MASTERS
FINAL ORAL EXAMINATION

Friday, November 18th
11:00a.m. @ 216 Atanasoff

Abhinav Yedla

Major Professors: Pavankumar Aduri and Shawn Dorius

Innovations in Research Methods Series (IRMS)

Data Science, since its inception has evolved as an extremely robust field equipped with multi-disciplinary techniques to extract, analyze and classify structured and unstructured data that give insights into new patterns and knowledge in multiple fields. This emergence has drawn the interest of pioneers and experts of different disciplines especially social scientists towards Big Data. The relatively obscure Data Science literature if simplified would help these pioneers in understanding the essence of Data Science and ease their process of learning new techniques in this field. This work is divided into two parts. The first part accomplishes two goals- Extract data science literature to define, classify and build relationship among the extracted terms and then to classify various techniques of data extraction and score them on the basis of their availability, data quality, ease of extraction, reproducibility, technical skills and Types of Data.

The second part of the work builds open source tools for web data extraction for a few novel purposes like google auto complete, wikipedia, social mentions and passport. Finally the tool for google auto complete text extraction is supported by experimental results. There are around 200 domains for google and we go to each particular domain and obtain auto complete data related to that particular country for specific queries like its geo politics and the opinion in that country about other neighboring countries. This data once extracted is subject to sentiment analysis and then the results are analyzed to know the polarity and emotions.

IOWA STATE UNIVERSITY
Department of Computer Science