Programming Homework 5

CS540

June 23, 2019

1 Instruction

Please submit your output files and code on Canvas → Assignments → P5. Please do not put code into zip files and do not submit data files. The homework can be submitted within 2 weeks after the due date on Canvas without penalty (50 percent penalty after that).
Please add a file named ”comments.txt”, and in the file, you must include the instructions on how to generate the output, for example:

- Data files required: train.csv, test.csv. Run: main.jar.
- Data folder required: data/train1.png ... data/train100.png . Compile and Run: main.java.

2 Details

All the requirements are listed on the course website. The following is only an example workflow to solve the problem.

1. Copy and paste the script to a text file and read the text file as a long string.
2. Replace all characters that are not ”a”, ”b”, ..., ”z” or ” ” by space ” ”, and remove all consecutive spaces, i.e. ” ” should become ” ”.
3. Count of the number of ”a”, ”b”, ..., ”z” and ”aa”, ”ab”, ”ac”, ..., ”zz” and ”aaa”, ”aab”, ”aac”, ..., ”zzz” in the document.
4. Estimate the transition probabilities using maximum likelihood estimation.

$$\begin{aligned}\hat{\mathbb{P}}\left\{x\right\} &= \frac{c_x + 1}{\text{total number of characters } + 27} \\ \hat{\mathbb{P}}\left\{y|x\right\} &= \left(\frac{c_{xy} + 1}{c_x + 27}\right) \\ \hat{\mathbb{P}}\left\{z|xy\right\} &= \left(\frac{c_{xyz} + 1}{c_{xy} + 27}\right)\end{aligned}$$

5. Start a sentence from ”a”, ”b”, ”c”, ..., ”z”. Generate the next character using $\hat{\mathbb{P}}\left\{y|x\right\}$. Generate 98 other chracters using $\hat{\mathbb{P}}\left\{z|xy\right\}$. For example, if you start with ”a”, use the $(p_a, p_b, ..., p_z, p_{\text{ space }})$

distribution with $p_x = \hat{\mathbb{P}}\left\{x|a\right\}$ to generate the next character. Given the next random character, say, "t", use the $(p_a, p_b, ..., p_z, p_{\text{ space }})$ distribution with $p_x = \hat{\mathbb{P}}\left\{x|at\right\}$ to generate the next character. Given the next random character, say, "e", use the $(p_a, p_b, ..., p_z, p_{\text{ space }})$ distribution with $p_x = \hat{\mathbb{P}}\left\{x|te\right\}$ to generate the next character. And so on.

In order to generate a character given a distribution $(p_a, p_b, ..., p_z, p_{\text{ space }})$, generate a uniform random variable u.

If $0 < u \leqslant p_a$, output "a".

If $p_a < u \leqslant p_a + p_b$, output "b".

If $p_a + p_b < u \leqslant p_a + p_b + p_c$, output "c".

If $p_a + p_b + p_c < u \leqslant p_a + p_b + p_c + p_d$, output "d". And so on.

If $\sum\limits_{i=a}^{z} p_i < u \leqslant 1$, output " ".

To implement this, you can compute the CDF first:

$$\left(p_a, p_a + p_b, p_a + p_b + p_c, ..., \sum_{i=a}^{z} p_i, 1\right)$$

Then search for the first time u is larger than a specific value in the CDF and output the corresponding character.

6. Output the transition matrix and the sentences. The transition matrix should have 27 lines, comma separated.

$$\mathbb{P}\left\{a|a\right\}, \mathbb{P}\left\{b|a\right\}, ..., \mathbb{P}\left\{z|a\right\}$$
$$\mathbb{P}\left\{a|b\right\}, \mathbb{P}\left\{b|b\right\}, ..., \mathbb{P}\left\{z|b\right\}$$
$$...$$
$$\mathbb{P}\left\{a|z\right\}, \mathbb{P}\left\{b|z\right\}, ..., \mathbb{P}\left\{z|z\right\}$$