

Jupyter notebook: [Assignment5.ipynb](#) / [Assignment5.html](#)

## Task 1:

```
[3]: topWordsDict = dict(topWordsDict)
    topWordsDict['applicant']
```

604

```
[4]: topWordsDict['and']
```

2

```
[5]: topWordsDict['attack']
```

515

```
[6]: topWordsDict['protein']
```

3681

```
[7]: topWordsDict['car']
```

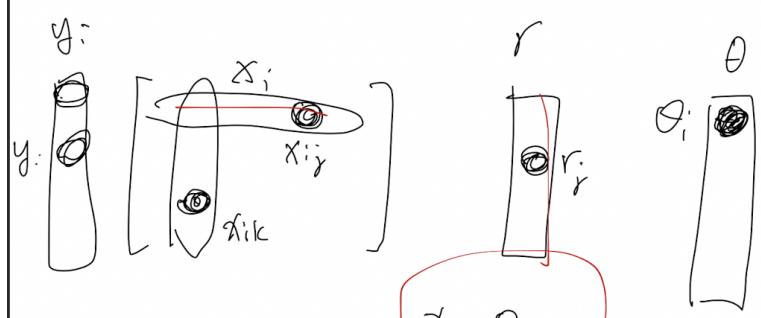
635

Task 2: (Trained on mid-size dataset, test on small dataset)

- Formula

$f$

$$\mathcal{L}(H) = \sum_i y_i \theta_i - \log(1 + e^{\theta_i}) - \lambda (\|r\|_2^2)$$

$$\frac{\partial \theta_i}{\partial r_k} = x_{i,k}$$


$$\frac{\partial f}{\partial r_k} = \sum_i \left[ y_i x_{i,k} - \frac{x_{i,k} \theta_i}{1 + e^{\theta_i}} - \lambda \cdot r_k \right]$$

$$\theta := \sum_i x_{i,j} r_j$$

$(x_1^2 + x_2^2 + x_3^2 + \dots + x_n^2)^{\frac{1}{2}}$

- 50 words:  
**matters**  
**magistrate**  
**court**  
**consider**  
**jurisdictional**  
**amp**  
**act**  
**regard**  
**respondents**

appellant  
judgment  
decision  
discretion  
notice  
orders  
sought  
ltd  
matter  
that  
proceedings  
any  
not  
relevantly  
respondent  
clr  
mr  
pursuant  
hearing  
submissions  
tribunal  
whether  
hca  
fcr  
gumnow  
fcafc  
pty  
relation  
circumstances  
applicant  
proceeding  
fca  
application  
relevant  
affidavit  
reasons  
claim  
appeal  
alr  
evidence  
satisfied

## Task 3

1.

- 3343 out of 3442 correct.
- Precision is: 0.4233128834355828
- Recall is: 0.9324324324324325
- F1 score is: 0.5822784810126582

2.

The false positive document IDs I found on the small dataset is:

```
array(['6272533', '6572113', '19179678', '13960455',  
      '39192923', '314107',  
      '11280044', '25063933', '6372080', '381369', '6164108',  
      '11200229',  
      '40949219', '13950560', '23846142', '3505201',  
      '15179722',  
      '12232081', '13722518', '27319027', '33692479',  
      '32323513',  
      'AU3577', 'AU3837', 'AU1816', '6803252', '29351291',  
      '15028621',  
      '7364668', '41722859', '30831822', '12294713',  
      '38708891',  
      '19371162', '28480970', '28340519', '3589511',  
      '12001341',  
      '3893075', '29898234', '34655588', '28079057',  
      '30034801',  
      '18621290', '6190688', '343299', '32997736', '31111239',  
      '20260638', '20338785', 'AU477', '18847030', '3385996',  
      '37599046',  
      '14563042', 'AU395', '35399712', '7285712', '12683320',  
      '19407815',  
      '38307325', '31725104', '3228611', 'AU1356', '3139076',  
      '12195728',  
      '314775', '7509365', '39565511', '12405631', '25094588',  
      '43469286', '14593350', '18825108', '36799193',  
      '14102393',  
      '25590036', '15300181', '6733026', '13566920',  
      '12542761',  
      '435161', '24753062', '3766266', '6653822', '32064267',  
      '31693217',  
      '3445338', '29289333', '20489435', '13737778',  
      '18555654',  
      '34069805', '38464259'], dtype='<U8')
```

3.

Looking at these docs, I'm not confident enough to pinpoint a certain reason for my model misclassifying these as positive. However, these articles are not like simple sports or entertainment articles but about novels, history, autography, politics. This suggests that my model probably doesn't have enough weight on the words that are actually linked to the court documents. In addition, the choice of cutoff point could also be a reason given the imbalance nature of the corpus.