CS550 - Machine Learning and Business Intelligence

Machine Learning - Text Classification

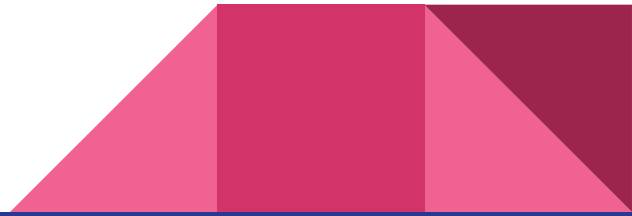
By : Prince Rajasekaran
Instructor: Dr. Chang, Henry

Table of content

- **Introduction**
- **Types of text classifiers**
- **Manual calculation**
- **Implementation of Code**
- **Conclusion**
- **References**

1. Introduction

Text classification is a machine learning technique that assigns a set of predefined categories to open ended text. Text classifiers can be used to organize, structure, and categorize pretty much any kind of text – from documents, medical studies and files, and all over the web.



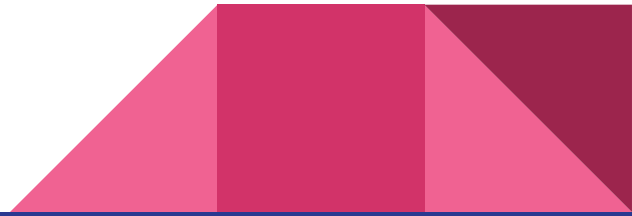
2. Types of text classifiers

There are various types of text classifiers, including rule-based classifiers, Naive Bayes classifiers, decision tree classifiers, support vector machines (SVM), and neural network classifiers. Each type has its own strengths and weaknesses, and the choice of classifier depends on the specific task and the nature of the text data being analyzed.

Overall, text classifiers are a powerful tool for automatically analyzing and categorizing large volumes of text data, which can help businesses and organizations make better decisions and improve their operations.

2. Data collection

- The quantity & quality of your data dictate how accurate our model is
- The outcome of this step is generally a representation of data which we will use for training
- Using pre-collected data



3. Manual calculation

- P(C): The probability of class C = 3/7
- P(W): The probability of class W = 2/7
- P(F): The probability of class F = 2/7
- P(W1|C): The probability that the word "W1" appears on the 3 class C documents
- = (count (W1, C) + 1) / (count(C)+|V|) = (4+1) / (12+6) = 5/18
- P(W1|W) : The probability that the word "W1" appears on the 3 class W documents
- = (count (W1, W) + 1) / (count(W)+|V|) = (1+1) / (8+6) = 2/14 = 1/7
- P(W1|F): The probability that the word "W1" appears on the 2 class F documents
- = (count(W1, F) + 1) / (count(F)+|V|) = (0+1) / (9+6) = 1/15
- P(W3|C) : The probability that the word "W3" appears on the 3 class C documents
- = (count(W3, C) + 1) / (count(C)+|V|) = (2+1) / (12+6) = 3/18 = 1/6

3. Manual calculation

- P(W3|W) : The probability that the word "W3" appears on the 3 class W documents
- = (count(W3, W) + 1) / (count(W)+|V|) = (1+1) / (8+6) = 2/14 = 1/7
- P(W3|F) : The probability that the word "W3" appears on the 2 class F documents
- = (count(W3, F) + 1) / (count(F)+|V|) = (2+1) / (9+6) = 3/15 = 1/5
- P(W4|C) : The probability that the word "W4" appears on the 3 class C documents
- = (count(W4, C) + 1) / (count(C)+|V|) = (2+1) / (12+6) = 3/18 = 1/6
- P(W4|W) : The probability that the word "W4" appears on the 3 class W documents
- = (count(W4, W) + 1) / (count(W)+|V|) = (1+1) / (8+6) = 2/14 = 1/7
- P(W4|F) : The probability that the word "W4" appears on the 2 class F documents
- = (count(W4, F) + 1) / (count(F)+|V|) = (2+1) / (9+6) = 3/15
- P(W5|C): The probability that the word "W5" appears on the 3 class C documents
- = (count(W5, C) + 1) / (count(C)+|V|) = (2+1) / (12+6) = 3/18 = 1/6

