Project Instruction
CSE 40647/60647 Data Science

Professor-in-charge:

Dr. Meng Jiang, mjiang2@nd.edu
Office: 326C Cushing Hall Phone: (574) 631-7454

Teaching Assistant (TA):

Qi Li, qli8@nd.edu

Project goal:

For the course project, students will be expected to collect a dataset (online or otherwise), formulate a question of interest, and perform aspects of data science to address that question by using whatever tools they find appropriate. The project will involve a **proposal**, **milestone**, and **final term paper** with **oral presentations** of the project.

Project introduction:

- The students may work in **team of 2 – 4 (minimum 2 members are required)** for the class project.
- The class project may involve some or all stages of the **knowledge discovery** process, depending on the chosen project. All project topics should be preapproved by the professor.
- The class project will require a **proposal and milestone assessment** during the semester with respect to the data science process.
    - **Proposal and milestone** will be presented and evaluated as **on-going term paper** and **oral presentation**.
- The students will be required to write a **term paper** and make a **class presentation** (poster required, oral encouraged) on their project.
    - The **term paper** will go through a **peer review process** among the classmates.
    - The **term paper** must be in PDF format and formatted according to the new Standard ACM Conference Proceedings **Template**.
    - The **term paper** should include **sections** about Introduction, Related Work, Problem Definition, Methodology, Experiments, Discussion, Conclusion and Future Work.
        - There is no page limit.
        - For LaTeX users: unzip acmart.zip, make, and use sample-sigconf.tex as a template; Additional information about formatting and style files is available online at: https://www.acm.org/publications/proceedings-template
        - For Word users: export into PDF format.

Grading policy: (30% of the final score)

Students are required to submit their **data and code package + "readme" (.ZIP) and term paper (.PDF)**.

Students are encouraged to **implement** algorithms such as Apriori, FP-Growth, Decision Trees, Naïve Bayes, SVM, and K-Means Clustering by themselves instead of calling Python packages. Students are also encouraged to **use Python packages** (e.g., Numpy and Scipy) when they use **advanced techniques** (e.g., Neural Networks, word2vec) to address challenging problems.

Graders should have **higher expectations on graduates** than undergraduates – not only on the project results but also on writing (a workshop-quality paper of strong reasoning). Undergraduates will be applied with a uniform grading policy no matter what majors they have.

The project final due (final term paper) is **05/03/2018 (11:59 pm)**. There will be absolutely NO extension!!!

Grading distribution: (100 points)

- **Proposal paper (10 points)**
- **Milestone presentation/paper (15 points)**
- **Final term oral presentation (25 points)**
- **Final term paper (25 points)**
- **Code package and data (25 points)**

The **project proposal** (proposal paper) will be graded as follows:

Title of Project:	5%	What's the title of the project?
Project Plan:	30%	What do you plan to do?
Data Sources:	20%	What data do you plan to use? From where will this data come?
Proposed Evaluation:	30%	How do you plan to evaluate your proposed method? How will you determine whether the method is successful?
Writing Quality:	15%	Clarity of expression (5%), organization (5%), and grammar (5%).

The **project presentation** (milestone presentation, final oral presentation) will be graded as follows:

Introduction:	15%	Provide context. What questions are being addressed?
Solution/Method:	30%	What did you do? Why did you choose this method? What tools and techniques did you use?
Data and Experiments:	10%	What data did you use? Are your experimental methods reliable?
Evaluation and Results:	30%	What evaluation did you do? Do your conclusions match your results?
Presentation Quality:	15%	Clarity of speaking (5%), organization (5%), and visuals (5%).

The **project paper** (milestone paper, final term paper) will be graded as follows:

Introduction:	15%	Provide context and motivation. What questions are being addressed? Why are these questions interesting or important?
Related Work:	10%	What other methods have addressed these or similar questions? How do these methods differ from your method?
Solution/Method:	25%	What did you do? What tools and techniques did you use? Was any innovation attempted?
Data and Experiments:	10%	What data did you use? Are your experimental methods reliable? What preprocessing was done the data?
Evaluation and Results:	25%	Did you properly evaluate your experiments? Did you test for statistical significance? Do your conclusions match your results?
Writing Quality:	15%	Clarity of writing (5%), organization (5%), and grammar (5%).

Academic Dishonesty:

- The CSE and du lac honor code will be strictly followed.
- All assignments are individual unless instructed. You can discuss the assignment at a high level, but you should independently and individually write down the answers and/or the program. The sharing and copying of homework solutions or programs or functions or exams will be considered cheating.
- All the references and sources should be carefully provided and cited.
- Entering Notre Dame you were required to study the on-line edition of the Academic Code of Honor, to pass a quiz on it, and to sign a pledge to abide by it. The full Code and a Student Guide to the Academic code of Honor are available at: http://honorcode.nd.edu.
- Perhaps the most fundamental sentence is the beginning of section IV-B: "The pledge to uphold the Academic Code of Honor includes an understanding that a student's submitted work, graded or ungraded – examinations, draft copies, papers, homework assignments, extra credit work, etc. - must be his or her own."

