Lab1 : Probability of Letters

- Markov approximation to probability of letter sequences

$$P(L) = P(l_1)P(l_2 \mid l_1)\cdots P(l_{|L|} \mid l_1,\ldots,l_{|L|-1}) \quad k-gram$$

$$\approx P(l_1)P(l_2 \mid l_1)\cdots P(l_k \mid l_1,\ldots,l_{k-1})\prod_{i=k+1}^{|L|} P(l_i \mid l_{i-1},l_{i-2},\ldots,l_{i-k})$$

Lab1: simulate Shannon’s study on English letters

1. Compute unigrams and bigrams of all letter events
2. List top and bottom 5 letters and their probabilities
3. List top and bottom 5 letter pairs and their probabilities
4. Do the above for 1000 & 10000 sentences, any difference?

Hint: compute conditional entropy given previous letters