(\texttt{FYP})}{count(\texttt{FY})} = (1-\alpha) * \dfrac{2}{21} = 0.085714286$

Thus, the probability of the sequence according to Uthlanga model is:

$$P(\texttt{KFYP}) = 0.01125 \cdot 0.000433006 \cdot 0.020322581 \cdot 0.085714286 = 0.8 \cdot 10^{-8}$$

Since the probability of the sequence for Thelmoth is an order of magnitude higher than for Uthlanga, we can report to the archeologists that the fragment corresponds most likely to the former.

Exercise 4.

Simplified Polinesian is an imaginary language with a reduced alphabet of six letters (`a,i,u,p,t,k`) plus the whitespace. We have the following text sample:

```
pituka tuika tatuk ku pika tuki pikata katuki tukapa
```

The task is the following:

1. Use the given sample to estimate via MLE the probability $P(xyz)$ of each observed character trigram.
2. Compute the smoothing of the obtained probabilities using Laplace's Law. Give also the probability for unseen events.
$$P_{LAP}(w_1 \dots w_n) = \frac{\text{count}(w_1 \dots w_n) + 1}{N + B}$$
3. Compute the smoothing of the obtained probabilities using Linear Discounting and $\alpha = 0.05$. Give also the probability for unseen events.
$$P_{LD}(w_1 \dots w_n) = \begin{cases} \frac{(1-\alpha)\text{count}(w_1 \dots w_n)}{N} & \text{if count}(w_1 \dots w_n) > 0 \\ \frac{\alpha}{N_0} & \text{otherwise} \end{cases}$$
4. Discuss which of both options is more appropriate for this problem and why.

SOLUTION

1. The sample text has 52 characters, thus we can extract $N = 50$ trigam occurences. MLE probability will be obtained by simple division of $count(xyz)/50$.
2. To use Laplace's Law we need to know the values of B and N. N is the number of observed trigrams $N = 50$. B is the number of potentially observable trigrams. Since our Simplified Polinesian alphabet has seven characters (6 letters plus the whitespace), the potential number of observable trigrams is 7^3 thus $B = 7^3 = 343$.
3. For linear discount smoothing, we need to know the number of observations ($N = 50$). We also need the number of unobserved potential trigrams N_0. Since $B = 343$ and the sample text contains 31 different trigrams, the number of unobserved possible trigrams is $N_0 = 343 - 31 = 312$.

The following table contains the results of computing the requested distributions:

