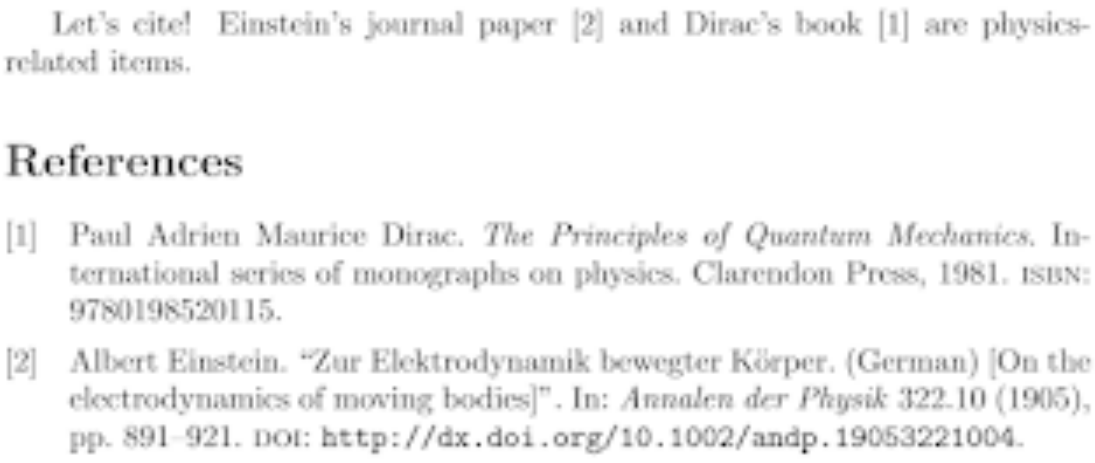
2023 Fall Data Mining Assignment #2

The goal of this assignment is to learn about the Naive Bayes Classifier (NBC).

- Build a NBC using jupyter notebook.
- Github
  - a. Create a repository in github and commit frequently.
  - b. Add readme.md for instructions about how to run your source code
- Blog
  - a. Explain NBC concepts and your contributions.
- Canvas
  - a. Create a zip file containing the following
    - source code
    - blog print (pdf)
- Add the link to the blog and github in the class directory
- For this homework, you are not allowed to use any library because NBC is very easy to implement except NLTK.
- We will use text dataset about LLM generated essay
  - a. https://www.kaggle.com/competitions/llm-detect-ai-generated-text
  - b. Given the essay, your goal is predicting whether the essay is written by human or LLM
- Below is the process.
  - a. You might want to generate some essays using ChatGPT given the prompt because the dataset contains mostly human essays.
  - b. Merge the dataset into one. And divide the dataset as train, development. In this project you don't need test dataset, because the competition has its own held-out dataset that can work as test dataset.
  - c. Build a vocabulary as list.
    - ['the' 'I' 'happy' …]
      - You may omit rare words for example if the occurrence is less than five times
    - A reverse index as the key value might be handy
      - {"the": 0, "I":1, "happy":2 , … }
  - d. Calculate the following probability
    - Probability of the occurrence
      - P["the"] = num of documents containing 'the' / num of all documents
    - Conditional probability based on the class (human or LLM)
      - P["the" | LLM] = # of positive documents containing "the" / num of all LLM documents
  - e. Calculate accuracy using dev dataset
  - f. Do following experiments
    - Compare the effect of Smoothing
    - Derive Top 10 words that predicts each class. Which word predicts the human essays most likely?

        - P[class | word]
    g. Using the test dataset
        - Use the optimal hyperparameters you found in the step e, and use it to calculate the final accuracy.
        - submit your result to kaggle and see your final accuracy.
- Documentation is the half of your work. Write a good blog post for your work and step-by-step how to guide.
- Add a reference
    a. You add a citation number in the contents and put the reference in the separate reference section



Let's cite! Einstein's journal paper [2] and Dirac's book [1] are physics-related items.

References

[1] Paul Adrien Maurice Dirac. *The Principles of Quantum Mechanics*. International series of monographs on physics. Clarendon Press, 1981. ISBN: 9780198520115.

[2] Albert Einstein. "Zur Elektrodynamik bewegter Körper. (German) [On the electrodynamics of moving bodies]". In: *Annalen der Physik* 322.10 (1905), pp. 891–921. DOI: http://dx.doi.org/10.1002/andp.19053221004.

    -
- Put the blog post and jupyter notebook link in your homepage
- Having a git commit history will save you when someone copy your work.
- Submit the jupyter notebook and blog posts as a single zip