14.

This a numerical calculation exercise, where we will use the posterior predictive of the Dirichlet-multinomial to predict the next letter in a sequence. We will work with the bag of words model seen in this chapter.

Intro

Solution

a)

$$p(x_{2001} = e|D) = \frac{\alpha_e + N_e}{\alpha + N} = \frac{10 + 260}{270 + 2000} = \frac{270}{2270} = 11.9\% \tag{1}$$

b)

$$p(x_{2001} = p, x_{2002} = a|D) = \frac{10 + 87}{270 + 2000}\frac{10 + 100}{270 + 2001} = \frac{97 \times 110}{2270 \times 2271} = \tag{2}$$

0.207%

Conclusion

In this exercise, we calculated the posterior predictive of the next letter in a sequence of letters modeled with bag of words. It is interesting to note the decrease in the probability when we work with a batch of data in item b. This is due to our assumption that letters are sampled indepently. In a more accurate model, we would have to take into account the relationship between the letters of the Roman alphabet.