xyz	$P_{MLE}(xyz)$	$P_{LAP}(xyz)$	$P_{LD}(xyz)$
a_k	$1/50 = 0.02$	$(1+1)/(50+343) = 0.0051$	$0.95 \times 1/50 = 0.0190$
apa	$1/50 = 0.02$	$(1+1)/(50+343) = 0.0051$	$0.95 \times 1/50 = 0.0190$
a_t	$3/50 = 0.06$	$(3+1)/(50+343) = 0.0102$	$0.95 \times 3/50 = 0.0570$
ata	$1/50 = 0.02$	$(1+1)/(50+343) = 0.0051$	$0.95 \times 1/50 = 0.0190$
atu	$2/50 = 0.04$	$(2+1)/(50+343) = 0.0076$	$0.95 \times 2/50 = 0.0380$
ika	$3/50 = 0.06$	$(3+1)/(50+343) = 0.0102$	$0.95 \times 3/50 = 0.0570$
i_p	$1/50 = 0.02$	$(1+1)/(50+343) = 0.0051$	$0.95 \times 1/50 = 0.0190$
i_t	$1/50 = 0.02$	$(1+1)/(50+343) = 0.0051$	$0.95 \times 1/50 = 0.0190$
itu	$1/50 = 0.02$	$(1+1)/(50+343) = 0.0051$	$0.95 \times 1/50 = 0.0190$
_ka	$1/50 = 0.02$	$(1+1)/(50+343) = 0.0051$	$0.95 \times 1/50 = 0.0190$
ka_	$3/50 = 0.06$	$(3+1)/(50+343) = 0.0102$	$0.95 \times 3/50 = 0.0570$
kap	$1/50 = 0.02$	$(1+1)/(50+343) = 0.0051$	$0.95 \times 1/50 = 0.0190$
kat	$2/50 = 0.04$	$(2+1)/(50+343) = 0.0076$	$0.95 \times 2/50 = 0.0380$
ki_	$2/50 = 0.04$	$(2+1)/(50+343) = 0.0076$	$0.95 \times 2/50 = 0.0380$
k_k	$1/50 = 0.02$	$(1+1)/(50+343) = 0.0051$	$0.95 \times 1/50 = 0.0190$
_ku	$1/50 = 0.02$	$(1+1)/(50+343) = 0.0051$	$0.95 \times 1/50 = 0.0190$
ku_	$1/50 = 0.02$	$(1+1)/(50+343) = 0.0051$	$0.95 \times 1/50 = 0.0190$
_pi	$2/50 = 0.04$	$(2+1)/(50+343) = 0.0076$	$0.95 \times 2/50 = 0.0380$
pik	$2/50 = 0.04$	$(2+1)/(50+343) = 0.0076$	$0.95 \times 2/50 = 0.0380$
pit	$1/50 = 0.02$	$(1+1)/(50+343) = 0.0051$	$0.95 \times 1/50 = 0.0190$
_ta	$1/50 = 0.02$	$(1+1)/(50+343) = 0.0051$	$0.95 \times 1/50 = 0.0190$
ta_	$1/50 = 0.02$	$(1+1)/(50+343) = 0.0051$	$0.95 \times 1/50 = 0.0190$
tat	$1/50 = 0.02$	$(1+1)/(50+343) = 0.0051$	$0.95 \times 1/50 = 0.0190$
_tu	$3/50 = 0.06$	$(3+1)/(50+343) = 0.0102$	$0.95 \times 3/50 = 0.0570$
tui	$1/50 = 0.02$	$(1+1)/(50+343) = 0.0051$	$0.95 \times 1/50 = 0.0190$
tuk	$5/50 = 0.10$	$(5+1)/(50+343) = 0.0153$	$0.95 \times 5/50 = 0.0950$
uik	$1/50 = 0.02$	$(1+1)/(50+343) = 0.0051$	$0.95 \times 1/50 = 0.0190$
uk_	$1/50 = 0.02$	$(1+1)/(50+343) = 0.0051$	$0.95 \times 1/50 = 0.0190$
uka	$2/50 = 0.04$	$(2+1)/(50+343) = 0.0076$	$0.95 \times 2/50 = 0.0380$
uki	$2/50 = 0.04$	$(2+1)/(50+343) = 0.0076$	$0.95 \times 2/50 = 0.0380$
u_p	$1/50 = 0.02$	$(1+1)/(50+343) = 0.0051$	$0.95 \times 1/50 = 0.0190$
unseen	**0**	**(0+1)/(50+343) = 0.0025**	**0.05/312 = 0.000016**

4. Using Laplace's Law we obtain a probability for unseen events that is about half the probability of events observed once. Since there are 312 unobserved possibilities, the total probability mass devoted to unobserved trigrams is $0.0025 \times 312 = 0.78$. That is, much larger than the total mass for observed events, which does not make sense.

   On the other hand, Linear Discount reassigns only 5% of the total mass to unseen events, thus obtaining probabilities for unseen events several orders of magnitude smaller than the probability for events observed once. This option is better, since it is closer to MLE, but overcomes the problem of zero probabilities for unseen trigrams.

Exercise 5.

We want to build a model of musical language. We will consider musical sequences formed by tones belonging to the set:

$$V = \{\texttt{do}, \texttt{do\#}, \texttt{re}, \texttt{re\#}, \texttt{mi}, \texttt{fa}, \texttt{fa\#}, \texttt{sol}, \texttt{sol\#}, \texttt{la}, \texttt{la\#}, \texttt{si}\}$$

The following sequence corresponds to Beethoven's 9th symphony theme *Ode to Joy*:

```
sol# sol# la si si la sol# fa# mi mi fa# sol# sol# fa# fa# fa#
 fa# sol# mi fa# sol# la# sol# mi fa# sol# la sol# fa# mi fa#
```

The task is the following:

1. Use the given sample to estimate via MLE the probabilities $P(xy)$ and $P(xyz)$ for each observed bigram/trigram.
2. Compute the smoothing of the obtained probabilities using Lidstones's Law with $\lambda = 0.1$. Justify the values selected for N and B. Give also the probability for unseen events.

$$P_{LID}(x_1 \dots x_n) = \frac{count(x_1 \dots x_n) + \lambda}{N + B\lambda}$$

3. Compute the parameters $P(z|xy) \quad \forall x, y, z \in V$ of the trigram language model that would result from using each of the two previous estimation methods.

