Master BeNeFri in Computer Science

Course: Digital Humanities
Fall 2022

Exercise #5: Probabilistic models

Instructions

Download from the ILIAS website the Federalist Papers dataset (filename: "federalist-papersNew2.csv". This file contains a large data frame with the 85 Federalist papers, where *row* = *paper* and *column* = *word-type*. The field "AUTHOR" indicates the author of the paper. Extract the papers written by "Hamilton" (51 papers), "Madison" (14 papers), and the test set (12 papers written by "Hamilton OR Madison"). This test set is {49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 62, 63}.
This exercise series consists of 6 practical questions. Upload your answers and the source code used for your computations to the ILIAS website. You can submit either a .pdf file or comment the source .py file. IPython/Jupyter notebook files (.ipynb) are allowed as well.

Practical Questions (Federalist Papers)

1) Considering the words "to", "upon" and "would", draw a graph representing the occurrences of those words in Hamilton and Madison's articles.

2) Model these three words as a Binomial distribution, to reflect either occurrences in Hamilton or Madison's writing style (you only need to estimate the parameters p and n).

3) If $p = 0.001$ and $n = 5000$, what is the probability (according to a Binomial) that we observe 5 occurrences of the underlying word-type?

4) Represent in a histogram the article lengths. Does it make sense to consider this distribution as a Gaussian one?

Practical Questions (Direct estimation vs Laplace smoothing)

In a small experiment, Mary has generated a language model with a training corpus. In this corpus, she counted 555 distinct word-types, and 5,050 tokens. In addition, she obtained the following information:

Unigram	Frequency	Bigram	Frequency
△	1000	△ today	3
the	1350	today the	4
big	320	the big	210
increase	10	big deal	11
deal	15	deal △	1
today	25	△ the	580
bank	8	the stock	8
stock	25	stock is	2
will	56		
is	132		

Mary wants to evaluate the probability of obtaining two sequences: "*today, the big deal*" and "*the stock is decreasing*". Mary is applying a direct estimation as a probability estimate (also called the "maximum likelihood principle"). On the other hand, Ann suggests that she can apply Laplace smoothing for better estimations.

5) Compute the probabilities of obtaining the two sequences with and without Laplace smoothing.

6) Provide an example of one drawback of applying the direct estimation as suggested by Mary. Provide as well two drawbacks related to Laplace smoothing.