Sabaragamuwa
University of Sri Lanka

Faculty of Computing

Department of Computing & Information Systems

Quiz 01

IS5114 – Data Mining & Analytics

Name : G.A.P. Pathum

Reg. no : 20APC4911

Academic Year : 2020/2021

Degree program : Information Systems

Due Date : 18/07/2024

1. Output after Stemming

   lemmat algorithm would know that the word better is deriv from the word good , and henc , the lemm is good . but a stem algorithm wouldn’t be abl to do the same . stem is a techniqu use to extract the base form of the word by remov affix from them . it is just like cut down the branch of a tree to it stem . for exampl , the stem of the word eat , eat , eaten is eat . search engin use stem for index the word . lemmat take a word and break it down to it lemma . for exampl , the verb `` walk " might appear as `` walk , " `` walk " or `` walk . "

   Output after Lemmatization

   lemmatization algorithm would know that the word better be derive from the word good , and hence , the lemme be good . but a stem algorithm would not be able to do the same . stemming be a technique use to extract the base form of the word by remove affix from them . it be just like cut down the branch of a tree to it stem . for example , the stem of the word eat , eat , eat be eat . search engine use stem for index the word . lemmatization take a word and break it down to it lemma . for example , the verb " walk " might appear as " walk , " " walk " or " walk . "

2. After replacing X and Y

   11, 75, 13, 73, 72, 45, 15, 15, 15, 16, 19, 20, 21, 22, 24, 30, 40, 45, 71, 20, 20, 21, 23, 6

   i. Equi-depth binning method

   Sort the dataset: 6, 11, 13, 15, 15, 15, 16, 19, 20, 20, 20, 21, 21, 22, 23, 24, 30, 40, 45, 45, 71, 72, 73, 75

   Number of bins: 4

   Bins:

   1. 6, 11, 13, 15, 15, 15
   2. 16, 19, 20, 20, 20, 21
   3. 21, 22, 23, 24, 30, 40
   4. 45, 45, 71, 72, 73, 75

a. Smoothing by Bin Mean

Bin 1 mean: (6 + 11 + 13 + 15 + 15 + 15) / 6 = 12.5

Bin 2 mean: (16 + 19 + 20 + 20 + 20 + 21) / 6 = 19.333 ≈ 19.33

Bin 3 mean: (21 + 22 + 23 + 24 + 30 + 40) / 6 = 26.666 ≈ 26.67

Bin 4 mean: (45 + 45 + 71 + 72 + 73 + 75) / 6 = 63.5

12.5, 12.5, 12.5, 12.5, 12.5, 12.5, 19.33, 19.33, 19.33, 19.33, 19.33, 19.33, 19.33, 26.67, 26.67, 26.67, 26.67, 26.67, 26.67, 63.5, 63.5, 63.5, 63.5, 63.5

b. Smoothing by Bin Boundaries

Bin 1 boundaries: 6 and 15

Bin 2 boundaries: 16 and 21

Bin 3 boundaries: 21 and 40

Bin 4 boundaries: 45 and 75

6, 15, 13, 15, 15, 15, 16, 19, 20, 20, 20, 21, 21, 22, 23, 24, 30, 40, 45, 45, 71, 72, 73, 75

ii. Equal Width binning

Min value: 6, Max value: 75

Width of each bin: (75 - 6) / 4 = 17.25 ≈ 17.25

Bins

1. Bin 1: 6 to 23.25
2. Bin 2: 23.25 to 40.5
3. Bin 3: 40.5 to 57.75
4. Bin 4: 57.75 to 75

Assign data to bins

1. 6, 11, 13, 15, 15, 15, 16, 19, 20, 20, 20, 21, 21, 22, 23
2. 24, 30, 40
3. 45, 45
4. 71, 72, 73, 75

a. Smoothing by Bin Mean

Bin 1 mean: (6 + 11 + 13 + 15 + 15 + 15 + 16 + 19 + 20 + 20 + 20 + 21 + 21 + 22 + 23) / 15 = 17.2

Bin 2 mean: (24 + 30 + 40) / 3 = 31.33

Bin 3 mean: (45 + 45) / 2 = 45

Bin 4 mean: (71 + 72 + 73 + 75) / 4 = 72.75

17.2, 17.2, 17.2, 17.2, 17.2, 17.2, 17.2, 17.2, 17.2, 17.2, 17.2, 17.2, 17.2, 17.2, 17.2, 31.33, 31.33, 31.33, 45, 45, 72.75, 72.75, 72.75, 72.75

b. Smoothing by Bin Boundaries

Bin 1 boundaries: 6 and 23

Bin 2 boundaries: 24 and 40

Bin 3 boundaries: 45 and 45

Bin 4 boundaries: 71 and 75

6, 11, 13, 15, 15, 15, 16, 19, 20, 20, 20, 21, 21, 22, 23, 24, 30, 40, 45, 45, 71, 72, 73, 75

iii. Min-Max Normalization

Min = 6, Max = 75

Normalized Value = (Value - Min) / (Max - Min)

0, 0.072, 0.101, 0.130, 0.130, 0.130, 0.145, 0.188, 0.203, 0.203, 0.203, 0.217, 0.217, 0.232, 0.246, 0.260, 0.348, 0.493, 0.565, 0.565, 0.942, 0.957, 0.971, 1.000