SOLUTION

1. The input sequence has 31 tokens, which produces 30 bigram occurrences and 29 trigram occurences. MLE probability will be computed by simple dount division.
2. To smooth using Lidstone's Law, we need the values of N and B. For bigrams $N = 30$ and for trigrams $N = 29$. There are 12 possible values (tones) for each token, thus the number of potentially observable combinations is $B = 12^2 = 144$ for bigrams and $B = 12^3 = 1728$ for trigrams.

The following tables contain the results of computing the requested distributions:

For bigrams we use $N = 30$ and $B = 12^2 = 144$, thus $N + B\lambda = 30 + 144 \times 0.1 = 44.4$

xy	$P_{MLE}(xy)$	$P_{LID}(xy)$
fa# fa#	$3/30 = 0.10$	$(3 + 0.1)/(30 + 144 \times 0.1) = 0.070$
fa# mi	$2/30 = 0.07$	$(2 + 0.1)/(30 + 144 \times 0.1) = 0.047$
fa# sol#	$4/30 = 0.13$	$(4 + 0.1)/(30 + 144 \times 0.1) = 0.092$
la si	$1/30 = 0.03$	$(1 + 0.1)/(30 + 144 \times 0.1) = 0.025$
la sol#	$2/30 = 0.07$	$(2 + 0.1)/(30 + 144 \times 0.1) = 0.047$
la# sol#	$1/30 = 0.03$	$(1 + 0.1)/(30 + 144 \times 0.1) = 0.025$
mi fa#	$4/30 = 0.13$	$(4 + 0.1)/(30 + 144 \times 0.1) = 0.092$
mi mi	$1/30 = 0.03$	$(1 + 0.1)/(30 + 144 \times 0.1) = 0.025$
si la	$1/30 = 0.03$	$(1 + 0.1)/(30 + 144 \times 0.1) = 0.025$
si si	$1/30 = 0.03$	$(1 + 0.1)/(30 + 144 \times 0.1) = 0.025$
sol# fa#	$3/30 = 0.10$	$(3 + 0.1)/(30 + 144 \times 0.1) = 0.070$
sol# la	$2/30 = 0.07$	$(2 + 0.1)/(30 + 144 \times 0.1) = 0.047$
sol# la#	$1/30 = 0.03$	$(1 + 0.1)/(30 + 144 \times 0.1) = 0.025$
sol# mi	$2/30 = 0.07$	$(2 + 0.1)/(30 + 144 \times 0.1) = 0.047$
sol# sol#	$2/30 = 0.07$	$(2 + 0.1)/(30 + 144 \times 0.1) = 0.047$
unseen	**0**	**(0+0.1)/(30+144×0.1) = 0.0023**

For trigrams we use $N = 29$ and $B = 12^3 = 1728$, thus $N + B\lambda = 29 + 1728 \times 0.1 = 201.8$

