Word Processing Assignment

Instructions: You may use any programming language to complete this assignment using the dataset provided below. Please submit the following:

1. A PDF report answering the questions mentioned below.
2. Link to a github repository containing your code (in the report).

Dataset and Resources

First, download the English **"Natural Stories corpus"** via the link provided below and explore the various folders:

- **Corpus Repository:** `https://github.com/languageMIT/naturalstories`
- **Reference Paper:** The Natural Stories corpus: a reading-time corpus...

Required Data Files:

- **Reading Times (RT):** processed_RTs.tsv
- **Word Frequencies:** freqs folder

Part I: Preliminary Data Analysis

(1 marks) **(2 marks)** For each word in the RT file, compute the average reading time (RT) across all subjects (henceforth referred to as *mean RT per word*).

(2 marks) **(5 marks)** Plot a graph with **word length** (in characters) on the X-axis and *mean RT* on the Y-axis.

(3 marks) **(5 marks)** Plot a graph with **word frequency** on the X-axis and *mean RT per word* on the Y-axis.

(4 marks) **(2 marks)** Compute Pearson's coefficient of correlation between *length* and *frequency*.

(5 marks) **(2 marks)** Compute Pearson's coefficient of correlation between *word length* and *mean RT per word*.

(6 marks) **(2 marks)** Compute Pearson's coefficient of correlation between *word frequency* and *mean RT per word*.

Deliverable: Write a short note summarizing the relationship between word length, frequency, and mean reading time per word.

Part II: Hypothesis Testing

Hypothesis 1 (10 marks)

Hypothesis: Language model probabilities are better predictors of reading time than word frequency. Create the two regression models shown below, compare their fit to the reading time data, and choose the better model:

- **Model 1:** Mean RT per word $\sim$ word freq $+$ word length
- **Model 2:** Mean RT per word $\sim -\log(\text{gpt3 probability}) +$ word length

Discuss your results briefly with appropriate visualizations.

Hypothesis 2 (15 marks)

Hypothesis: Content words are processed differently than function words. Create the regression models shown below, compare their fit, and choose the better model:

- **Model 1:** Mean RT (content) $\sim$ word freq $+$ word length
- **Model 2:** Mean RT (content) $\sim -\log(\text{gpt3 probability}) +$ word length
- **Model 3:** Mean RT (function) $\sim$ word freq $+$ word length
- **Model 4:** Mean RT (function) $\sim -\log(\text{gpt3 probability}) +$ word length

Discuss your results briefly with appropriate visualizations.

Part III: Frequency Ordered Bin Search (FOBS)

(5 marks) Create a Frequency Ordered Bin Search (FOBS) model of human memory using the mean reading time data. Use a standard lemmatizer/lexicon to get the root of each word and arrange roots and surface forms by frequency.

Hypothesis 1 (10 marks)

Hypothesis: Root frequency predicts reading times better than surface frequency. Compare the following models:

- **Model 1:** Mean RT per word $\sim$ word freq $+$ word length
- **Model 2:** Mean RT per word $\sim$ lemma freq $+$ lemma length

Hypothesis 2 (5 marks)

Hypothesis: Pseudo-affixed words like "finger" take more processing time compared to words with regular affixes like "driver".

Task: Take 5 words (of approximately same word length and frequency) containing real and pseudo affixes and test the above hypothesis comprehensively.

Write a short note summarizing your findings pertaining to the above 2 hypotheses.

Reference: Refer to Chapter 3 of the Traxler textbook posted on Moodle for details on the FOBS model.