Crossparsing

1. Algorithm Description

Defination pf crossparsing

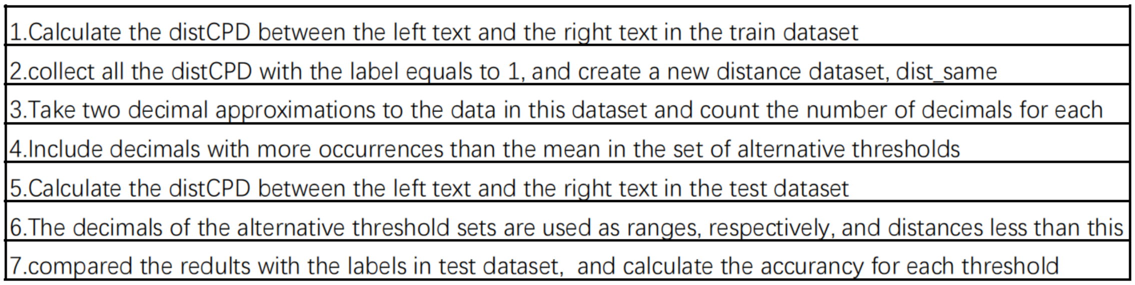
Giving two separated sequences x and y, the characters in x are $x_1, x_2, x_3, \ldots, x_m$ and the characters in y are $y_1, y_2, y_3, \ldots y_n$. Compare x and y to find the largest number t satisfying: $x_i, x_{i+1}, ..., x_{i+t} = y_j, y_{j+1}, ..., y_{j+t}$ For example, the crossparsing of x = 'abcdegj' and y='bcdggggj' is a set of $s(x|y) = \{a, bcd, e, gj\}$ and if y is included in x then we get $s(x|y)/\{y\} = s(x|y) - 1$ else $s(x|y)/\{y\} = s(x|y)$, the same can be said for,$s(x|y)/\{y\} = s(x|y) = 4$, $s(y|x) = \{bcd, g, g, g, gj\}$,$s(y|x)/\{x\} = s(x|y) = 5$

Defination of Crossparsing Distance

Given two words x and y, the crossparsing distance distCPD(x, y) between x and y is $distCPD(x, y) = \frac{\frac{|s(x|y)/\{y\}|}{|x|} + \frac{|s(y|x)/\{x\}|}{|y|}}{2}$

2.Data Implement

Based on data sets: DDP, Amazon_Google, abt_buy, in order to understand the meaning of sentences and match the sentences more accurately, first align the case of the words and remove any punctuation marks present.



1.Calculate the distCPD between the left text and the right text in the train dataset
2.collect all the distCPD with the label equals to 1, and create a new distance dataset, dist_same
3.Take two decimal approximations to the data in this dataset and count the number of decimals for each
4.Include decimals with more occurrences than the mean in the set of alternative thresholds
5.Calculate the distCPD between the left text and the right text in the test dataset
6.The decimals of the alternative threshold sets are used as ranges, respectively, and distances less than this
7.compared the redults with the labels in test dataset, and calculate the accurancy for each threshold