xyz	$P_{MLE}(xyz)$	$P_{LID}(xyz)$
fa# fa# fa#	$2/29 = 0.069$	$(2+0.1)/(29+1728\times 0.1) = 0.0104$
fa# fa# sol#	$1/29 = 0.034$	$(1+0.1)/(29+1728\times 0.1) = 0.0055$
fa# mi fa#	$1/29 = 0.034$	$(1+0.1)/(29+1728\times 0.1) = 0.0055$
fa# mi mi	$1/29 = 0.034$	$(1+0.1)/(29+1728\times 0.1) = 0.0055$
fa# sol# la	$1/29 = 0.034$	$(1+0.1)/(29+1728\times 0.1) = 0.0055$
fa# sol# la#	$1/29 = 0.034$	$(1+0.1)/(29+1728\times 0.1) = 0.0055$
fa# sol# mi	$1/29 = 0.034$	$(1+0.1)/(29+1728\times 0.1) = 0.0055$
fa# sol# sol#	$1/29 = 0.034$	$(1+0.1)/(29+1728\times 0.1) = 0.0055$
la si si	$1/29 = 0.034$	$(1+0.1)/(29+1728\times 0.1) = 0.0055$
la sol# fa#	$2/29 = 0.069$	$(2+0.1)/(29+1728\times 0.1) = 0.0104$
la# sol# mi	$1/29 = 0.034$	$(1+0.1)/(29+1728\times 0.1) = 0.0055$
mi fa# sol#	$3/29 = 0.103$	$(3+0.1)/(29+1728\times 0.1) = 0.0154$
mi mi fa#	$1/29 = 0.034$	$(1+0.1)/(29+1728\times 0.1) = 0.0055$
si la sol#	$1/29 = 0.034$	$(1+0.1)/(29+1728\times 0.1) = 0.0055$
si si la	$1/29 = 0.034$	$(1+0.1)/(29+1728\times 0.1) = 0.0055$
sol# fa# fa#	$1/29 = 0.034$	$(1+0.1)/(29+1728\times 0.1) = 0.0055$
sol# fa# mi	$2/29 = 0.069$	$(2+0.1)/(29+1728\times 0.1) = 0.0104$
sol# la si	$1/29 = 0.034$	$(1+0.1)/(29+1728\times 0.1) = 0.0055$
sol# la sol#	$1/29 = 0.034$	$(1+0.1)/(29+1728\times 0.1) = 0.0055$
sol# la# sol#	$1/29 = 0.034$	$(1+0.1)/(29+1728\times 0.1) = 0.0055$
sol# mi fa#	$2/29 = 0.069$	$(2+0.1)/(29+1728\times 0.1) = 0.0104$
sol# sol# fa#	$1/29 = 0.034$	$(1+0.1)/(29+1728\times 0.1) = 0.0055$
sol# sol# la	$1/29 = 0.034$	$(1+0.1)/(29+1728\times 0.1) = 0.0055$
unseen	**0**	**(0+0.1)/(29+1728×0.1) = 0.00049**

3. MLE estimation for the trigram transition probability is $P_{MLE}(z|xy) = count(xyz)/count(xy)$

   For Lidstone Law, we need to consider that we are estimating the probability of z but restricting our observations to cases where the first two tokens were xy. Thus, the formula will be:

   $$P_{LID}(z|xy) = \frac{count(xyz) + \lambda}{count(xy) + |V|\lambda}$$

   that is, the total number of observations considered is $count(xy)$ (the same than in P_{MLE} above). The number of potentially observable events is the number of possible values for z, that is the number of different tones in V.

   Thus, we get the distributions in the following table. Note that each conditional probability is a separate distribution, so each of them has a different probability for the unseen events (in bold).

