University of the Witwatersrand
School of Computer Science and Applied Mathematics
COMS4054A & COMS7066A: Natural Language Processing/Technology
Lab 2 (Naive Bayes Classifier)

1 Instructions

For this lab you will be implementing the Naive Bayes Classifier discussed in the lecture. You may work in groups of two or three people. The topic which we will be experimenting on this time is tokenization and the effect of using N-grams. To complete the base version of the lab you must:

1. Read in a corpus of text either from a text file (we will be using the Harry Potter books again).
2. Split the text by whitespace and remove punctuation so that only the words are left.
3. Tokenize the words in the text (tokenizers will be allocated) - you may use libraries to help with tokenization, but your discussion must demonstrate that you understand the tokenizer.
4. Extract the N-grams of tokens from the text and collect them in an array (in order of the word's first appearance in the text).
5. Split your data into a training-validation-test split.
6. Train a Naive Bayes Classifier to predict which book a given page of HP comes from.

You may use as many pages from each book as you like to build your classifier (since this is a lot of data), but your accuracy will be taken into account for successfully completing the lab. You may **not** use libraries which implement Naive Bayes for you (like sklearn).

In addition to the above you will be expected to contribute to the class discussion where we will talk about the effect of tokenization and N-gram representations which we collectively explore throughout the lab. You will be marked within your group for the discussion but any member of the group may be asked to answer. You will be allocated a tokenizer and/or N-gram setting in the google sheet. If you have an interesting idea for a subtopic in line with this you may still recommend it. Suggested sub-topics are not final and I may still change your sub-topic but I will do my best to not be prescriptive.

You will also be expected to write a one-page (double column) report on the topic you covered (a second page may be used for figures, tables, your contribution statement and references). Please use the IEEE format for this. Please see the rubric below for more details on what should be included in the write-up. This write-up should focus on your methodology, results and insight for your particular topic. Naturally there will be overlap in what is written by each group, however, your aim should be to spend as much time discussing decisions or results for your particular setting.

2 Submission

Due Date: 4 September 2024 at 18:00.

For the submission you may work in pairs. You will be required to:

1. Submit your full code implementation (in a single file called “nbc.py”).
2. Submit a one-page report which includes a small statement on the contribution of each person if you worked in pairs. This must be completed on a new paragraph.
3. Partake in the class discussion. It is likely that you/your group will be asked to briefly explain what you found so being prepared is recommended. That said it will be fairly informal. To receive full marks for this you will need to contribute to the discussion meaningfully.

The following is an indication of the rubric which will be used to assess your write-up and discussion. Once again, negative marking will be used for poor formatting and language, and false statements. See the comment on ChatGPT below for more one this.

Mode	0% to 20%	20% to 40%	40% to 60%	60% to 80%	80% to 100%
Write-up Structure	Adequate use of language and structure.	Fair use of language and structure.	Well written and clear	Good use of language and structure.	Excellent use of language, very well written and structured
Method	No clear description of the architecture, data or approach to training	Some description of high-level details	Fair description on the details of the base model but little elaboration of the approach to answering the sub-topic	Base approach is described in detail and steps are well justified. Adequate discussion on the approach to answering the sub-topic	Excellent description and motivation for all parts of the method including on how the sub-topic was answered
Results	Results are not given or irrelevant	Some results given with inappropriate metrics	General results presented with appropriate metrics	General and sub-topic results presented with appropriate metrics	Thorough results which are appropriate for answering the sub-topic and display the general correctness of the model
Discussion Section	No interpretation of results or incorrect interpretation. Little knowledge of the topic displayed	Some general interpretation of results broadly	Results are interpreted which display some understanding of the mechanics of the model. Displays a basic understanding of the topic	Results are interpreted and contextualized to begin to answer the sub-topic. Clear demonstration of knowledge of the general topic	Results are interpreted and contextualized to answer the sub-topic and display insight into the working of the model or original thought on the concepts
Verbal Assessment (Class Discussion)	No display of knowledge on the topic.	Adequate knowledge of the topic but unable to coherently answer the question (gives random facts or irrelevant information)	Provides a concise and correct but shallow answer	The answer is clear, concise and correct. Demonstrates some deeper understanding of the topic	Demonstrates insight into the topic. Answer is clear, concise and correct. Relates the topic to other material or contextualizes the topic more broadly.

3 On the Use of ChatGPT

These labs do not count much individually for the course, and yet they are crucial to understand the material and prepare for other assessments. Thus, I urge you to take them seriously and engage with the material. I am aware though that there is an incentive to just complete these labs as quickly as possible which would result in the use of generative tools to speed up the process. This is an NLP course though and the use of ChatGPT-like software is at least a learnable experience for us. Thus, you are welcome to use generative models for the written portion of these assessments. However, greater emphasis will now be placed on the factual correctness of what is said and the degree of insight which is shown in the write-ups. If I detect something resembling a clear "hallucination" (a generative model making up facts) you will receive 0. Negative marking will also be implemented for poor writing or formatting and incorrect information or incoherent reasoning. You will also receive less marks for extremely generic facts or information. The strategy then is to use these tools to get started but then add insight in afterwards - particularly insight directed towards the sub-topic that you are investigating. Equally, clever prompt-engineering will likely go a long way here. Appropriate use of both strategies will be rewarded.

Similarly, using generative models to help you code is fine but the usual plagiarism rules for the school remain (it is not tolerated). The class discussion will then be used to check that **all** students are engaging with the course material.