Email Filtering (Email Classification)

Type of email:

1. Good Emails (Hams)
2. Bad Emails (Spam)

Let us do some analysis on some (self-made) emails.

1. Regular email sample: Hello Navid, How are you? When would you be able to give me some to tomorrow to discuss about the content that was taught in the class yesterday? Regards, Shakhawat.
2. Spam email sample: Hi, we are offering a great deal. This is the very last chance. Please click the following link. Don’t miss this opportunity. Bye.

Question: How many words are there in English language? Roughly 172000 words.

For simplicity, let us consider that English has only 10 words.

1. How
2. You
3. Give
4. Discuss
5. Deal
6. Explain
7. Chance
8. Opportunity
9. Regards
10. Thank

The above words can be used to create a dictionary.

Let us do some field work.

Let us assume that we are given 100 sample emails. Out of them, 50 are spam and 50 are Ham emails.

Ham Dictionary

|  |  |  |
| --- | --- | --- |
| **Word** | **Frequency** | **Probability** |
| How | 35 | |  | | --- | | 0.11824324 | |
| You | 40 | |  | | --- | | 0.13513514 | |
| Give | 39 |  |
| Discuss | 41 |  |
| Deal | 12 |  |
| Explain | 33 |  |
| Chance | 8 |  |
| Opportunity | 3 |  |
| Regards | 40 |  |
| Thank | 45 |  |

Spam Dictionary

|  |  |
| --- | --- |
| **Word** | **Frequency** |
| How | 20 |
| You | 22 |
| Give | 8 |
| Discuss | 12 |
| Deal | 42 |
| Explain | 15 |
| Chance | 48 |
| Opportunity | 52 |
| Regards | 10 |
| Thank | 12 |