Problem 1
The “Little Amazon” company sells books on the internet. “Little Amazon” has the following titles for sale: 0 – “War and Peace”, 1 – “Harry Potter & the Deathly Hallows”, 2 – “Winnie the Pooh”, 3 – “Get rich NOW”, 4 – “Probability”. “Little Amazon” has collected data on the sales of each title over the last 3 months. This data is collected in the file hw2-little-amazon.dat which is available on the assignments web page.

For all the following questions, please give the “literal” expression of the answer as well as the numeric value.

a. Denote by $\theta_i$ the probability that a customer buys title $i$. Assume that each purchase of a book is independent of the other purchases by the same customer or by other customers. Estimate $\theta = (\theta_0, \ldots, \theta_4)$ from the data. What are the sufficient statistics?

b. A customer buys 3 books. What is the probability that he buys “War and Peace”, “Harry Potter”, “Probability” in this order?

c. A customer buys 4 books. What is the probability that she buys only non-fiction, that is, $N = \{3, 4\}$?

d. A customer buys 2 “Probability” books and 3 fiction (i.e 0 or 1 or 2) books. What is the probability of this event?

e. A customer buys $n$ books. What is the probability that he buys at least one “Probability”?

Problem 2
This problem is a language identification experiment (will also be demoed in class).

We assume that sentences in a language are generated by sampling letters independently from the alphabet $\{A, B, C, \ldots, Z\}$. Spaces and punctuation are ignored. For instance, the probability of the sentence “’Who’s on first?’” is

$$\theta_W \theta_H \theta_O^2 \theta_S \theta_N \theta_F \theta_I \theta_R \theta_T$$

because the sentence contains (W, H, O, S, O, N, . . ., T) in this order. The parameters $\theta_{A,Z}$ of this simple model depend on the language. The files english.dat,
french.dat, german.dat, spanish.dat are ASCII files containing the probabilities of the letters A–Z in each of the languages, multiplied by 1000. For example, below is the beginning of english.dat:

A 81.51
B 14.40
C 27.58

The data mean that for the English language \( \theta_A = 0.08151, \theta_B = 0.0144, \theta_C = 0.02758 \). These estimates are obtained by taking a long text, eliminating all the spaces and punctuation (and other non-literals like numbers), turning everything to lower case, and treating the obtained sequence as the outcome of a series of independent trials.

a. Use the above language models to decide on the language of the following sentences by the Maximum Likelihood (ML) method. The sentences are:

1. "As far as the laws of mathematics refer to reality, they are not certain, as far as they are certain, they do not refer to reality."
   –Albert Einstein
2. ‘‘Freude, schöner Götterfunken, Tochter aus Elysium.’’ (Joy, beautiful divine spark, daughter from Elysium.)
   –From Schiller’s “Ode to Joy”.
3. ‘‘Chi trova un amico, trova un tesoro.’’ (He who finds a friend finds a treasure.)
   –Italian Proverb
4. “El que busca encuentra” (Who looks for something will find it)
   –Spanish proverb
4. ‘‘From what we get, we can make a living; what we give, however, makes a life.’’
   –Arthur Ashe

For each sentence, do the following:

• Preprocess: Turn all letters to lower case, eliminate spaces and punctuation.

• Get the sufficient statistics: Count the number of times each letter appears in the sentence. These are the counts \( n_a, n_b, \ldots n_z \).

• For each language, compute the log-likelihood of the sentence in that language \( l_{E,G,S,F}(\text{sentence}) = \log_2 P_{E,G,S,F}(\text{sentence}) \). Print out these log-likelihoods. Make sure to convert into base 2 logarithms or to indicate the basis of the logarithm if it is not base 2.

• Output the best guess according to the ML method, i.e the language that gives highest likelihood to the data.

---

1 The source of this data is http://www.santacruzpl.org/readyref/files/g-l/ltfrqeng.shtml, ltfrqger.shtml, ltfrqsq.shtml, ltfrqfr.shtml.
Comment on what you observe: are the guesses correct? If not, why do you think not? How does the likelihood of the best guess depend on the length of the sentence? How does the difference in log-likelihoods between the best guess and the second best guess depend on the length of the sentence? What do you think of the probability models defined here as description of how language is produced?

Here is a short matlab code that computes the statistics of a sentence, typed all in lower case. It ignores all characters different from “a–z”.

```matlab
alphabet='abcdefghijklmnopqrstuvwxyz';
sentence = input('Type a sentence (lower case only): ');
for ii = 1:26;
    counts(ii) = length(find(sentence == alphabet(ii)));
end;
```

Note: The task that you just performed, deciding which of a given set of sources has generated an observation (in this case a sentence) is called classification or pattern recognition. Classification is very important both in Artificial Intelligence and in Statistics. We will talk more about classification later in this course.