Knowledge Discovery through text mining in the United States Data Science Curriculum

By Hammad Khan
University of North Texas
hammadkhan@my.unt.edu
Universities rush to add data science majors as demand explodes

By Isha Salian | September 5, 2017 | Updated: September 5, 2017 6:00am

3 reasons why data scientist remains the top job in America

Why is the position in such high demand and will it stay there in the years to come?


News

Data Scientist Number 1 Again - Best 50 jobs in America for 2017

February 6, 2017


American Library Association President Maureen Sullivan was not amused by Forbes’ recent designation of the library and information science degree as the worst master’s for a student to pursue.
2030: Librarians In Demand

THE TOP 10 OCCUPATIONS

Below are the occupations classifications most likely to experience increased demand in 2030 out of the 772 tracked by the U.S. government.

1. Preschool, Primary, Secondary, and Special Education School Teachers
2. Animal Care and Service Workers
3. Lawyers, Judges, and Related Workers
4. Postsecondary Teachers
5. Engineers
6. Personal Appearance Workers
7. Social Scientists and Related Workers
8. Counselors, Social Workers, and Other Community and Social Service Specialists
9. Librarians, Curators, and Archivists
10. Entertainers and Performers, Sports and Related Workers

Source: The Future of Skills: Employment in 2030

New Responsibilities

- Data Librarian
- Data Coordinator
- Data Curation Specialist
- Digital Content Management
- Research Data Services
Research Data Management Plan

Data Management elements

- Goals and motivation for managing data
- Creating documentation and metadata, metadata for discovery
- Tracking Data Usage
- Data Formats
- Handling sensitive data
- Backing up data
- Data Management Plan (DMP)
Tasks Related to Data Science

- Data analysis
- Modeling/statistics
- Visualization of analytics and statistics
Previous Research


Previous Research

• Harris-Pierce and Liu (2012) surveyed the web sites of 52 LIS schools in North America to identify data curation courses. Results indicated that although LIS schools are beginning to respond to the demand for data curation professionals, they need to add more such courses to their curriculum and continue to work collaboratively to determine the optimal course objectives and learning outcomes.
More Data Less Insight

Open Government Data
Health Data
Business Data
Method

- Text mining is an algorithmic process of extracting meaningful information from text in an effort to uncover linkages between text objects. Unlike conventional data mining tasks that extract patterns from structured databases, text mining is intended to explore relationship among objects, generally stored in unstructured formats.
Data

• The data for this research was collected from many different universities located in the United States using their campus URL’s that were made available through the website http://www.mastersindatascience.org/schools/. The website offers a comprehensive directory of data science programs being offered in the United States.

• 534 data science programs were recorded in the United States. The programs are being made available at 258 different universities that offer either a certificate, master’s, or Ph.D in the discipline of data science. Total document word count resulted in 15,101 words.
Association Rules

[predictive] --> [analytics] (confidence: 0.667)
[management, information] --> [systems] (confidence: 0.667)
[technology] --> [information] (confidence: 0.676)
[science] --> [data] (confidence: 0.691)
[learning] --> [machine] (confidence: 0.741)
[visualization] --> [data] (confidence: 0.868)
[care] --> [health] (confidence: 0.929)
[courses] --> [core, list] (confidence: 0.936)
[intelligence] --> [business] (confidence: 0.942)
[courses] --> [list] (confidence: 0.945)
[mining] --> [data] (confidence: 0.949)
[core] --> [list] (confidence: 0.965)
[core] --> [courses, list] (confidence: 0.965)
[courses] --> [core] (confidence: 0.966)
[courses, core] --> [list] (confidence: 0.969)
[list] --> [core] (confidence: 0.987)
[list] --> [courses, core] (confidence: 0.987)
[courses, list] --> [core] (confidence: 0.991)
[list] --> [courses] (confidence: 0.996)
[core] --> [courses] (confidence: 0.996)
[big] --> [data] (confidence: 1.000)
[warehousing] --> [data] (confidence: 1.000)
[machine] --> [learning] (confidence: 1.000)
[analytics, big] --> [data] (confidence: 1.000)
[core, list] --> [courses] (confidence: 1.000)
Conclusion

This preliminary research investigated the use of association rules for text mining in the titles of core courses in Data Science programs across universities in the United States.

The research provided text associations of terms that are frequently linked together in core course titles in the discipline of data science.
References