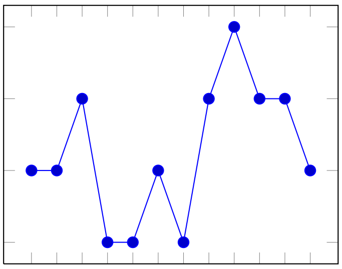
xyz	$P_{MLE}(z\mid xy)$	$P_{LID}(z\mid xy)$
fa# fa# fa#	$2/3 = 0.67$	$(2+0.1)/(3+12\times 0.1) = 0.500$
fa# fa# sol#	$1/3 = 0.33$	$(1+0.1)/(3+12\times 0.1) = 0.262$
fa# fa# others	**0/3=0**	**(0+0.1)/(3+12×0.1) = 0.024**
fa# mi fa#	$1/2 = 0.50$	$(1+0.1)/(2+12\times 0.1) = 0.344$
fa# mi mi	$1/2 = 0.50$	$(1+0.1)/(2+12\times 0.1) = 0.344$
fa# mi others	**0/2=0**	**(0+0.1)/(2+12×0.1) = 0.031**
fa# sol# la	$1/4 = 0.25$	$(1+0.1)/(4+12\times 0.1) = 0.212$
fa# sol# la#	$1/4 = 0.25$	$(1+0.1)/(4+12\times 0.1) = 0.212$
fa# sol# mi	$1/4 = 0.25$	$(1+0.1)/(4+12\times 0.1) = 0.212$
fa# sol# sol#	$1/4 = 0.25$	$(1+0.1)/(4+12\times 0.1) = 0.212$
fa# sol# others	**0/4=0**	**(0+0.1)/(4+12×0.1) = 0.019**
la si si	$1/1 = 1.00$	$(1+0.1)/(1+12\times 0.1) = 0.500$
la si others	**0/1=0**	**(0+0.1)/(1+12×0.1) = 0.045**
la sol# fa#	$2/2 = 1.00$	$(2+0.1)/(2+12\times 0.1) = 0.656$
la sol# others	**0/2=0**	**(0+0.1)/(2+12×0.1) = 0.031**
la# sol# mi	$1/1 = 1.00$	$(1+0.1)/(1+12\times 0.1) = 0.500$
la# sol# others	**0/1=0**	**(0+0.1)/(1+12×0.1) = 0.045**
mi fa# sol#	$3/4 = 0.75$	$(3+0.1)/(4+12\times 0.1) = 0.596$
mi fa# others	**0/4=0**	**(0+0.1)/(4+12×0.1) = 0.019**
mi mi fa#	$1/1 = 1.00$	$(1+0.1)/(1+12\times 0.1) = 0.500$
mi mi others	**0/1=0**	**(0+0.1)/(1+12×0.1) = 0.045**
si la sol#	$1/1 = 1.00$	$(1+0.1)/(1+12\times 0.1) = 0.500$
si la others	**0/1=0**	**(0+0.1)/(1+12×0.1) = 0.045**
si si la	$1/1 = 1.00$	$(1+0.1)/(1+12\times 0.1) = 0.500$
si si others	**0/1=0**	**(0+0.1)/(1+12×0.1) = 0.045**
sol# fa# fa#	$1/3 = 0.33$	$(1+0.1)/(3+12\times 0.1) = 0.262$
sol# fa# mi	$2/3 = 0.67$	$(2+0.1)/(3+12\times 0.1) = 0.500$
sol# fa# others	**0/3=0**	**(0+0.1)/(3+12×0.1) = 0.024**
sol# la si	$1/2 = 0.50$	$(1+0.1)/(2+12\times 0.1) = 0.344$
sol# la sol#	$1/2 = 0.50$	$(1+0.1)/(2+12\times 0.1) = 0.344$
sol# la others	**0/2=0**	**(0+0.1)/(2+12×0.1) = 0.031**
sol# la# sol#	$1/1 = 1.00$	$(1+0.1)/(1+12\times 0.1) = 0.500$
sol# la# others	**0/1=0**	**(0+0.1)/(1+12×0.1) = 0.045**
sol# mi fa#	$2/2 = 1.00$	$(2+0.1)/(2+12\times 0.1) = 0.656$
sol# mi others	**0/2=0**	**(0+0.1)/(2+12×0.1) = 0.031**
sol# sol# fa#	$1/2 = 0.50$	$(1+0.1)/(2+12\times 0.1) = 0.344$
sol# sol# la	$1/2 = 0.50$	$(1+0.1)/(2+12\times 0.1) = 0.344$
sol# sol# others	**0/2=0**	**(0+0.1)/(2+12×0.1) = 0.031**

Exercise 6.

We want to devise a method to predict stock exchange evolution using n-gram models.
For this, we model the stock value daily behaviour as a sequence of *movements*. Possible movements are in the set $M = \{=, +, -, \wedge, \vee\}$, where:

$=$	no change
$+$	small value increment
$-$	small value decrement
$\wedge$	big value increment
$\vee$	big value decrement

For instance, the sequence $= + \vee = + - \wedge + - = -$ encodes the following stock share behaviour:



We have the following historic data that we want to use as a training set to build a trigram model:

$$= + == ++ = - \wedge - \wedge + == - = + = \vee \vee + =$$

1. Compute the following probability values corresponding to a MLE trigram model, and those corresponding to a MLE model smoothed with Lidstone Law using $\lambda = 0.1$.

