Universität Regensburg
Lehrstuhl für Informationswissenschaft
Prof. Dr.-Ing. Bernd Ludwig

Übung Informationslinguistik 1
Wintersemester 2025/26
Studienleistung 2

Use the file `imdb.zip` (on GRIPS) for the tasks below! For each task, hand in your code (e.g. as a Jupyter Notebook) and data files containing the computed output via GRIPS!

Deadline: December 15, 2025 24:00

Task 1: N-Grams for Generative AI

Use the solution to the home work on the N-grams-assignment sheet as a basis for this task.

1. Implement a function that takes 2-, 3-, or 4-grams to generate text, if an initial $N-1$-gram is given, by sampling a next token according to the conditional distribution $P(w_N|w_1, ..., w_{N-1})$ (the solution to the home work serves as a reference)!

2. If the initial $N-1$-gram is unknown, there is no distribution to sample from and the generation process cannot start.

   A way out is to "reverse" the backoff-idea for N-grams: if $P(w_N|w_1, ..., w_{N-1})$ is unknown, I can try $P(w_N|w_2, ..., w_{N-1})$, $P(w_N|w_3, ..., w_{N-1})$, $P(w_N|w_{N-1})$, or $P(w_N)$. These distributions have already been estimated in subtask 1.

   Update your code such that it implements this idea!

   Does it help only for the initial $N-1$-gram or also later in the generation process?

Task 2: Classification with Naive Bayes

Use N-grams as features for a Naive Bayes-classifier that can distinguish between reviews from the corpus and reviews generated by your solution to Task 1!

We have two classes `orig` for reviews from the corpus and `gen` for generated reviews.

1. Implement a Naive-Bayes-classifier that takes the 3-grams in a sample (either from `orig` or `gen`) as input and maps them to `orig` or `gen`!

   Take $5,000$ random reviews from the corpus and generate $5,000$ reviews using 3-grams to estimate the conditional distribution you need for the Naive Bayes-classifier!

2. For evaluation, implement the following loop:

    - Randomly (with uniform distribution) decide whether you take a random review for the corpus that has not been used for training or generate a new review using your Generative AI!

      Your decision determines the ground truth for the sample.
    - Classify the sample using your Naive Bayes-classifier!
    - If the sample was misclassified, add it to the training data and re-estimate the conditional distributions

3. Iterate this loop $1,000$ times and report the classifier performance always after 100 iterations!

4. Can you find a trend in the performance over time?