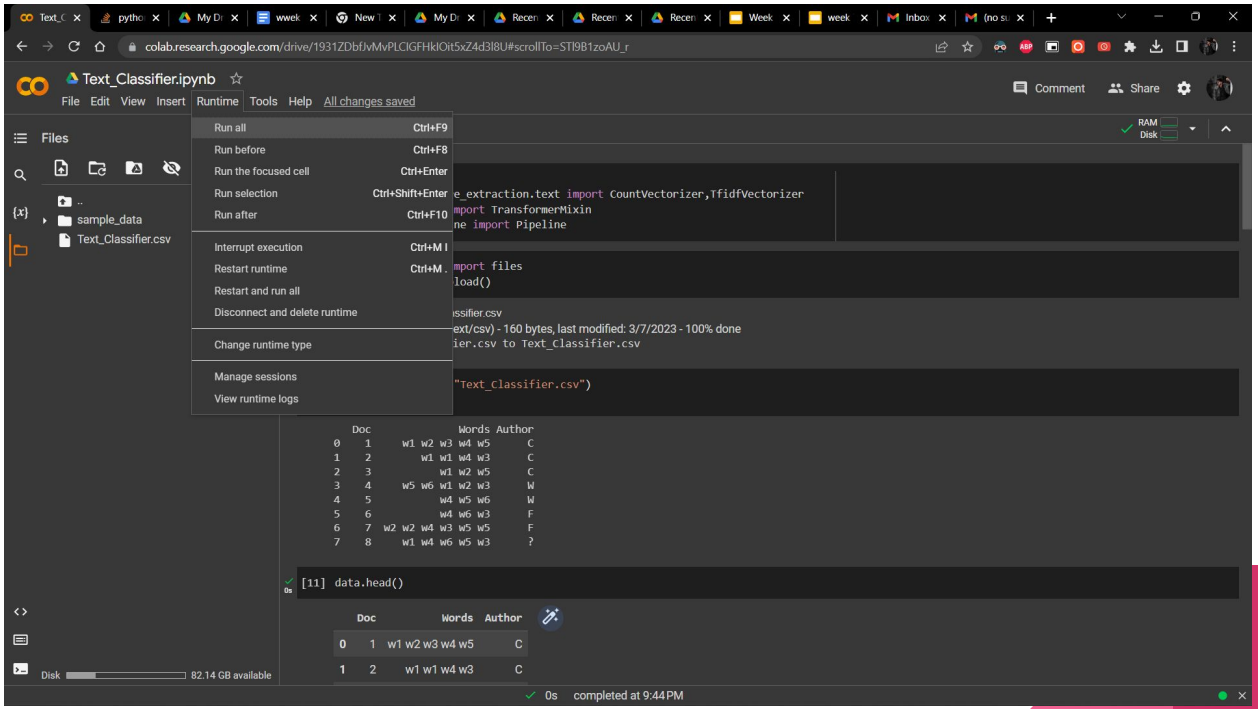
3. Manual calculation

- P(W5|F): The probability that the word "W5" appears on the 2 class F documents
- = (count(W5, F) + 1) / (count(F)+|V|) = (2+1) / (9+6) = 3/15
- P(W6|C): The probability that the word "W6" appears on the 3 class C documents
- = (count(W6, C) + 1) / (count(C)+|V|) = (0+1) / (12+6) = 1/18
- P(W6|W): The probability that the word "W6" appears on the 2 class W documents
- = (count(W6, W) + 1) / (count(W)+|V|) = (2+1) / (8+6) = 3/14
- P(W6|F) : The probability that the word "W6" appears on the 2 class F documents
- = (count(W6, F) + 1) / (count(F)+|V|) = (1+1) / (9+6) = 2/15