   - $P(+ =)$
   - $P(+ = \vee)$
   - $P(+ = +)$
   - $P(+ \mid + =)$
   - $P(\wedge \mid = -)$

   Justify the values chosen for B and N in each case. You can leave the probability value as a fraction. Do not provide just a final numeric result.

2. What is the most likely continuation of the sequence $\vee - = \vee == +$ according to each model? Justify your answer.

SOLUTION

1.

	MLE prob.	Smoothed prob.	Justification
$P(+ =)$	5/21	(5+0.1)/(21+25*0.1)	$N = 21$ observed bigrams; $B = \lvert M\rvert^2 = 5^2 = 25$ possible bigrams
$P(+ = \vee)$	1/20	(1+0.1)/(20+125*0.1)	$N = 20$ observed trigrams $B = \lvert M\rvert^3 = 5^3 = 125$ possible trigrams
$P(+ = +)$	0/20	(0+0.1)/(20+125*0.1)	$N = 20$ observed trigrams $B = \lvert M\rvert^3 = 5^3 = 125$ possible trigrams
$P(+ \mid + =)$	0/5	(0+0.1)/(5+5*0.1)	$N = 5$ occurrences of '$+ =$' $B = \lvert M\rvert = 5$ possible values after '$+ =$'
$P(\wedge \mid = -)$	1/2	(1+0.1)/(2+5*0.1)	$N = 2$ occurrences of '$= -$' $B = \lvert M\rvert = 5$ possible values after '$= -$'

2. Since it is a trigram model, the probability of the possible continuations of the given sequence is determined by its two last elements, i.e. $= +$.

   The continuation probabilities after $= +$ are:

	MLE prob.	Smoothed prob.
$P(=\|= +)$	2/3	(2+0.1)/(3+5*0.1)
$P(+\|= +)$	1/3	(1+0.1)/(3+5*0.1)
$P(-\|= +)$	0/3	(0+0.1)/(3+5*0.1)
$P(\wedge\|= +)$	0/3	(0+0.1)/(3+5*0.1)
$P(\vee\|= +)$	0/3	(0+0.1)/(3+5*0.1)

   Thus, the most likely continuation is '=' in both models.

Exercise 7.

We want to build a probabilistic model for sentiment analysis on tweets, and for that we have gathered the following training data:

- We have a collection of 1,000 tweets with gold standard annotations. 300 tweets are annotated as *positive* (POS), 250 as *negative* (NEG), and the remaining 450 are annotated as *neutral* (NEU).
- We collected some statistics about words appearing in these tweets:
  - 140 tweets contain the word *big*. 70 of them are annotated as NEU, 60 as POS, and 10 as NEG.
  - 90 tweets contain the word *great*. 50 of them are annotated as POS.
  - 100 tweets contain the word *awful*. 80 of them are annotated as NEG.
  - 40 tweets contain the word *terrific*. All of them are annotated as POS.

1. Compute the following probability values corresponding to a MLE model, and those smoothed using linear discount with $\alpha = 0.1$. Justify your answer and the values chosen for N and N_0 in each case where it applies.
   - Probability that a tweet is positive, $P(\text{POS})$
   - Probability that a tweet contains word *big*, $P(big)$
   - Probability that a tweet is positive and contains word *big*, $P(\text{POS} \wedge big)$
   - Probability that a negative tweet contains word *awful*, $P(awful|\text{NEG})$
2. Compute the following probability values corresponding to a MLE model, those corresponding to a model smoothed with Laplace's Law, and those smoothed using linear discount with $\alpha = 0.1$. Justify your answer and the values chosen for B, N, and N_0 in each case where it applies.
   - Probability that a tweet containing word *great* is positive, $P(\text{POS}|great)$
   - Probability that a tweet containing word *terrific* is positive, $P(\text{POS}|terrific)$
   - Probability that a tweet containing word *terrific* is negative, $P(\text{NEG}|terrific)$

   Develop your computations, do not provide just a numeric result. You may leave probabilities as fractions.

SOLUTION

1.

