

DATA.STAT.840 Statistical Methods for Text Data Analysis

Exercises for Lecture 1: Introduction

Answered to problems 1.1, 1.2, 1.4, 1.3, 1.5

1.1:

- Emails
 - Spam mail detection.
- Bug reports
 - What the bug report is about? Automatic detection to eg. to categorize reports to certain type or to point to certain code file.
- Real-time text
 - How to detect what user is looking for (eg. in case of chatbots in online store). Or alternatively, how to detect if the person typing is the bot.
- Log files
 - What has happened on the server on last week?
- Device names connected to internet
 - Eg. how many Cisco routers in use can be found. Or which type of devices are connected to local network.

1.2:

I chatted with Mitsuku (<http://www.square-bear.co.uk/mitsuku/nfchat.htm>). Here is the conversation:

- Happy Thanksgiving to all my friends in the USA. Enjoy your day.
 - im not from usa
- Oops! Sorry. Where are you located?
 - Finland. But can i still enjoy my day?
- Finland is a country in northern Europe. Interesting. I want to find out more about you.
 - Yes i know. But can I enjoy my day?
- How do you know? Interesting. I want to find out more about you.

Pretty obvious that I'm discussing with a bot. It couldn't answer my question targeted to the original welcome message so it seemed to lack of long term memory. Also perfectly repeated sentence feels suspicious.

1.3:

1.4:

By the chain rule of probability we know that:

$$\begin{aligned} p(w_1, \dots, w_N) &= p(w_1 | w_2, \dots, w_N) p(w_2, \dots, w_N) \\ &= p(w_1 | w_2, \dots, w_N) p(w_2 | w_3, \dots, w_N) p(w_3, \dots, w_N) \end{aligned}$$

$$\begin{aligned}
&= \dots \\
&= p(w_1 | w_2, \dots, w_N) p(w_2 | w_3, \dots, w_N) \dots p(w_{N-1} | w_N) p(w_N) \\
&= \prod_{i=1}^N p(w_i | \bigcap_{j=i+1}^N w_j)
\end{aligned}$$

Also, by noting that $p(w_1, \dots, w_N) = p(w_N, \dots, w_1)$ we get:

$$p(w_1, \dots, w_N) = p(w_N, \dots, w_1) = \prod_{i=1}^N p(w_i | \bigcap_{j=i+1}^N w_j) = \prod_{i=1}^N p(w_i | \bigcap_{j=1}^{i-1} w_j)$$

It was defined that

$$\mathbf{Left}_i = [w_1, \dots, w_{i-1}] = \bigcap_{j=1}^{i-1} w_j \text{ and } \mathbf{Right}_i = [w_{i+1}, \dots, w_N] = \bigcap_{j=i+1}^N w_j.$$

Therefore:

$$\prod_{i=1}^N \frac{p(w_i | \mathbf{Left}_i)}{p(w_i | \mathbf{Right}_i)} = \frac{\prod_{i=1}^N p(w_i | \mathbf{Left}_i)}{\prod_{i=1}^N p(w_i | \mathbf{Right}_i)} = \frac{\prod_{i=1}^N p(w_i | \bigcap_{j=1}^{i-1} w_j)}{\prod_{i=1}^N p(w_i | \bigcap_{j=i+1}^N w_j)} = \frac{p(w_1, \dots, w_N)}{p(w_N, \dots, w_1)} = 1$$

1.5:

Code output:

```
[1] 0.001371799 0.002609034 0.005724151
```