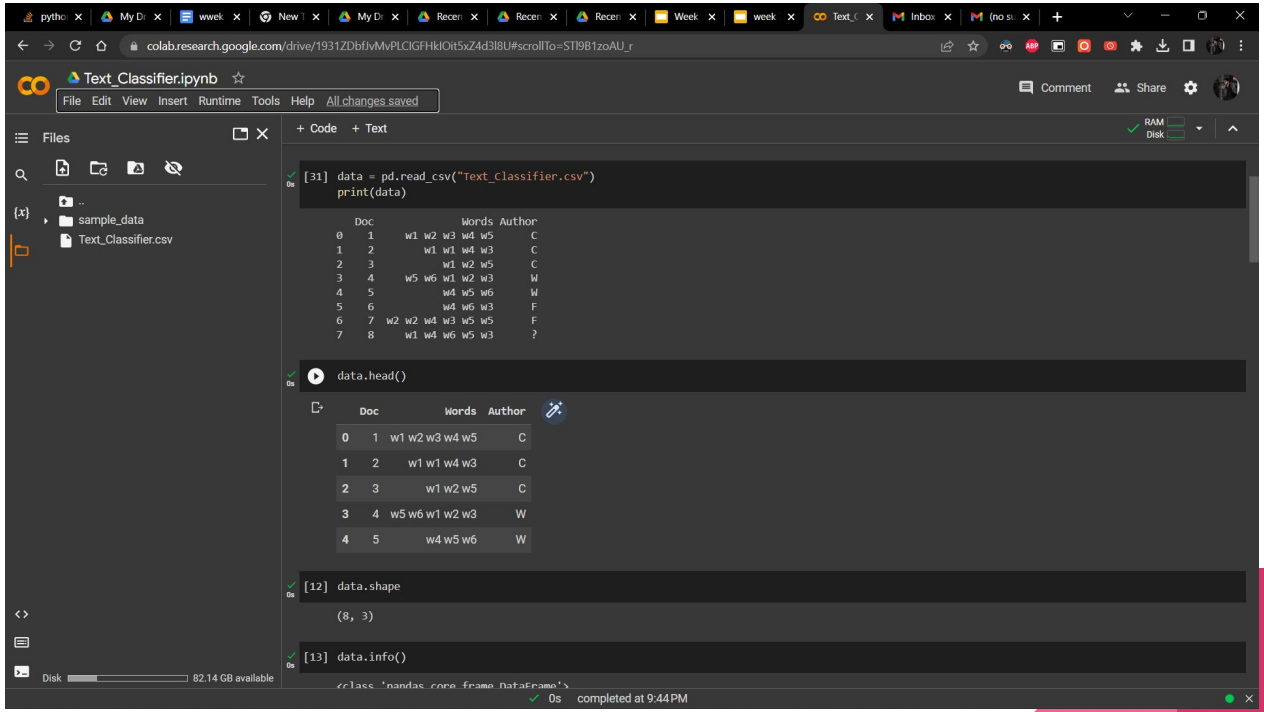
3. Manual calculation

- P(C|d8) : P(C) * P(W1|C) * P(W4|C)* P(W6|C) * P(W5|C) * P(W3|C)
- = ((3/7) * (5/18) * (1/6) * (1/18) * (1/6) * (1/6)
- = 0.00003061924 , approx. 0.00004
- P(W|d8) = P(W) * P(W1|W) * P(W4|W)* P(W6|W) * P(W5|W) * P(W3|W)
- = (2/7* 2/14 * 2/14 * 3/14 * 3/14 * 2/14)
- = 0.00002824936, approx. 0.00003
- P(F|d8) = P(F) * P(W1|F) * P(W4|F)* P(W6|F) * P(W5|F) * P(W3|F)
- = ((2/7) * (1/15) * (3/15) * (2/15) * (3/15) * (3/15))
- = 0.00002031746, approx. 0.00002
- The probability calculations show that Document 8 should be in Class C because it has the highest
- probability calculation.