	P_{MLE}	P_{LD}	justification
$P(\text{POS})$	$\frac{\#\text{POS}}{N} = \frac{300}{1,000}$	$0.9\frac{300}{1,000}$	We have 1,000 tweets, i.e. 1,000 cases where a label may be observed, thus $N = 1,000$.
$P(big)$	$\frac{\#big}{N} = \frac{140}{1,000}$	$0.9\frac{140}{1,000}$	We have 1,000 tweets, i.e. 1,000 cases where a word may be observed, thus $N = 1,000$.
$P(\text{POS} \wedge big)$	$\frac{\#(\text{POS}\wedge big)}{N} = \frac{60}{1,000}$	$0.9\frac{60}{1,000}$	We have 1,000 tweets, i.e. 1,000 cases where a word-label co-occurrence may be observed, thus $N = 1,000$.
$P(awful\|\text{NEG})$	$\frac{\#(\text{NEG}\wedge awful)}{\#NEG} = \frac{80}{250}$	$0.9\frac{80}{250}$	We are counting word ocurrences conditioned to the tweet being NEG, so, $N = 250$.

2.

	P_{MLE}	P_{LAP}	P_{LD}	justification
$P(\text{POS}\vert great)$	$\frac{\#(\text{POS}\wedge great)}{\#great} = \frac{50}{90}$	$\frac{50+1}{90+3}$	$0.9\frac{50}{90}$	We are counting events conditioned to the occurrence of *great* in a tweet, so $N = 90$. For P_{LAP}, Possible labels are (POS, NEG, NEU), thus $B = 3$.
$P(\text{POS}\vert terrific)$	$\frac{\#(\text{POS}\wedge terrific)}{\#terrific} = \frac{40}{40}$	$\frac{40+1}{40+3}$	$0.9\frac{40}{40}$	We are counting events conditioned to the occurrence of *terrific* in a tweet, so $N = 40$. For P_{LAP}, Possible labels are (POS, NEG, NEU), thus $B = 3$.
$P(\text{NEG}\vert terrific)$	$\frac{\#(\text{NEG}\wedge terrific)}{\#terrific} = \frac{0}{40}$	$\frac{0+1}{40+3}$	$\frac{\alpha}{N_0} = \frac{0.1}{2}$	We are counting events conditioned to the occurrence of *terrific* in a tweet, so $N = 40$. For P_{LAP}, , Possible labels are (POS, NEG, NEU), thus $B = 3$. For P_{LD}, N_0 is the number of outcomes unobserved with *terrific*. There are 3 possible outcomes (POS, NEG, NEU) but only one (POS) was observed, thus $N_0 = 2$

Exercise 8.

Worldwide online retail seller Papazom.com asked us to design a model that classifies products into a set of categories, given the product description. The target categories are: *Electronics* (ELEC), *Computers* (COMP), *Fashion* (FASH), and *Tools* (TOOL).

We have a sample of 1,500 annotated product descriptions. 200 are classified as ELEC, 350 as COMP, 650 as FASH, and 300 as TOOL.

- 70 products contain the word *display*. 40 are annotated as ELEC, 20 as COMP, and 10 as FASH.
- 150 products contain the word *Gigabytes*. 90 are annotated as ELEC, and 60 as COMP.
- 160 products contain the word *waterproof*. 80 are annotated as FASH, 30 as ELEC, and 50 as TOOL.
- 110 products contain the word *handmade*. 95 are annotated as FASH, and 15 as TOOL.

1. Compute the following probability values corresponding to a MLE model
   - Probability that a product belongs to category *Electronics*, $P_{MLE}(\text{ELEC})$
   - Probability that a product containing word *handmade* is in *Fashion*, $P_{MLE}(\text{FASH}|handmade)$
   - Probability that a product containing word *handmade* is in *Computers*, $P_{MLE}(\text{COMP}|handmade)$
   - Probability that a *Computer* product contains word *display*, $P_{MLE}(display|\text{COMP})$
   - Probability that a *Computer* product contains word *handmade*, $P_{MLE}(handmade|\text{COMP})$
2. Compute the following probability values corresponding to a model smoothed using absolute discount with $\delta = 0.3$
   - Probability that a product description contains word *waterproof*, $P_{AD}(waterproof)$
   - Probability that a product is in *Tools* and contains word *handmade*, $P_{AD}(\text{TOOL} \wedge handmade)$
   - Probability that a product containing word *display* is in *Electronics*, $P_{AD}(\text{ELEC}|display)$
   - Probability that a product containing word *display* is in *Tools*, $P_{AD}(\text{TOOL}|display)$
3. Compute the following probability values corresponding to a model smoothed with Lidstone's Law with $\lambda = 0.2$.

