Ho Vinh Thinh (2562630)
Prabal Agarwal (2561419)
Nurzat Rakhmanberdieva (2561148)

Problem 1

For this exercise, you should select three texts, from a source such as NLTK or Project Guten- berg. Choose two in the same language (for example, English), and one in a different language (for example, German). We will consider the token-level distribution, using the maximum likelihood esti- mation with Lidstone smoothing

Three texts was chosen from Project Gutenberg: 2 of them in English and 1 in Spanish:
English: "EN_ATaleOfTwoCities.txt", "EN_Emma.txt"
Spanish: "ES_LaNarizDeUnNotario.txt"

The results of entrophy for three given texts:

- Entropy of text "EN_ATaleOfTwoCities.txt": 9.48271256
- Entropy of text "EN_Emma.txt": 9.23650977
- Entropy of text "ES_LaNarizDeUnNotario.txt": 10.06911031

The value for entropy provided above explains how many bits needed to store information content for every word in texts.

The results of KL divergence D(p||q) for three given texts:

- D("EN_ATaleOfTwoCities.txt"||"EN_Emma.txt") = 1.05025079
- D("EN_Emma.txt"||"EN_ATaleOfTwoCities.txt") = 0.98926600

- D("EN_ATaleOfTwoCities.txt"||"ES_LaNarizDeUnNotario.txt") = 4.67393790
- D("ES_LaNarizDeUnNotario.txt"||"EN_ATaleOfTwoCities.txt") = 7.96479022

- D("EN_Emma.txt"||"ES_LaNarizDeUnNotario.txt") = 5.09320823
- D("ES_LaNarizDeUnNotario.txt"||"EN_Emma.txt") = 8.40508592

In KL divergence, D(p||q) and D(q||p) are not same. That is the reason why KL divergence values for both orders are calculated.

The results above show that the difference in KL divergence between two same languages are far less than between two different languages. As words distribution in one language is more similar than words distribution in different languages. That is the reason why the results that we received for KL divergence drastically are different between same and different language texts.

Instructions to run the the code (for Linux):

```
tar -xzvf code.tar.gz
cd code
bash build
bash run file1[file2[file3[...]]]
```

bash run takes 1 or more parameters:
[file1],[file2],[file3],...: input files containing texts.

To run the code with 3 given documents:
`bash run EN_ATaleOfTwoCities.txt EN_Emma.txt ES_LaNarizDeUnNotario.txt`

The code will output the entropy of each documents as well as KL divergence between each pair of them (in both order).

Ho Vinh Thinh (2562630)
Prabal Agarwal (2561419)
Nurzat Rakhmanberdieva (2561148)

Problem 2

aaabacddbabacdd

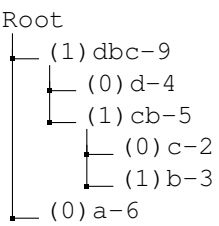
a)Use the Huffman coding algorithm (explained here: wikipedia to calculate, by hand, the encoding for the string above. Report the code obtained for each character, and use it to encode the string above.

Table 1: Huffman Code

Letter	Frequency	Code
a	6	0
b	3	111
c	2	110
d	4	10

Encoded string: 000111011010101110111011010 10

Huffman Tree



b)Using the formula in the slides, calculate the optimal length of the code. Comment on how this relates to the length of the code you found in the previous question.

The optimal length of code symbols:

$$l_i = -log_D P(s_i) \tag{1}$$

$$\begin{aligned} P(s_a) &= 6/15 \\ P(s_b) &= 3/15 \\ P(s_c) &= 2/15 \\ P(s_d) &= 4/15 \\ D &= 2 \end{aligned} \tag{2}$$

The expected length (per symbol) of a text after coding:

$$\begin{aligned} \sum_i l_i * P(s_i) &= \sum_i -log_D P(s_i) * P(s_i) \\ &= -(log_2(6/15) * (6/15) + log_2(3/15) * (3/15) + log_2(2/15) * (2/15) + log_2(4/15) * (4/15)) \\ &\approx 1.88924643 \end{aligned} \tag{3}$$

The expected length of the whole text after coding:

$$\begin{aligned} Length(text) * \sum_i l_i * P(s_i) &\approx Length(\text{"}aaabacddbabacdd\text{"}) * 1.88924643 \\ &= 15 * 1.88924643 \\ &= 28.33869643 \end{aligned} \tag{4}$$

See that this expected length is only a little bit different from the actual length of the encoded string in Huffman coding algorithm.

$$Length(\text{"}000111011010101110111011010\text{"}) = 29 \approx 28.33869643 \tag{5}$$

Ho Vinh Thinh (2562630)
Prabal Agarwal (2561419)
Nurzat Rakhmanberdieva (2561148)

Bonus

The joint entropy H(X,Y) measures the entropy of a joint probability distribution, and is given by the following formula:
Prove the following identity between the joint and conditional entropy: H(X,Y) = H(X) + H(Y|X).
We have:

$$H(X) = -\sum_x p(x).log_2(p(x)) \tag{6}$$

We also have:

$$H(Y|X) = -\sum_x \sum_y p(x,y).log_2(p(y|x)) \tag{7}$$

$$= -\sum_x \sum_y p(x,y).log_2\left(\frac{p(x,y)}{p(x)}\right) \tag{8}$$

$$= -\left(\sum_x \sum_y p(x,y).log_2(p(x,y)) - \sum_x \sum_y p(x,y).log_2(p(x))\right) \tag{9}$$

Using margin distribution, we have:

$$-\sum_x \sum_y p(x,y).log_2(p(x)) = -\sum_x p(x).log_2(p(x)) = H(X) \tag{10}$$

So:

$$H(Y|X) = -\left(\sum_x \sum_y p(x,y).log_2(p(x,y)) + H(X)\right) \tag{11}$$

$$H(Y|X) = H(X,Y) - H(X) \tag{12}$$

$$H(X,Y) = H(X) + H(Y|X) \tag{13}$$