CS571-Hashtag Segmentation

Chen Lin

10 February 2019

Abstract

Hashtags found in social media do not include whitespaces, which make them hard to interpret. My task is to write a program that takes a hashtag and returns a list of tokens representing the most likely sequence of the input hashtag (e.g., 'helloworld' → ['hello', 'world']).

1 Introduction

I used 1gram.txt and 2gram.txt to do the hashtag segmentation.
The main idea is to find the maximum generation probability of hashtag segmentation, thus dynamic programming was used in this project.
For each word in 1gram.txt, its frequency(probability) was calculated as log(the number of the word/total number of words)
For each pair of words in 2gram.txt, its probability was calculated as log(the number of the pair/the number of the first word)

2 Dynamic Programming

For the input sequence, every character in the sequence was treated as node.
Record the previous node and probability of the node in a dictionary.
Use dynamic programming to find the node list to maximize the probability of segmentation.
For specific node j, calculate the probability of previous node (from 0 to j-1), select the maximum probability as the probability of node j, and record the previous node.
Use backtracking on the node list to find the segmentation with maximum

probability.

3 Result

Input 7 hashtags, the program correctly segmented 6 of them.
However, for hashtag "helloworld", it failed to segment the sequence.
There is no pair "hello world" in 2gram.txt, thus another method should be used to improve the segmentation result.