- Probability that a product belongs to category *Tools*, $P_{LID}(\text{TOOL})$
- Probability that a product containing word *Gigabytes* is in *electronics*, $P_{LID}(\text{ELEC}|Gigabytes)$
- Probability that a product containing word *Gigabytes* is in *Tools*, $P_{LID}(\text{TOOL}|Gigabytes)$

Justify your answer and the values chosen for B, N, and N_0 where it applies.

Develop your computations, do not provide just a numeric result. You may leave probabilities as fractions.

SOLUTION

1. Maximum Likelihood

$$\begin{aligned}
P_{MLE}(\text{ELEC}) &= \tfrac{\#\text{ELEC}}{N} = \tfrac{200}{1500} && (1)\\
P_{MLE}(\text{FASH}|handmade) &= \tfrac{\#(\text{FASH}\wedge handmade)}{\#handmade} = \tfrac{95}{110} && (2)\\
P_{MLE}(\text{COMP}|handmade) &= \tfrac{\#(\text{COMP}\wedge handmade)}{\#handmade} = \tfrac{0}{110} && (2)\\
P_{MLE}(display|\text{COMP}) &= \tfrac{\#(\text{COMP}\wedge display)}{\#COMP} = \tfrac{20}{350} && (3)\\
P_{MLE}(handmade|\text{COMP}) &= \tfrac{\#(\text{COMP}\wedge handmade)}{\#COMP} = \tfrac{0}{350} && (3)
\end{aligned}$$

2. Absolute Discount

$$\begin{aligned}
P_{AD}(waterproof) &= \tfrac{\#waterproof-\delta}{N} = \tfrac{160-0.3}{1500} && (1)\\
P_{AD}(\text{TOOL} \wedge handmade) &= \tfrac{\#(\text{TOOL}\wedge handmade)-\delta}{N} = \tfrac{15-0.3}{1500} && (1)\\
P_{AD}(\text{ELEC}|display) &= \tfrac{\#(\text{ELEC}\wedge display)-\delta}{\#display} = \tfrac{40-0.3}{70} && (4)\\
P_{AD}(\text{TOOL}|display) &= \tfrac{(B-N_0)\delta/N_0}{N} = \tfrac{(4-1)0.3/1}{70} && (4)(5)(6)
\end{aligned}$$

3. Lidstone's Law

$$\begin{aligned}
P_{LID}(\text{TOOL}) &= \tfrac{\#ELEC+\lambda}{N+B\lambda} = \tfrac{300+0.2}{1500+4\times0.2} && (1)(5)\\
P_{LID}(\text{ELEC}|Gibabytes) &= \tfrac{\#(ELEC\wedge Gigabytes)+\lambda}{N+B\lambda} = \tfrac{90+0.2}{150+4\times0.2} && (5)(7)\\
P_{LID}(\text{TOOL}|Gibabytes) &= \tfrac{\#(TOOL\wedge Gigabytes)+\lambda}{N+B\lambda} = \tfrac{0+0.2}{150+4\times0.2} && (5)(7)
\end{aligned}$$

Justification of chosen values:

(1) $N = 1500$ because we have 1500 products, i.e. 1500 cases where a word or a class may be observed.

(2) We are conditioning on the occurrece of *handmade*, which occurrs in 110 products, so that is our number of observations N.

(3) We are conditioning on the occurrece of COMP, which occurrs in 350 products, so that is our number of observations N.

(4) We are conditioning on the occurrece of *display*, which occurrs in 70 products, so that is our number of observations N.

(5) We are computing the probability of the class, so there are 4 possible outcomes, thus $B = 4$.

(6) Only 3 classes (out of 4 possible) were observed with *display*, thus there is one unobserved class: $N_0 = 1$.

(7) We are conditioning on the occurrece of *Gigabytes*, which occurrs in 150 products, so that is our number of observations N.