You have already written a piece of code which can compute the frequency of each appearing words in a dictionary.

In English, we have more than 170000 words but our task, we have taken a smaller test case.

Let’s say we have a dictionary as follows:

1. Hello
2. You
3. Thank
4. Please
5. Buy
6. Shipment
7. Purchase
8. Receive
9. How
10. Are

**Sample Input: Hello You Buy Receive Thank You**

**Sample Output:**

**Hello 1**

**You 2**

**Thank 1**

**Please 0**

**Buy 1**

**Shipment 0**

**Purchase 0**

**Receive 1**

**How 0**

**Are 0**

Let us formulate our problem:

Assuming, we have 100 emails and 50 of them are spam emails and the remaining 50 are ham emails.

At first we need to create **two dictionaries**. One for **ham** and another for **spam**.

Spam Emails:

1. Buy Thank
2. Receive Thank You
3. …

**Join Spam Emails:** Buy Thank Receive Than You …

Ham Emails:

1. Hello How Are You
2. Thank You
3. How Are You
4. …

**Join Ham Emails:** Hello How Are You Thank You How Are You …

At first, your program will input two strings one each for spam and ham respectively. Then your program will count the frequency of the appearing words. By that way you will create two dictionaries.

**Spam Dictionary**

|  |  |  |
| --- | --- | --- |
| **Word** | **Frequency** | **Likelihood** |
| How | 40 | =40/660 |
| You | 30 | 30/660 |
| Thank | 20 |  |
| Please | 70 |  |
| Buy | 100 |  |
| Shipment | 130 |  |
| Purchase | 120 |  |
| Receive | 90 |  |
| How | 40 |  |
| Are | 20 |  |
| **Total** | 660 |  |

**Ham Dictionary**

|  |  |  |
| --- | --- | --- |
| **Word** | **Frequency** | **Likelihood** |
| How | 70 | =70/815 |
| You | 90 | =90/815 |
| Thank | 85 |  |
| Please | 120 |  |
| Buy | 30 |  |
| Shipment | 40 |  |
| Purchase | 20 |  |
| Receive | 20 |  |
| How | 140 |  |
| Are | 200 |  |
| **Total** | 815 |  |

**Prior probability:**

P(spam) = 0.4

P(ham) = 0.6

After having all of the above information, we are ready to compute hamity and spamity of each incoming email. In order to do that we need to compute the posterior probability using Bayes’ Rule.

Lets say we have a new incoming email which contains the following words: **Please Buy Shipment**

Then, you need to computer the followings:

P(spam|**Please**) = P(Please|spam)p(spam)/P(Please|spam)p(spam) + p(Please|ham)p(ham)

P(ham|**Please**) = P(Please|ham)p(ham)/P(Please|ham)p(ham) + p(Please|spam)p(spam)

P(spam|**Buy**) = P(Buy|spam)p(spam)/P(Buy|spam)p(spam) + p(Buy|ham)p(ham)

P(ham|**Buy**) = P(Buy|ham)p(ham)/P(Buy|ham)p(ham) + p(Buy|spam)p(spam)

P(spam|Shipment) =

P(ham|Shipment) =

Simple calculation:

Spamity = P(spam|Please) \* P(spam|Buy) \* P(spam|Shipment)

Hamity = P(ham|Please) \* P(ham|Buy) \* p(ham|Shipment)

If Spamity > Hamity, the email is a spam email.

Otherwise, the email is a ham email.