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.

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.

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 $x_1, \dots, x_n$ according to a trigram model is computed as:

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

   (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:

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

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

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

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.

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:

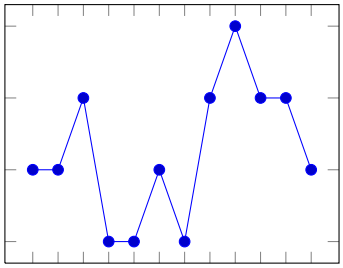
1. Use the given sample to estimate via MLE the probabilities $P(xy)$ and $P(xyz)$ for each observed bigram/trigram.
2. Compute the smoothing of the obtained probabilities using Lidstones's Law with $\lambda = 0.1$. Justify the values selected for N and B. Give also the probability for unseen events.
$$P_{LID}(x_1 \dots x_n) = \frac{count(x_1 \dots x_n) + \lambda}{N + B\lambda}$$
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.

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.

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.

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.