Department of MACS, NITK Surathkal

MA855: Big data and Analytics

Feb-June 2021

Seminar Schedule and Project allotment

Course Evaluation plan:	**Seminar(20)**	Report: 05 Presentation: 10 Teamwork: 05
Mid sem: 20% End Sem exam : 30% Quiz : 10 % Seminar : 20 % Project : 20%	**Project (20%)**	Report:05 Demonstration:10 Teamwork:05

Group. No.	**Team No**	**Seminar Topic**	**Date**
1	Team 1,	**MongoDB:** Features, Comparison with RDBMS, Data Types, MongoDB Query Language, Installation and Query execution, Practice Examples Exercise: Qn 1 (covid 19 data set)	12 April 2021, 2:00 pm
	Team 7		12 April 2021, 2:30 pm
2	Team 2	**Cassandra:** Apache Cassandra – Introduction, Features of Cassandra, CQL Data Types CQLSH, Keyspaces CRUD (Create, Read, Update, and Delete) Operations, Collections, Using a Counter , Time to Live (TTL) , Alter Commands, Import and Export, Querying System Tables, Installation and Practice Examples Exercise: Qn 2 (Placement data Set)	13 April 2021, 2:00 pm
	Team 8		13 April 2021, 2:30 pm

3	Team 3	MAPREDUCE Programming: Introduction, Mapper, Reducer, Combiner, Partitioner, Searching, Sorting, Compression, Practice examples, MapReduce Use Case: KMeans Algorithm or similar Exercise: Qn 1 (covid 19 data set)	14 April 2021, 2:00 pm
	Team 9		14 April 2021, 2:30 pm
4	Team 4	Hive: Introduction, History of Hive and Recent Releases of Hive, Hive Architecture, Hive Data Types, Hive File Format, Hive Query Language (HQL), RCFile Implementation, User-Defined Function (UDF), Hive Integration and Work Flow Exercise: Qn 2 (Placement data Set)	15 April 2021, 2:00 pm
	Team 10		15 April 2021, 2:30 pm
5	Team 5	**Pig :** Introduction to Pig, Key Features of Pig, The Anatomy of Pig, Pig on Hadoop, Pig Philosophy, Use Case for Pig, ETL Processing, Pig Latin Overview, Usage of Pig , Pig at Yahoo, Pig versus Hive Exercise: Qn 1 (covid 19 data set)	16 April 2021, 2:00 pm
	Team 11		16 April 2021, 2:30 pm
6	Team 6	R programming: Basic data manipulation, Basic plotting, Loops and functions, Basic stats, Advanced data manipulation, Example Project. Exercise: Qn 2 (Placement data Set)	19 April 2021, 2:00 pm
	Team 12		19 April 2021, 2:30 pm

Exercise Question for seminar.

1. Refer Placement_Data_Set.csv file and answer the following questions.

Data Description:
ssc_p – Percentage in Class 10
ssc_b – Board studied in Class 10
hsc_p – Percentage in Class 12
hsc_b – Board studied in Class 12
hsc_s – Stream studied in Class 12
degree_p – Percentage in Degree
degree_t – Type of Degree (Science & Technology / Commerce & Management / Others)
workex – Work experience
etest_p – Percentage in Entrance Test
status – Placement status

a Filter the students from Science and Technology (Sci&Tech) degree (degree_t) having degree percentage (degree_p) ≥ 75. Find out number of such students present?.

b From the above filter students with gender as F. How many observations do you get?

c How will you select a subset of the data of students from Solution (b) with degree_p >= 80? Assign the dataset to p_upper. How many observations does p_upper contain?

d Calculate he median of degree_p of all students? What is the mean value?

Qn 2. Refer covid_19_india.csv file and answer the following questions.

a. Filter the month in which heighest people are get infected to Covid-19 virus?
b. Obtain state in which survival rate is high.
c. Check for state in which death rate is more than 1% .

MA855 Big data and Analytics Mini Project:

Note:1. This project is for the whole class, i.e., all Teams.
2. Implement the task using Python Programming.

Qn.1. The task is to create and analyse Co-Authorship network using Scopus or Web of science databases.

For a selected keyword or a combination of key words (List of keywords are given below), find:

a) Highest cited author and his h-index (from the world)

b) Highest publication author

c) Highest cited authors avg. citations, and the country name.

d) Total number of publications of the highest cited author

e) Total publication in year

f) Total citation per year

g) Author(country) having highest co-authorship with indian authors.

h) Highest cited author from India and the university.

i) Comparative year wise article publication analysis of india, china and usa.

j) Total number of grants given to the field
k) Country wise total number of publication

Qn. 2. Co-Author Relationship Prediction and Citation prediction using known machine learning techniques.

Note:
Co-Author: If Authors A and B said to be co-author, then they written research article together.
Citation: If author A refers author B's journal, then A cites B' s work, and in a graph a directed link between node A to node B indicates the co-authorship relation.

Steps to download data(may vary between the databases):

STEP 1. Go to the homepage of SCOPUS https://www.scopus.com or https://apps.webofknowledge.com/
STEP 2. Type your relevant topic (KEYWORD) in Search box
STEP 3. Click source type- check to JOURNALS.
STEP 4- Document type- ARTICLE
STEP 5- Click Subject Area (i.e Computer Science , Information systems etc)
STEP 6. After all the steps- Click "LIMIT TO" or refine in web of science

STEP 7. find out best articles by click sort based on Number of citations
STEP 8. Click Export button after selecting articles based on method adopted,
Choose CSV (comma separated value) Excel to export the data

STEP 9.From the above data Create new Excel file with required columns to analyze given set of attributes.

List of keywords (put with in double quotes “ ”):

Not exhaustive

1. Wireless Sensor networks
2. biogeography-based optimization
3. block-chain
4. quantum computing
5. game theory
6. internet of things
7. deep learning
8. big data analytics
9. fuzzy logic
10. reinforcement learning
11. India
12. USA
13. China
14. artificial intelligence
15. computer science
16. meta heuristic
17. Evolutionary
18. bio-inspired
19. graph theory
20. graph coloring
21. ill posed problems
22. robotics
23. corona virus
24. cancer

Links:

Web of science and Scopus record:

1.https://www.youtube.com/watch?v=dQgmIcVXqu8&feature=youtu.be

2. http://networksciencebook.com/translations/en/resources/data.html

Instructions for Seminar/Project:

1. The duration of presentation for each team (3 members) is 30 mins. The weightage for project and seminar are as given in table above.
2. The presentation slides should be neatly prepared, which needs to be shared among fellow course mates, which will be included in syllabus
3. The group members should work together to prepare the report and presentation slides to avoid any repetition.

Course Instructor: Dr. Pushparaj Shetty D.