Ethics in Natural Language Processing 2024 Homework 3

TECHNISCHE UNIVERSITÄT DARMSTADT

Due until Wednesday, 03.07.2024 at 11:59pm

Submission Guidelines for Homework

- This homework is worth 20 points
- Use the .ipynb file as a template.
- Submit your code in a single .ipynb notebook. Submit subjective answers using the given latex template.
- The dataset is a bit large this time. So upload it to your Google Drive and mount it on Google Colab. Download it from: https://hessenbox.tu-darmstadt.de/getlink/fiCd4Ym2C61sGD2dVd1RqFNh/simple_wikipedia.zip
- Extra credit shall be given to well-structured submissions.
- In case of questions or remarks, please contact:
  - Aishik Mandal, aishik.mandal@tu-darmstadt.de

Before you start, make sure you read the Submission Guidelines instructions associated with this homework for important setup and submission information. Additionally, we encourage you to use the notebook provided with this homework as a template, as we have already put a lot of code in it. Also, this will give you a head start on the assignment.

1 Toy Bi-gram LM (8 points)

You will build a toy Bi-gram LM using the plain text dump of Simple Wikipedia – a subset of Wikipedia written in simple English language, that is widely used in NLP for text simplification and other tasks. Using this Bi-gram model, you already can (1) estimate probability of sentences (2) generate sentences. However, this model is very simple and does not do a great job neither at (1), nor at (2). You will perform the following tasks:

- Find the Bi-gram probabilities.
- Find the probability of a given sequence using the Bi-gram probabilities. If you have a sequence $s = w1, w2, w3, w4$ where $w1, w2, w3, w4$ are words in the sequence, then the sequence probability is $p_s = p(w1, w2) * p(w2, w3) * p(w3, w4)$.
- Generate two sequences of length 10, one starting with the word "we", another one starting with the word "cats". Use the Bi-gram probabilities you created to generate the sequences. Select the next word based on the highest Bi-gram probability. For example, if you have a sequence $s = w1, w2, w3$, then $w4$, which has the highest probability $p(w3, w4)$, will be selected as the next word.
- The sequences generated this way will not look very useful. What is the issue with these sequences? Can you think of one way to mitigate this issue?

2 Watermarking (4 points)

Your next task is to create a toy example for watermarking. Watermarking is done so that AI generated text can be distinguished from human generated text. One approach to watermarking that we discussed in the lecture is "red-listing": in a simple version, for each word in the vocabulary, we randomly decide which other words it can("green list") and can not be followed by ("red list"). This table is used by the language model while generating text, and is later shared with the person who wants to check if the text was AI-generated. The intuition is that while a language model will not violate the "red list rule", human authors will violate it half of the time. In this task you are given the "red list / green list"(red_list.data) table and two sentences.
Your tasks for are as follows:

- Calculate the number of red list violations by each sentence by writing a function. Based on the violations, which one do you think was human-generated and which one was AI-generated?
- Calculate the probability of a human writing a sentence of the same length as s1 without ever violating the red list. Calculate the same for a sentence of the same length as s2.

3 LLM Performance (4 points)

Consider the following case based on real events. The year is 2024. You are working on a sentiment analysis task using a large language model (LLM), like ChatGPT. You are trying to reproduce the results from a paper published in 2023. The paper introduced a new sentiment analysis benchmark and made it publicly available on GitHub. The paper reports that the LLM has achieved 0.7 accuracy on the test set of the benchmark. The paper used version 1 of the LLM, which is no longer available, so you have to use version 2 of the LLM, which is the current version.
When you prompt the LLM, using the same benchmark, in exactly the same way as the paper from 2023 did, instead of 0.7, you get an accuracy of 0.95. This is a very high accuracy, and you get suspicious. There are two possible reasons for this dramatic improvement. The newer version of the LLM might have better sentiment analysis capabilities. What could be the other reason? What could you do to find out which reason behind the improvement is more likely?

4 Problems with LLMs (4 points)

You are given an LLM-generated text that illustrates a common issue encountered in machine-generated texts, which occurs when the model's knowledge about the world is inaccurate. Name the issue and give three examples in the text where this issue is present. Tip: feel free to use search engines for this task.
Prompt: "Describe the most important works of Iryna Gurevych, a computer scientist from Darmstadt, Germany".
Response:
Iryna Gurevych is a computer scientist and professor at the Technical University of Darmstadt, Germany. She is known for her work in the field of natural language processing, particularly in the areas of text analysis, information retrieval, and machine learning. Some of her most important works include:

1. "Text Classification and Clustering" (2006) - This book is a comprehensive introduction to text classification and clustering, and has been widely used by researchers and students in the field.
2. "Natural Language Processing and Information Retrieval" (2011) - This book is a fundamental text on natural language processing and information retrieval, and has been widely used by researchers and students in the field.
3. "Machine Learning for Text Analysis" (2017) - This book is a comprehensive treatment of machine learning for text analysis, and has been widely used by researchers and students in the field.

Iryna Gurevych has also published numerous papers on various topics in natural language processing, information retrieval, and machine learning, including text classification, clustering, topic modeling, and named entity recognition. Her work has had a significant impact on the development of natural language processing and information retrieval, and has influenced many researchers and students in the field. Some of her notable research contributions include:

- Development of the concept of "topic modeling" for text analysis
- Design of the " Latent Dirichlet Allocation" (LDA) algorithm for topic modeling
- Development of the "Gurevych-Model" for named entity recognition
- Research on the application of machine learning to text analysis and information retrieval

Iryna Gurevych is also a leading researcher in the field of natural language processing and has received numerous awards and honors for her work.