Some Background on the `stylest` package

Elisa Wirsching

February 16, 2023

For simplicity, suppose we have 2 speakers: $S_c = \{s, t\}$.

Likelihood of Word Frequency from Data

For a speech i by speaker s and randomly chosen word w, the (log) probability that this word is $v \in V_c$ is

$$\log Pr(w = v|s) = \eta_{sv} \tag{1}$$

Can get a speaker-specific vector η_s for all words and speeches.

Posterior Probability of Authorship

Suppose we randomly pick a word type v and a word token w from speech i. Then, the posterior probability that speech i is given by speaker s

$$Pr(s|w = v) = \frac{Pr(w = v|s) \times Pr(s)}{Pr(w = v|s) \times Pr(s) + Pr(w = v|t) \times Pr(t)} \tag{2}$$

Measure of Distinctiveness

Intuition: distinctive speaker if we can determine speaker's authorship of a given speech with relatively high probability (i.e. high posterior probability)

With equal prior probabilities for whether s or t is the speaker of a speech i, we can obtain the posterior log-odds of authorship for speech i for a word type v and token w as:

$$\log\left(\frac{Pr(s|w = v)}{Pr(t|w = v)}\right) = \log\left(\frac{Pr(w = v|s) \times Pr(s)}{Pr(w = v|t) \times Pr(t)}\right) = \log\left(\frac{Pr(w = v|s)}{Pr(w = v|t)}\right) = \eta_{sv} - \eta_{tv} \tag{3}$$

The expected value across word types and word tokens of this is the *distinctiveness of speaker s* for speech i (with x_i being the number of word tokens in speech i equal to v and n_i being the length of speech i):

$$E\left(\log\left(\prod_{w \in n_i}\prod_{v \in V_c} Pr(s|w) - Pr(t|w)\right)\right) = \frac{1}{n_i}\sum_{v \in V_c} x_{iv}(\eta_{sv} - \eta_{tv}) \tag{4}$$

This can be generalized to larger reference speaker sets and larger number of speeches. See Huang, L., Perry, P., & Spirling, A. (2020). A General Model of Author "Style" with Application to the UK House of Commons, 1935–2018. Political Analysis, 28(3), 412-434. doi:10.1017/pan.2019.49.