Cross Entropy Analysis of Literary Texts

Course Details

- **Course Code**: COMP41730
- **Course Name**: Generative AI
- **Student Name**: Yash Singh Pathania
- **Student Number**: 24204265

This document is my formal submission for the coursework in Text Analytics, which involves calculating the entropy and cross-entropy of three literary texts: *The Great Gatsby* (En), *Faust* (De), and *Candide* (Fr). This analysis uses Excel for calculations of character distributions and entropy. NOTE: Chat gpt chat link in the end

Definitions and Mathematical Formulas

Probability

The probability of each character in a text is calculated as:

$$ P(x_i) = \frac{\text{Number of occurrences of } x_i}{\text{Total number of characters}} $$

Where:

- $P(x_i)$ is the probability of character x_i.
- The total number of characters is the sum of occurrences for all characters in the text.

Entropy

Entropy quantifies the unpredictability or randomness of a distribution. The entropy $H(X)$ of a text X is calculated as:

$$ H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i) $$

Where:

- $P(x_i)$ is the probability of encountering character x_i in the text.

Explanation of the Negative Sign in Entropy

The negative sign in the entropy formula ensures that the entropy value remains positive. Since the logarithm (base 2) of any probability (a value between 0 and 1) is negative, multiplying by -1 converts the sum into a positive value. This positive value effectively represents the uncertainty or randomness inherent in the text's character distribution.

Cross-Entropy

Cross-entropy measures the divergence between two probability distributions, p and q, over the same alphabet A. The formula is:

$$ H(p, q) = -\sum_{i=1}^{n} p(x_i) \log_2 q(x_i) $$

Where:

- $p(x_i)$ is the true probability distribution (from one text).
- $q(x_i)$ is the estimated probability distribution (from another text).

Application of Cross-Entropy

Cross-entropy helps us compare how well one text's character distribution serves as a model for another. For instance, the cross-entropy of *The Great Gatsby*'s character distribution when used as a model for *Faust* can be computed to understand how similar or different their character distributions are.

Excel Implementation

The calculations were done using Excel for each of the following:

1. **Character Count**: For each of the texts, I calculated the frequency of all characters (including letters and punctuation marks) using the formula:

```
=COUNTIF(range, character) / SUM(COUNTIF(range, unique_characters))
```

2. **Entropy Calculation**: The entropy for each text was calculated by summing the individual entropies of all characters:

```
= -SUM((COUNTIF(range, unique_characters)/total_characters) *
LOG(COUNTIF(range, unique_characters)/total_characters, 2))
```

3. **Cross-Entropy Calculation**: The cross-entropy between each pair of texts was calculated using:

```
= -SUM((COUNTIF(range1, unique_characters1)/total_characters1) *
LOG(COUNTIF(range2, unique_characters1)/total_characters2, 2))
```

Results

The analysis included the following steps:

- Probabilities were calculated for each character in each text.
- Entropy values were computed for each text using the character distributions.
- Cross-entropy values were calculated between every pair of texts, helping assess the similarity between their character distributions.

Conclusion

This coursework applied entropy and cross-entropy principles to analyze three literary texts. The results provided insights into the unpredictability and informational efficiency of each text's character distribution. Furthermore, cross-entropy comparisons between different texts revealed how one text could model the character distribution of another.

Have a good read! ^-^

> **Disclaimer:**
> The formulas presented in this document for use in Excel were generated by ChatGPT. [Link] The negative sign in the entropy formula is essential to ensure that the computed entropy remains positive, as the logarithm (base 2) of a probability (a number between 0 and 1) yields a negative value. If error had to be added to the formulaes in excel as sometimes i ran into division by zero errors when there was zero probability.