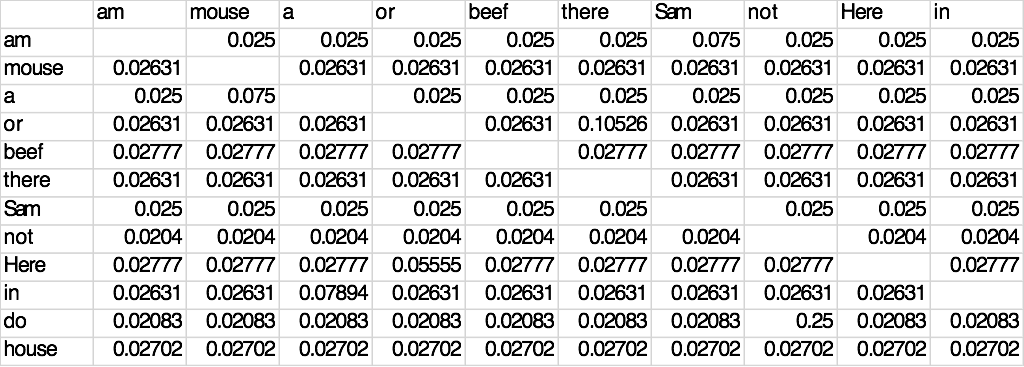
Assignment 1

Computational Linguistics

3) Double click on the table below to see the table of probabilities (spreadsheet format)



1. I do not like them in a mouse.

P(I|<s>) P(do|I) P(not|do) P(like|not) P(them|like) P(in|them) P(a|in) P(mouse|a) P(.|mouse) P(</s>|.)

From the probability table we get,

= 0.3448 x 0.21428 x 0.25 x 0.2857 x 0.2363 x 0.0612 x 0.0789 x 0.075 x 0.0526 x 0.2857

= 6.78654196e-9

1. I am Sam I am Sam

P(I|<s>) P(am|I) P(Sam|am) P(I|Sam) P(am|I) P(Sam|am) P(</s>|Sam)

From the probability table we get,

= 0.3448 x 0.07142 x 0.075 x 0.05 x 0.0714 x 0.075 x 0.075

= 3.7088e-08

1. I do like them anywhere.

From the probability table we get,

P(I|<s>) P(do|I) P(like|do) P(them|like) P(anywhere|them) P(.|anywhere) P(</s>|.)

= 0.3448 x 0.21428 x 0.02083 x 0.2363 x 0.06122 x 0.0789 x 0.2857

= 5.01859853e-7

1. I would like green ham and beef.

From the probability table we get,

P(I|<s>) P(would|I) P(like|would) P(green|like) P(ham|green) P(and|ham) P(beef|and) P(.|beef) P(</s>|.)

= 0.3448 x 0.01785 x 0.02631 x 0.9090 x 0.025 x 0.025 x 0.025 x 0.0277 x 0.2857

= 1.8201e-12

We have estimated the bigram probabilities using the Maximum Likelihood Estimation. We have taken counts

From the file nonsense.txt and normalized them so that they lie between 0 and 1. We have computed the

Probability using the following equation:

P(Wn|Wn-1) = C(Wn-1Wn)/ΣwC(Wn-1W)

This can be further simplified since the sum of all bigram counts that start with a given word Wn-1 must be equal to the unigram count for that word Wn-1:

P(Wn|Wn-1) = C(Wn-1Wn)/C(Wn-1)

For example, In the first sentence given, ‘I do not like them in a mouse.’

The probability that ‘I’ occurs after ‘<s>’ is given by P(I|<s>). This probability value can be obtained from the table

given above. Similarly, we find the probabilities of each word occurring given the previous word in the sentence. It is calculated as follows:

P(I|<s>) P(do|I) P(not|do) P(like|not) P(them|like) P(in|them) P(a|in) P(mouse|a) P(.|mouse) P(</s>|.)

From the probability table we get,

= 0.3448 x 0.21428 x 0.25 x 0.2857 x 0.2363 x 0.0612 x 0.0789 x 0.075 x 0.0526 x 0.2857

= 6.78654196e-9