STA141C Hw1 Report

Ruriko_Imai, id:912212313

4/17/2017

Problem 2

1. Overlapping scores for the first 10 lines

```
0    0.961538
1    0.523810
2    0.000000
3    0.400000
4    0.676471
5    0.500000
6    0.750000
7    0.444444
8    0.107143
9    0.588235
```

2. Maximum, minimum, median overlapping score

Minimum is 0.0. Maximum is 1.0. Median is 0.5.

3. Findings

The first 10 overlapping scores indicate that there are a range (0 to 1) of matches found in the pair of questions presented. The minimum overlapping score shows us that there are pairs of questions that are completely different from one another and the maximum score of 1 shows us that exact matches exists. It is interesting that the median results is exactly 0.5, since that would mean exactly 50% is above and below it. This indicates that most sentences seem to have an approximate match of half the words (more or less) in the sentences. (Those words are probably the stopwords that we will remove them later in the problem.)

Problem 3

1. Accuracy with threshold = 0.1,0.2,0.3,0.4,0.45,0.5,0.55,0.6,0.65,0.7,0.8,1.0 Which threshold gives you the best result?

Accuracy Result	threshold
0.43478099468843673	0.1
0.49890334934742325	0.2
0.5739148726864384	0.3
0.6350303319030988	0.4
0.6558125578100397	0.45
0.6638897224190041	0.5
0.6646847554731994	0.55
0.6601001679778009	0.6
0.6529912731974868	0.65
0.6464268368511742	0.7
0.6332330003681281	0.8
0.6638897224190041	1.0

The threshold of 0.55 gives the best result of 0.6646847554731994.

2. Using the best threshold on validation.csv.

The threshold of 0.55 gives the result of 0.6634786257281075 for validation.csv.

3. Findings

The results for both data (training and validation) gave approximately the same accuracy score. This is a good indicator because the threshold should be similar for any given data.

Looking at the accuracy results and their corresponding thresholds, all the results are within the 0.6 range from threshold of 0.4 onward and differs only by a small amount. This might mean that the accuracy scores from threshold of 0.4 onward do not differ significantly from one another.

Problem 4

1. Accuracy with threshold = 0.1,0.2,0.3,0.4,0.45,0.5,0.55,0.6,0.65,0.7,0.8,1.0 Which threshold gives you the best result?

Accuracy Result,	threshold
0.5549052302038316	0.1
0.6166827013800166	0.2
0.665074538215723	0.3
0.6737425639661322	0.4
0.6700520019674748	0.45
0.6686661077718348	0.5
0.6571118336184522	0.55
0.6520044422858592	0.6
0.6446728145098173	0.7
0.6447563393832152	0.8
0.6565735622121098	1

   The 0.4 threshold gives the best result of 0.6737425639661322.

2. Using the best threshold on validation.csv.

Using the best threshold of 0.4, the validation accuracy score is 0.6844703327080659.

3. Findings

   The accuracy results after removing the stopwords are now smaller in diffrences. By taking out the stopwords, all the words that do not hold unique meaning was taken out. Therefore, it makes sense that the range of accuracy is smaller compared to when the stopwords are still present since only unique words that are left in the questions will be compared to another to calculate its score.

   The range of the accuracy results mostly reside within the 0.6 range execept for 0.1. This means that the accuracy results following its thresholds do not differ significantly. In this result, the best threshold is 0.4. Looking back at the results before the stopwords were removed, the best threshold is 0.55. However, the threshold of 0.4 is the starting range where most results lie in (0.6). This is an interesting outcome since the compression of the ranges after the stopwords were removed and the range where the accuracy results starts to matter, both points to 0.4. Concluding that the threshold of 0.4 is indeed an optimum threshold.