DECODING REAL-WORLD DATA

Mihiliya Jayasiri (23113610)

5FTC2151 Coursework 1

TASK 1

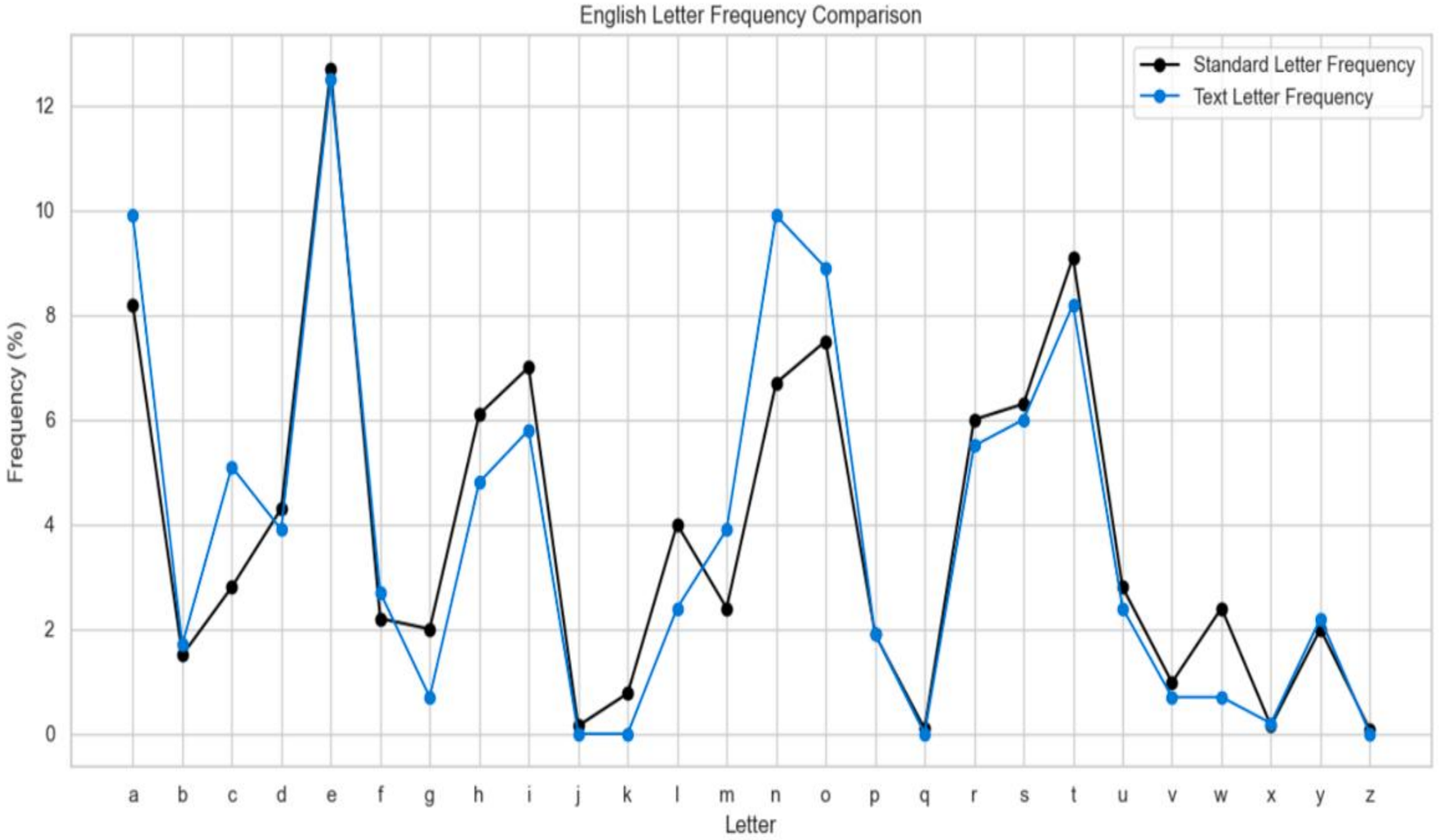
- Dataset – Wikipedia excerpt on Entropy
- Frequency – **517 characters**
- Shannon's Entropy:

$$H(X) = -\sum p(x) \log_2 p(x)$$

  - H(x) = **4.3159 bits/char**

→ Indicates moderate predictability

→ suitable for compression & encoding



- Predictability Analysis (Letter Frequency Comparison with Standard English)

→ Both have **'e' as most frequent**; other common letters align well.

