# A Naïve Librarian

(Time Limit=1 second)

Jabez is a librarian at *Maranatha Christian University*[1]. One day his supervisor, Mr. August receives a large amount of free books in one package from *Periplus Online Bookstore*[2]. Once the package is opened, there are two collections of books: the one with labels (**the labeled collection**) and the other without labels (**the unlabeled collection**). A book in each collection is accompanied with a sentence written in a post-it note. Apparently, the sentence is a summary of a book. Moreover, **each book in the labeled collection has a label which is a category of the book**. For example, one of the books in the labeled collection, entitled *Alice in Wonderland*, has a sentence, "*A girl who is trapped in a fantasy world*" and a category, "*fiction*".

The feeling of joy upon receiving free books mixed with confusion happen to Mr. August as **he does not know how to give categories to all books in the unlabeled collection**. Therefore, he asks Jabez to solve this categorisation problem. Unbeknown to all library workers, Jabez is a truly great grandchildren of Thomas Bayes[3] (1701-1761), an English statistician, philosopher, and Presbyterian minister. Notwithstanding his unpublished works, Thomas' notes were edited and published after his death by Richard Price (McGrayne, 2011).

During his childhood, Jabez Bayes was trained austerely by his parents; therefore, Jabez has mastered the only one theorem from his great grandfather's legacy (Bayes, 1970). Mr. August has stumbled upon a fortuitous moment in which Jabez, his assistant, is an extremely talented statistician instead of just a decent librarian.

Evidently, a sentence consists of words; in this categorization problem, a word is named a *term*. For example, a sentence, "*A girl who is trapped in a fantasy world*", has nine terms as follows: *a*, *girl*, *who*, *is*, *trapped*, *in*, *a*, *fantasy*, and *world*. In order to give a category to a book, Jabez needs to compute the probability of a category given a sentence. For example, there is an unlabeled book with a summary, "*A guide to study machine learning*". If Jabez want to choose a label either *fiction* or *non-fiction* for the book, he should compute both probability of *fiction* given the sentence and probability of *non-fiction* given the sentence. After computing both probabilities, Jabez shall assign the label with the largest probability to the book. Again, for example, Jabez give *non-fiction* to the book with a summary, "*A guide to study machine learning*".

Before embarking computing the probabilities, Jabez states his great grandfather's theorem that is: the probability of event $A$ happens, given that event $B$

[1] http://library.maranatha.edu

[2] https://www.periplus.com

[3] https://en.wikipedia.org/wiki/Thomas_Bayes

has occurred, is

$$\begin{aligned} P(A|B) &= \frac{P(A\cap B)}{P(B)} \\ &= \frac{P(A)\cdot P(B|A)}{P(A)\cdot P(B|A) + P(A')\cdot P(B|A')}. \end{aligned} \tag{1}$$

with

$$\begin{aligned} P(A\cap B) &= \text{probability of both event } A \text{ and } B \text{ happen simultaneously,} \\ P(B|A) &= \text{probability of event } B \text{ happens, given that event } A \text{ has occurred,} \\ P(A) &= \text{probability of event } A \text{ happens,} \\ P(B) &= \text{probability of event } B \text{ happens,} \\ P(A') &= \text{probability of event } A' \text{ (complement of } A\text{), happens.} \end{aligned}$$

For example, ***if we want to know the probability of a label is fiction, given that a term is fantasy***, equation (1) can be denoted with $A =$ fiction (F), $A' =$ non-fiction (N)[4], and $B =$ fantasy as

$$P(\text{F}|\text{fantasy}) = \frac{P(\text{F})\cdot P(\text{fantasy}|\text{F})}{P(\text{F})\cdot P(\text{fantasy}|\text{F}) + P(\text{N})\cdot P(\text{fantasy}|\text{N})}. \tag{2}$$

However, what Jabez shall compute is the probability of a label given a sentence instead of a term. Therefore, for computing this kind of probability, Jabez shall compute as follows: given a sentence consisting $n$ terms, $t_1$, $t_2$, …, $t_n$, and a label $c$,

$$\begin{aligned} P(c|t_1t_2\ldots t_n) &= \frac{P(c\cap t_1t_2\ldots t_n)}{P(t_1t_2\ldots t_n)} && \text{we use Bayes theorem} \\ &= \frac{P(c)\cdot P(t_1t_2\ldots t_n|c)}{P(t_1t_2\ldots t_n)} && \text{again Bayes theorem} \\ &= \frac{P(c)\cdot P(t_1|c)\cdot P(t_2|c)\ldots P(t_n|c)}{P(t_1t_2\ldots t_n)} && \text{Bayes' simplicity} \\ &\approx \underbrace{P(c)}_{\text{prior probability}} \cdot \underbrace{\prod_{k=1}^{n} P(t_k|c)}_{\text{posterior probability}} && \text{final version} \end{aligned} \tag{3}$$

Prior and posterior probability in Equation (3) will be calculated based on the **labeled collection** as follows:

$$P(c) = \frac{\text{number of books that have label } c}{\text{number of books in all collection}}, \tag{4}$$

$$P(t_k|c) = \frac{\text{number of occurrences of } t_k \text{ in books that have label } c}{\text{number of occurences of all terms that have label } c} \tag{5}$$

[4] $P(\text{fiction}) + P(\text{non-fiction}) = 1$

To determine labels for books in the unlabeled collection, Jabez need to compute Equation (3) from all books in labeled collection. The process of computing $P(c)$ and $P(t_k|c), \forall t_k \in$ labeled collection is called *training phase* and the process of utilizing $P(c)$ and $P(t_k|c)$ in order to assign a label to an unlabeled book is defined as *testing phase*.