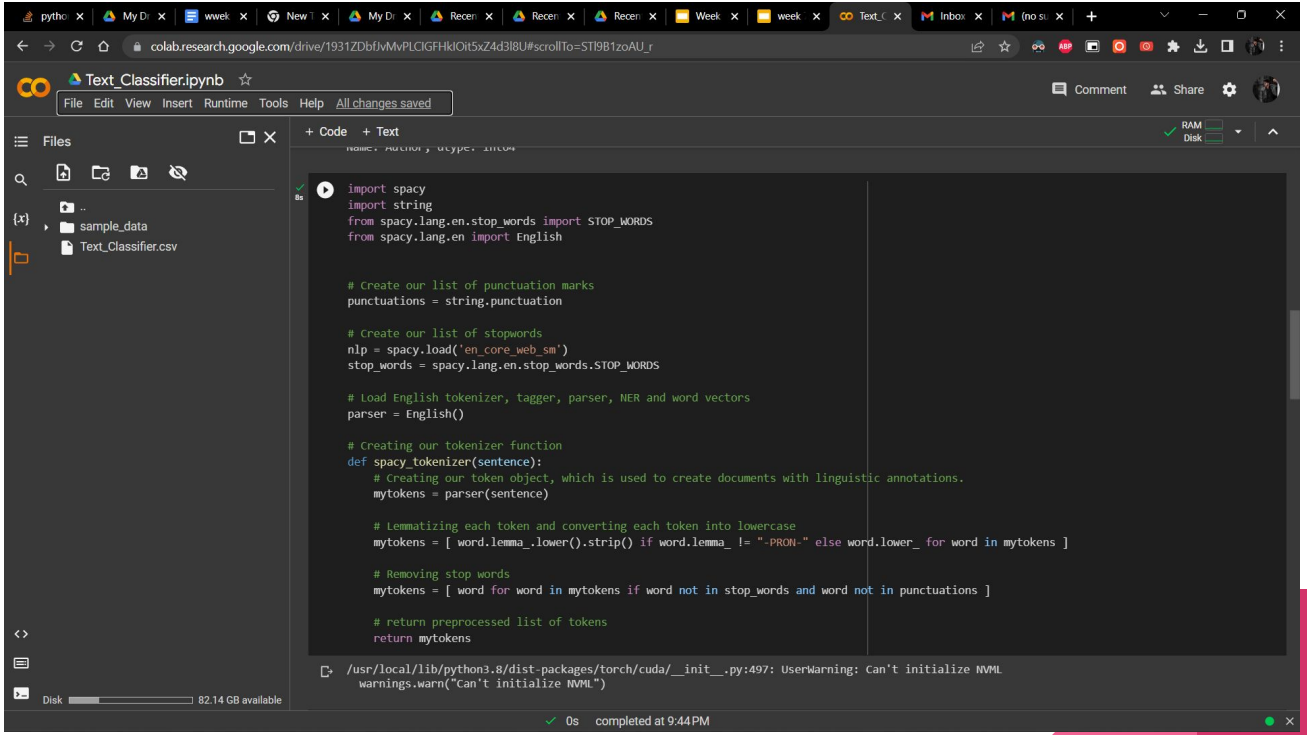
4. Implementation of Code



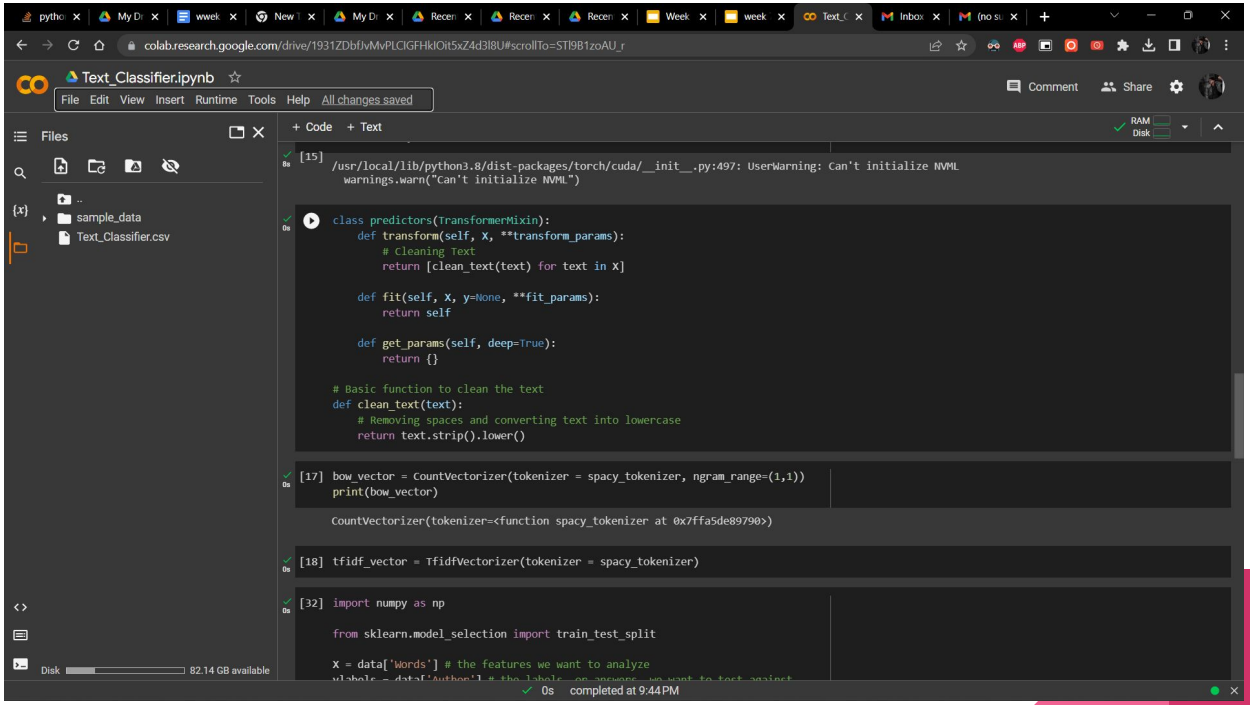
Implementation



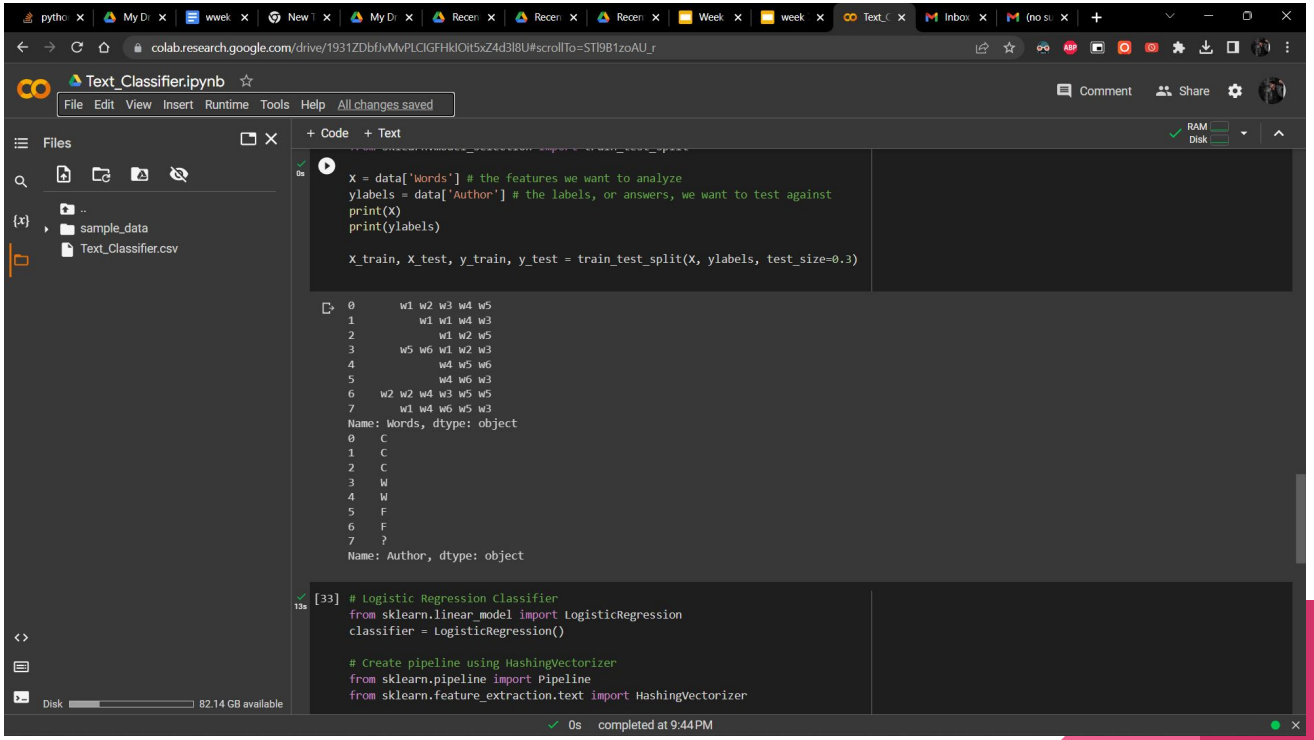
Implementation



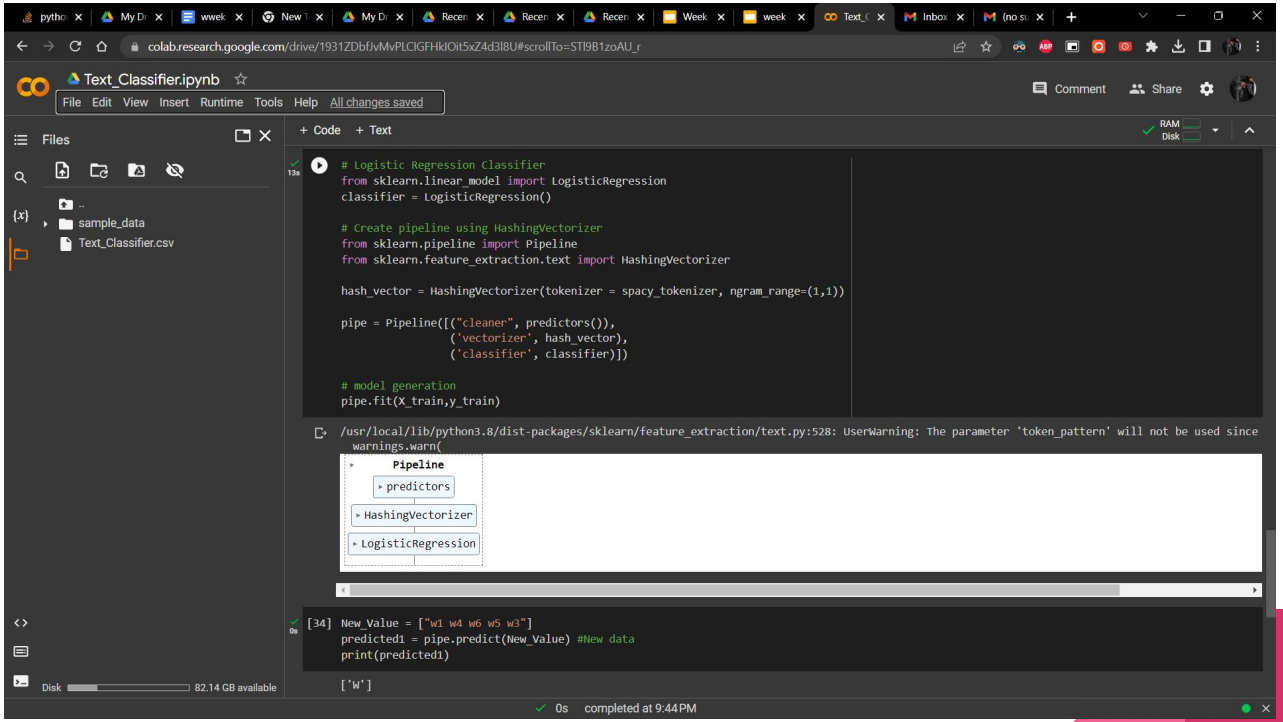
Implementation



Implementation



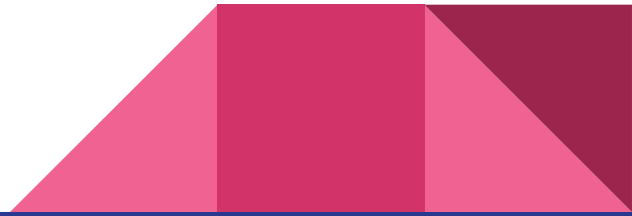
Implementation



Conclusion

In conclusion, text classifiers are an essential component of modern machine learning systems. They enable the automatic analysis and categorization of large volumes of text data, which can be used for various applications such as sentiment analysis, spam detection, topic categorization, and language identification.

Text classifiers have become increasingly important in today's data-driven world, where organizations generate and collect vast amounts of text data. With the help of text classifiers, businesses can extract valuable insights from this data and make better-informed decisions.



References

https://levity.ai/blog/text-classifiers-in-machine-learning-a-practical-guide

https://monkeylearn.com/text-classification/#:~:text=Tutorial-,What%20is%20Text%20Classification%3F,and%20all%20over%20the%20web.