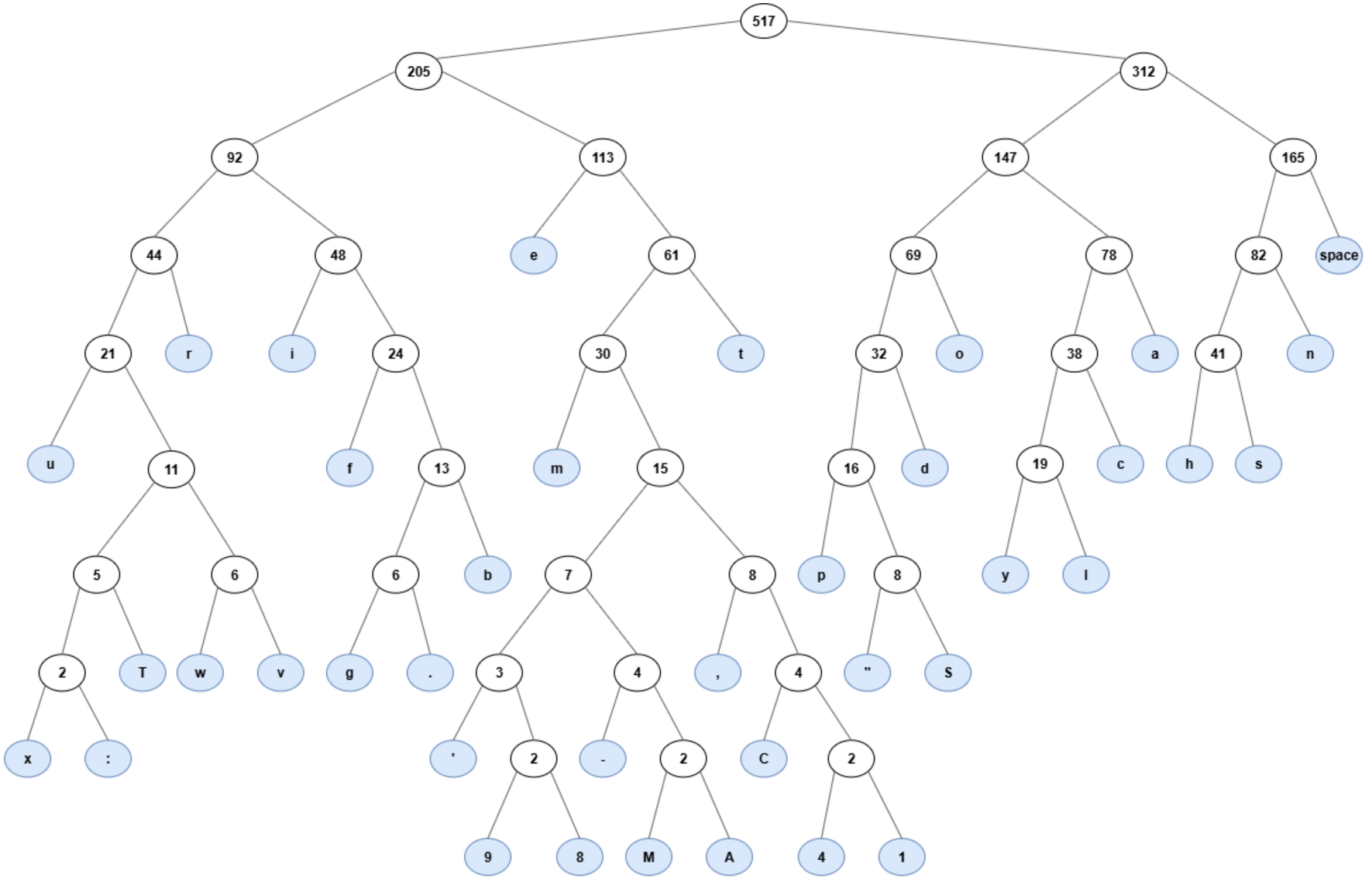
→Confirms text is **linguistically typical** and entropy value valid.

