Report

Goal: develop and test text classifications with Naive Bayes algorithm.

Input DataBase

['0', 'The Chinese In Beijing The Chinese']
['0', 'Chinese Chinese In The In Shanghai']
['0', 'The Chinese The In In Macao']
['1', 'In Tokyo In The Japan The In Chinese']

Preprocessing:

Lowercase:
['0', 'the chinese in beijing the chinese']
['0', 'chinese chinese in the in shanghai']
['0', 'the chinese the in in macao']
['1', 'in tokyo in the japan the in chinese']

Spliten DataBase:
['0', 'the', 'chinese', 'in', 'beijing', 'the', 'chinese']
['0', 'chinese', 'chinese', 'in', 'the', 'in', 'shanghai']
['0', 'the', 'chinese', 'the', 'in', 'in', 'macao']
['1', 'in', 'tokyo', 'in', 'the', 'japan', 'the', 'in', 'chinese']

Deleting stopwords:
['0', 'chinese', 'beijing', 'chinese']
['0', 'chinese', 'chinese', 'shanghai']
['0', 'chinese', 'macao']
['1', 'tokyo', 'japan', 'chinese']

“Predict” method:

Test sample:
Chinese The Chinese In Chinese Tokyo The In The Japan

Probability of sample to all classes:
[0.00030121377997263036, 0.00013548070246744226]
With log:
[-8.10769031284391, -8.906681345001262]

Answer of test:
['Chinese The Chinese In Chinese Tokyo The In The Japan', '0']

Inference:
Algorithm work correct.

“Fit” method:

Returned '100%' accuracy

Inference:
Algorithm work correct on the test case.