### To what extent can the Naïve Bayes machine learning algorithm be used to help filter email spam ?

Candidate Name: AYAAN MECKLAI

Centre Number: IN693

Candidate Number: 0006

Syllabus Code: 9980

Word Count: 4405

**Research Log**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | Date | Details | URL to websites visited | Student reflection |
| 1 | September 17, 2021 | Introduction to the IPQ | <https://us02web.zoom.us/j/84473512045> | Callido session - Introduction to the Research Process  1. Selecting an area of interest  2. Narrowing the to the question  3. Discussion of Callido Research Tools  4. Source Evaluation |
| 2 | September 17, 2021 | Callido Meeting | https://us02web.zoom.us/j/89411343637 | The zoom meeting discussed the process of forming a research question. We were introduced to different sample research question to get us familiar with the process |
| 3 | October 22, 2021 | Callido Meeting | https://us02web.zoom.us/j/89411343637 | In the zoom session I was able understand the different ways of researching as well as methods citing various sources |
| 4 | October 23, 2021 | Researching the topic for RQ |  | I zeroed down on 2 ideas, one being the use of artificial intelligence to translate sign language, and using the naïve Bayes algorithm to filter spam emails. |
| 5 | October 27, 2021 | Researching the topic for RQ |  | After going through various online resources I found that not a lot of quality resources were present for the first topic, so decided to choose my second research topic (Using Naïve Bayes algorithm to filter spam emails) |
| 4 | Nov 1, 2021 | This is a website which describes methods of spam filtering | https://www.sciencedirect.com/science/article/pii/S2405844018353404 | This website was very concise and had good information about my research topic. This article describes the different methods to filter and identify spam emails. Using this resource I was able to understand Naive Bayes classifiers and their use in many companies who try to prevent the distribution of spam messages. |
| 5 | Nov 10, 2021 | This website discusses the different types of spam | https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/meet.14504201146 | This is not a very well-known source, but it refers to the foundation of the spam emails as well as an explanation to the uses/ misuse of spam emails and how they can affect people in their day to day life. |
| 6 | Nov 11, 2021 | Zoom Link | <https://us02web.zoom.us/j/89496104117?pwd=NjVJVm5GckkrbjVab0w5bmpoNjZtdz09> | A zoom call with my supervisor to finalise the topic and research question. |
| 7 | Nov 12, 2021 |  | <https://scikit-learn.org/stable/modules/naive_bayes.html>  <https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c>  <https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html> | Improvising and editing the final research question.  Testing the research question through literature review, by reading articles on Naïve Bayes on the internet. |
| 8 | Nov 20, 2021 |  |  | Discussed different research papers with my supervisor, made necessary changes to the survey and worked further on writing introductions. |
| 9 | Nov 28, 2021 |  |  | Revised and improved introduction |
| 10 | Dec 3, 2021 |  |  | Discussed Naive Bayes theorem with my supervisor, understanding the method and use in a code. |
| 11 | Dec8, 2021 |  | <https://www.simplilearn.com/tutorials/machine-learning-tutorial/naive-bayes-classifier> | This video explained what exactly Naive Bayes was, and gave a brief explanation of its uses, advantages, and its limitations. |
| 12 | Dec 10, 2021 |  | http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.4.6385&amp;rep=rep1&amp;type=pdf | This research paper had many analytical methods and results which it had conducted in python on the Naives Bayes algorithm. I was able to understand the complete algorithm from here. |
| 13 | Jan 10, 2022 |  | https://www.yesware.com/blog/\_next/image/?url=https%3A%2F%2Fwww.yesware.com%2Fwp-content%2Fuploads%2Femail-spam-example.png&w=992&q=75 | This link consists of an image identifying features of a spam email, which could explain the general idea of filtering these mails. |
| 14 | Jan 11, 2022 |  | https://c3.ai/glossary/data-science/classifier/#:~:text=In%20data%20science%2C%20a%20classifier,%E2%80%9D%20or%20%E2%80%9Cperson%E2%80%9D). | This link defines classifiers and how they can be implemented in machine learning code. |
| 15 | Feb 12, 2022 |  | https://www.geeksforgeeks.org/naive-bayes-classifiers/ | This link describes the different types of Naive Bayes classifiers, it also explains how naives Bayes can be used practically( in code), rather than just explaining the formula and discussing it theoretically. |
| 16 | Feb 13, 2022 |  | https://www.javatpoint.com/machine-learning-naive-bayes-classifier | This link describes the different types of Naive Bayes classifiers, it also explains how naives Bayes can be used practically( in code), rather than just explaining the formula and discussing it theoretically. |
| 17 | Feb 16, 2022 |  |  | After thinking and looking about various research methods for my research paper I decided to use secondary sources for this research paper as it is testing the efficiency of the Naïve Bayes algorithm, it would require large data sets to answer the problem. Using primary data in the form of interviews ad surveys would not be applicable as not everyone is familiar with the Naïve Bayes algorithm and the primary data will not be able to come to a conclusion. |
| 17 | March 1, 2022 |  | Aladdinpersson. “Aladdinpersson/Machine-Learning-Collection: A Resource for Learning about ML, DL, Pytorch and Tensorflow. Feedback Always Appreciated :).” *GitHub*, https://github.com/aladdinpersson/Machine-Learning-Collection. | This link is the code my main code is based on. It describes the cleaning of data, creating and mapping of arrays as well as the implementation of Naive Bayes. This website also allowed me to view their raw data of over 6000 emails which I was then able to clean and filter. |
| 18 | March 2, 2022 |  |  | Debugging the main code and analysing the problems in the code  (zero frequency error) |
| 19 | April 5, 2022 |  | https://towardsdatascience.com/laplace-smoothing-in-na%C3%AFve-bayes-algorithm-9c237a8bdece | Read this website which resolved the zero frequency problem using Laplace Smoothing technique, which I implemented in the code |
| 20 | April 10, 2022 |  | https://www.datarobot.com/wiki/scoring/. | This website explained how machine learning algorithms were able to increase their accuracy through the use of a score based system. |
| 21 | May 1, 2022 |  |  | Started writing the conclusion |
| 22 | June 11, 2022 |  |  | Improvised the report after receiving feedback from my advisor about gaps in my report (mainly about the analysis and results from the code) |
| 23 | June 12, 2022 |  |  | Correcting citations, spelling errors |
| 24 | July 1, 2022 (and Aug 12, 2022 ) |  |  | submitting first draft ipq |
| 25 | Sept 3, 2022 |  |  | Worked on the feedback given by my supervisor |
| 26 | Sept 6, 2022 |  |  | Formatting the IPQ, by making an index and headings and subheading |
| 27 | Oct 5, 2022 |  |  | Submission of final IPQ and research log |