NLP Assignment: Adding Periods to Supergood Transcripts

Justo E. Karell

Due: 11:59pm, December 16th, 2024

Objective

Your task is to develop a model that adds missing periods to raw transcripts from comedy routines on the YouTube channel **Supergood**. The dataset, provided as unpunctuated text, must be processed to output transcripts with correctly placed periods.

This is a challenging task, and the emphasis is on your methodology and explanation of your process rather than perfect results.

Dataset

The raw transcripts are provided in the file `transcripts_separated.txt`, available for download from Canvas. The data is unpunctuated and unstructured, reflecting the spontaneous nature of live performances. There is no provided labeled or punctuated data; you may label parts of the data yourself to create training labels.

Requirements

1. **Preprocessing and Data Cleaning**:
    - Clean and preprocess the raw text to make it suitable for modeling.
    - Document all preprocessing steps clearly.
2. **Model Development**:
    - Develop a model to identify sentence boundaries and insert periods.
    - You may use any relevant NLP techniques, such as tokenization, n-grams, embeddings, or language models.
3. **Output**:

- Produce a text file with the same transcripts, but with correctly placed periods.
- The output should closely resemble punctuated English transcripts.

4. **Report (2-5 Pages)**:
   - Explain your methodology, including:
     a) Preprocessing and data cleaning steps.
     b) Any manual labeling or rule creation for training.
     c) Approach for sentence boundary detection.
   - Discuss the challenges you faced and how you addressed them.
   - Provide reflections on what worked well and what could be improved.

Grading Breakdown (100 Points)

The majority of the points are allocated to the report and explanation of your methodology:

Category	**Points**
Report and Methodology	70
Preprocessing and Data Cleaning Explanation	30
Modeling and Sentence Boundary Explanation	30
Challenges and Reflections	10
Implementation and Results	30
Output File with Inserted Periods	15
Documentation of Process in Code	10
Adherence to Assignment Instructions	5

Table 1: Grading Breakdown

Submission Instructions

Submit the following on Canvas:

- A **PDF report** (2-5 pages) explaining your methodology and reflections.
- A **text file** with the processed transcripts, now punctuated (`output_transcripts.txt`).
- Your **code files**, with sufficient comments documenting your process.

Additional Notes

This assignment encourages creativity and critical thinking. While achieving perfect accuracy is not expected, focus on a clear, well-documented methodology to approach this problem.