राष्ट्रीय प्रौद्योगिकी संस्थान रायपुर

NATIONAL INSTITUTE OF TECHNOLOGY RAIPUR

Mid Sem Exam-2022, 7th Semester

B.Tech, Information Technology

Text Mining

Duration: 2 Hours **Max Marks: 30**

Roll No. 19111 8068

Note:-All questions are compulsory.

Q.1. Consider a very small collection C that consists following three documents: d1: "new york times" d2: "new york post" d3: "los angeles times". Given the following query: "new new times" rank the documents of C using TF-IDF method. **(5)**

Q.2. Give the brief summary of lexical analysis for textual data. **(10)**

Q.3. Give Bayes' theorem. Describe Naive Bayes classification for both discrete and continuous valued features. Consider the following training examples of PlayTennis and apply Naive Bayes classification for predicting the class label of new instance *X'=(Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=Strong)*. **(10)**

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Q.4. Make the decision tree for above PlayTennis training example using ID3 decision tree classifier. Show your all workings in making decision tree and predict the class label of new instance *X'=(Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=Strong)* using ID3 classifiers. **(10)**