COMP 330 Assignment 5 Written

Thomas Herring

November 29, 2018

Task 1. Give us the frequency position of the words applicant, and, attack,protein, and car. These should be values from 0 to 19,999, or -1 if the word is not in the dictionary,because it is not in the to 20,000.

Frequency positions:
application - 448
and - 2
attack - 514
protein - 3167
car - 652

Task 2. Once you have completed this task, you will get credit by (a) writing up your gradient update formula,and (b) giving us the fifty words with the largest regression coefficients. That is, those fifty words that are most strongly related with an Australian court case.

$$
\begin{aligned}
\theta_i &= \sum_j r_i * x_{i,j} \\
LLH &= \sum_i -y_i\theta_i + log(1 + e^{\theta_i} + \zeta \sum_j r_j^2 \\
\frac{\partial f}{\partial r_j} &= -x_{i,j}y_i + x_{i,j}\frac{e_i^\theta}{1 + e_i^\theta} + 2\zeta r_j
\end{aligned}
$$

Fifty words with largest coefficients:
[u’satisfied’] [u’applicant’] [u’hca’] [u’pursuant’] [u’fca’] [u’respondent’] [u’relevantly’] [u’tribunal’] [u’whether’] [u’relevant’] [u’clr’] [u’pty’] [u’submissions’] [u’alr’] [u’circumstances’] [u’reasons’] [u’relation’] [u’fcafc’] [u’affidavit’] [u’proceedings’] [u’respect’] [u’mr’] [u’interlocutory’] [u’honour’] [u’consideration’] [u’proceeding’] [u’ltd’] [u’relied’] [u’notice’] [u’jj’] [u’jurisdictional’] [u’appeal’] [u’respondents’] [u’error’] [u’evidence’] [u’appellant’] [u’multicultural’] [u’decision’] [u’judgment’] [u’gummow’] [u’subsection’] [u’hearing’] [u’claim’] [u’fmca’] [u’consider’] [u’sought’] [u’delegate’] [u’discretion’] [u’leave’] [u’application’]

Task 3. To get credit for this task,you need to compute for us the F1 score obtained by your classifierwe will use the F1 score obtained as one of the ways in which we grade your Task 3 submission.I am going to ask you to actually look at the text for three of the false positives that your model produced (that is, Wikipedia articles that your model thought were Australian court cases). Write paragraph describing why you think it is that your model was fooled. Were the bad documents about Australia? The legal system?If you dont have three false positives, just use the ones that you had (if any).

F1 = 0.935657

I found that the I had several false positives while classifing the data. Three of these are 3617964, 35399712, and 29893732. All of these are related to law and describe legal proceedings in the US and other countries. This makes sense that it threw off the model as many of the words that classify a text as an Australian legal case would match in an article on law.