TASK 2

- Huffman Coding (lossless compression technique)
- Compression Analysis
  - Encoded text length = **2252 bits (vs 4136 bits in ASCII)**
  - Average bits/char = **4.3559**
  - Compression ratio = **1.84**

(≈ twice as compact as ASCII)

→ Key Insight:
Huffman ≈ Shannon Entropy (efficient compression)



Critical Reflection on Information Content

1

- Entropy (4.3159) shows **balance between redundancy and novelty**.
- Text is **informative, coherent, and efficient** for both humans and machines.
- Wikipedia content: **moderately predictable, rich in information, and reliable** for compression and analysis.

Reflection: Comparing to Theoretical Entropy

2

- Entropy (H): **4.3159** Average bits/char (L): **4.3559**
- Very close to theoretical bound:

$$H \leq L \leq H+1 \Rightarrow 4.3159 \leq 4.3559 \leq 5.3159$$

- Small overhead (0.04 bits/char)
- Confirms Huffman coding is **near-optimal**, efficient, and aligns with entropy theory.

TASK 3

- Binary Segment – **10011011**
- Encoded with Hamming (7, 4):
  1001100 1010101
- Single-Bit Error
  - Error : 1001100 → **1011100**
  - Syndrome: **101 (binary 5)**
  - Correction: **1001100**

Detected & corrected

- Two-Bit Error
  - Error : 1010101 → **1110111**
  - Syndrome: **001 (binary 1)**
  - Result: **1110110**

Miscorrected

Feature	Huffman Coding	LZW Compression
Compression Type	Statistical	Dictionary-based
Optimization Target	Symbol frequency	Pattern repetition
Preprocessing Required	Yes	No
Adaptability	Static (unless extended)	Highly adaptive
Best Data Types	Skewed text, grayscale images	Source code, markup, structured logs
Entropy Alignment	Direct (Shannon entropy)	Indirect (pattern entropy)

- Comparison Key Points

→ **Huffman**: Best for symbol-level optimization with known distributions.

→ **LZW**: Best for larger, pattern-rich datasets and real-time compression.

→ **Choice depends on dataset characteristics** (distribution vs. repetition) and operational needs (preprocessing vs. streaming).

TASK 4

- Non-Latin Language – **Sinhala**
- Dataset – Sinhala Wikipedia article *“Sri Lanka”*.
- Text Encoding used – **UTF-8**
- Frequency – **741 characters**
- Shannon’s Entropy:

  H(x) = 4.8561 bits/char

→ Entropy is higher than English (4.3159) by 0.5402 bits

→ Reason:
Larger alphabet + complex graphemes, diacritics and ZWJ

Feature	Sinhala	English
Alphabet Size	Over 60 characters	26 characters
Character Distribution	Evenly used across the alphabet	Some letters used more often
Script Complexity	Complex shapes and ligatures	Simple, basic letter forms

- Outcome:

→ Sinhala characters carry more information per symbol; encoding is denser despite visually shorter words.

- Challenges

→ Issue: mismatch in counts (total = 741 vs. tallied = 730).
→ Cause: Zero Width Joiner (ZWJ)

- Key Insight:

→ Grapheme clusters ≠ code points.
→ Unicode-aware analysis is essential for non-Latin scripts.