9318 Project Report

2019 Term1

Z5219071 Yu Han

Question 1: Viterbi_algorithm

1. Calculate the matrix of state file.
    a) Read the first line and get the number of states;
    b) Record state's IDs;
    c) Use following data to calculate the transition rate and record into matrix (add smoothing for probabilities).
2. Calculate the matrix of symbol file.
    a) Read the first line and get the number of states;
    b) Read following symbols and create a symbol dictionary;
    c) Use following data to calculate the emission rate and record into matrix (add smoothing for probabilities).
3. Record Query from Query file.
    a) Read query file and use regular expression to split each word and special punctuations;
    b) Use symbol dictionary to translate these queries to symbol number;
4. Calculate the probability and record path by using Viterbi Algorithm.
    a) Use state 'BEGIN' probability to initial the matrix;
    b) Use recursive method to calculate the emission rate * transition rate of the matrix and find the maximum probability;
    c) Use the maximum probability to record the path;
    d) Canonical the output format (path + maximum probability)

Question 2: top_k_viterbi

1. 3D Matrix
    a) Created a 3D matrix to record the probability and path. In the first question, the matrix only saves the best path and probability for the query.
    b) In the second question, the 3D matrix will save the best k path and probabilities.
2. Survival of the fittest

a) If there is a better path with higher probability, the matrix will drop the worse one and add the better one. Finally, the matrix will have the top k path and probabilities.

Question 3: advanced _decoding

1. Change the smoothing

    a) I tried to change the smoothing of question 3. But the result is not good. If the smooth is higher than 1, the result will have a little more wrong label. If the smooth is less than 1, the wrong label result will be much higher than 1.