CSCI 375 HW 2

Tongyu Zhou

September 26 2017

1 Written Exercise [30 points]

1.

With linear interpolation, the bigram probability of P(Sam|am) is:

$\lambda_1 P(Sam|am) + \lambda_2 P(Sam)$
$= \frac{1}{2} P(Sam|am) + \frac{1}{2} P(Sam)$ since $\lambda_1 = \lambda_2 = \frac{1}{2}$
$= \frac{1}{2} \times \frac{2}{3} + \frac{1}{2} \times \frac{4}{24}$
$= \frac{5}{12}$

2.

Katz backoff of P(Sam|I), where d = 1, is:

$\alpha(I) \times P_{\text{katz}}(Sam)$
$= \alpha(I) \times \frac{4}{24}$
where $\alpha(I) = \frac{1-(P^*(am|I)+P^*(do|I)}{P^*(<s>)+P^*(I)+P^*(Sam)+P^*(</s>)+P^*(not)+P^*(like)+P^*(eggs)+P^*(and)}$
$= \frac{1-\frac{3-d}{4}+\frac{1-d}{4}}{\frac{4}{24}+\frac{4}{24}+\frac{4}{24}+\frac{4}{24}+\frac{1}{24}+\frac{1}{24}+\frac{1}{24}+\frac{1}{24}}$
$= \frac{\frac{d}{2}}{\frac{5}{6}} = \frac{3}{5} d$
so $\frac{3}{5} \times 1 \times \frac{4}{24} = \frac{1}{10}$

2 Programming Exercise [70pts]

3.

Unigram Katz-backoff

precision: 78.74% (P), 25.97% (R)
recall: 21.73% (P), 82.40% (R)
f1: 34.06% (P), 39.50% (R)

Bigram Katz-backoff

precision: 87.99% (P), 59.96% (R)
recall: 85.53% (P), 65.00% (R)
f1: 86.74% (P), 62.38% (R)

Unigram Bayes Classification

precision: 99.69% (P), 47.61% (R)
recall: 63.53% (P), 99.40% (R)
f1: 77.61% (P), 64.37% (R)

In terms of best f1 scores, the bigram model with katz-backoff ranked the highest with 86.64% for movie plot and 62.38% for movie review. The second best was the unigram Bayes' classification with 77.6% for movie plot and 64.37% for movie review. The unigram model with katz-backoff did not perform well at all, with f1 scores less than 1/2 for both (a randomization would probably even be better in this case).

It seems reasonable for bigram models to perform significantly better than unigram models for katz-backoff, as the whole point of backing off is to rely on previous, more reliable history to generate probabilities if the ngram does not exist. A bigram would back-off to a 2 - 1 = unigram model, whereas an unigram would back-off to the whole corpus, which does not improve estimates of conditional probability at all. Even if the ngram does exist, estimation using bigrams would still be more accurate than estimation using unigrams, as the number of false positive matches would be reduced.

Comparing the unigram bayes classification with the unigram katz-backoff, it also seems probable that the former would perform better. While both models use unigrams, the former also takes into account the distribution of input for the training set (there were 3 times as many movie plots in comparison to movie reviews). This already bumps up the probability of any test set for movie plot as $\prod_{i=1}^{\infty} P(t_i|c)$ is multiplied by 3/4 instead of 1/4.

It is unclear whether the unigram bayes classification or the bigram katz-backoff performed better–alternate tests with different distributions of training sets could shed some light on this problem.

Design choices:
- preprocessing: the 'p:' and 'r:' in the beginning of each sentence was used to classify the training set, but stripped from the testing set before it was tested
- threshold for out of vocabulary words was set to 1
- each classification method (uncomment to run) prints to terminal and the output was stored in separate text files (katz1.txt, katz2.txt, bayes.txt). These were compared with the original test file in the check method for evaluation.