To determine the best label for an labeled book, Jabes will choose the most likely or *maximum a posteriori* (MAP) label $c_{\text{map}}$ defined as:

$$c_{\text{map}} = \arg\max_{c \in \mathbb{C}} \left[ \hat{P}(c) \prod_{k=1}^{n} \hat{P}(t_k|c) \right]. \tag{6}$$

where

$$\begin{aligned} c &= \text{a label, for example, } \textit{fiction} \text{ and } \textit{non-fiction} \\ \mathbb{C} &= \text{collection of labels, for example, } \{\textit{fiction}, \textit{non-fiction}\}, \\ t_k &= k\text{-th term, for } k = 1, \ldots, n. \end{aligned}$$

Jabez writes $\hat{P}(c)$ for $P(c)$ and $\hat{P}(t_k|c)$ for $P(t_k|c)$ because Jabez does not know the true values of the parameters $P(c)$ and $P(t_k|c)$; both $\hat{P}(c)$ and $\hat{P}(t_k|c)$ are estimated from the labeled collection. There are many conditional probabilities multiplied in Equation (6). This can result in a disastrous event called *floating point underflow*. Therefore, Jabez prefers adding logarithms of probabilities to multiplying them. Hence, the problem of finding the MAP can be cast as

$$c_{\text{map}} = \arg\max_{c \in \mathbb{C}} \left[ \log_{10} \hat{P}(c) + \sum_{k=1}^{n} \log_{10} \hat{P}(t_k|c) \right]. \tag{7}$$

Yabes also forecast that there are many zeros in calculating Equation 7. Therefore, he will use *add-one* or *Laplace smoothing*, which ***simply adds one to each count of terms*** in $\hat{P}(t_k|c)$ from $k = 1$ to $n$.

Unfortunately, at the same time the secret of Jabez as the sole survivor of Thomas Bayes has been compromised; consequently, the US Goverment will promptly pick him up and bring him back to US soil. Accordingly, Jabez ask you to implement the algorithm in order to categorize sentences from the unlabeled collection. Additionally, all floating numbers should be rounded to six decimal places with respect to Jabez' recommendation.

## *Input*

The input will consist of information for a number of categorization problems. Each categorization problem has some labeled sentences and one unlabeled sentence. The information for each categorization problem will consist of:

a) a line containing 1 integer ($0 < n \leq 50$) and 1 *label* (a string up to 80 characters long): $n$ is divided into 2 collections—i.e., $(n - 1)$ sentences in the labeled collection and 1 sentence in unlabeled collection. *label* is the name of category/label for the labeled collection. A line `0 0` indicates there are no more categorization problem.

b) $2(n-1)$ lines where each two-line is a description of a book in labeled collection. Each description consists of two lines. The first line contains a binary integer (`0` denotes the book is NOT categorized in the *label* and `1` denotes the opposite) and the second line shows a one-sentence summary of the book. Each sentence is a string up to 80 characters long, terminated by the end of line. All strings are case sensitive.

c) A line describes a one-sentence summary of a book in an unlabeled collection. The sentence is is a string up to 80 characters long, terminated by the end of line.

## *Output*

For each categorization problem, give information whether or not the *last summary* is categorized into the *label*. Leave a blank line between the output for each pair of categorisation problems.

## *Sample Input*

```
5 Indonesia
1
Indonesian Jakarta Indonesian
1
Indonesian Indonesian Bandung
1
Indonesian Medan
0
Singaporean Singapore Indonesian
Indonesian Indonesian Indonesian Singaporean Singapore
5 Marvel
1
Quicksilver Spider-Man Quicksilver
1
Quicksilver Quicksilver Thor
1
Quicksilver Daredevil
0
Superman Batman Quicksilver[5]
Batman Superman Batman Quicksilver Quicksilver
0 0
```

[5] One of superheroes appearing in both DC and Marvel universe.

*Sample Output*

```
Classifying Document #1 with max value -3.521125
Unlabeled Document:  Indonesian Indonesian Indonesian Singaporean
Singapore has category:  Indonesia

Classifying Document #2 with max value -3.868123
Unlabeled Document:  Batman Superman Batman Quicksilver Quicksilver
does NOT have category:  Marvel
```

## References

Bayes, T. (1970). An essay towards solving a problem in the doctrine of chances. *Studies in the History of Statistics and Probability*, 1:134–153.

McGrayne, S. B. (2011). *The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, & emerged triumphant from two centuries of controversy*. Yale University Press.