Schedule:

Date	Lecture#	Topic	Goals
01-16 (T)	1	Introduction	Understand what is data science research Know project grading policy and schedule Start looking for your **teammates** and find them ASAP Start looking for **interesting and doable** topics ASAP
02-06 (T)	7	Proposal: Teaming and proposal	Write down your teammate names in HW1 (due Feb. 6) and proposal paper (due Feb. 5) Submit your proposal paper: • **What is your project topic/research problem?** • **How will you find your dataset?** • **What may be your proposed method?** You will listen to proposals from your classmates. This may help you if you still want to improve your idea.
03-06 (T)	14	QA	In case that you need to discuss about your project and you don't have time to come to office hours, we offer a great chance for you to briefly introduce your idea in class – everybody in the class will be happy to help you! **Keep in mind: In two days, you'll submit your milestone paper and give a presentation.**
03-08 (R)	15	Milestone	Submit your milestone paper: • **Your topic, dataset, and method** • **Milestone progress: Some preliminary results** • **Challenges and proposed solutions** • **Plan for the next two months** You will give milestone presentations in class. Believe me: Audience will help you, not argue with you.
04-26 (R)	27	Oral 1 (up to 20% additional credits)	Every team gives an oral presentation. **Classmates, instructor, and invited faculty will evaluate your presentation.**
05-01 (T)	28	Oral 2	
05-03 (R)			Project final paper due: **You have to submit your code package, data, and term paper at 11:59PM this date.**

Data Portals:

- Kaggle: https://www.kaggle.com/
- DATA.GOV: https://www.data.gov/
- City of Chicago Data Portal: https://data.cityofchicago.org/
- City of South Bend Open Data: http://data-southbend.opendata.arcgis.com/
- Index of Complex Networks: https://icon.colorado.edu/
- The Koblenz Network Collection: http://konect.uni-koblenz.de/
- Stanford Large Network Dataset Collection: http://snap.stanford.edu/data/

Other Resources

Data Sources

KDnuggets Data Repositories List — Data repository list maintained by KDnuggets, a popular data mining website
UCI Datasets — The UC Irvine Machine Learning Repository, a popular source of machine learning datasets
mldata.org — A public repository for machine learning data
Wikipedia Database — Webpage for access to complete Wikipedia database dumps
IMDb Datasets — Webpage for access to IMDb datasets
Last.fm Datasets — Webpage for access to Last.fm datasets
Census.gov — US government source of data about the nation's people and economy
Data.gov — Source of machine readable datasets generated by the US government
UK's Office for National Statistics — Source of datasets generated by the UK's Office for National Statistics
UK's Met Office Data — Climate station records from the UK's National Weather Service
CDC Data — Medical data from the Centers for Disease Control and Prevention
World Bank Catalog — World Bank data
RealClimate Data — Aggregator for selected sources of code and data related to climate science
Google Public Data Explorer — Google's public data portal to explore, visualize, and communicate large datasets
Dataverse Network — Repository for research datasets
Linked Data — Linkage site for distributed data
Datamob — Aggregator for public datasets
Quandl — Search engine for financial, economic, and social datasets
Data Market — Portal for shared business data
CKAN — Open-source data portal platform
Hilary Mason (bitly) Data Links — Hilary Mason's bookmarked research-quality datasets
Peter Skomoroch (LinkedIn) Data Links — Peter Skomoroch's bookmarked machine learning data resources

Jake Hofman Data Links — Jake Hofman's bookmarked computational social science data resources
Reddit Open Data — Forum on the social news site reddit for open APIs and datasets
Guardian DataBlog — Data journalism and data visualization from the Guardian
Free SVG Maps — Website for free geographic maps
StateMaster — Reference site for data on US states
Wolfram|Alpha — Computational knowledge engine or answer engine

Data Visualization Resources
Many Eyes — Web community that connects visualization experts, practitioners, academics, and enthusiasts
Visual Complexity — Resource space for anyone interested in the visualization of complex networks
Thumbs Up Viz — Collection of elegant, efficient, and (above all) effective data visualizations
WTF Visualizations — Visualizations that make no sense

Python
Python.org — The Official Python Website
The Python Tutorial — The Python.org Python tutorial
Learn Python in X Minutes — Whirlwind tour of Python programming
Learn Python the Hard Way — Teaches Python by slowly building and establishing skills through practice and application
Learn Python (interactive) — Engaging Python tutorials
Google's Python Class — Teaches Python via written materials, lecture videos, and lots of code exercises
pyvideo.org — Python-related video index
yhat Data Science in Python Tutorial — Uses IPython to teach data science
Anaconda Python Distribution — Free Python distribution for large-scale data processing and predictive analytics
The Python Package Index — Repository of Python software
pip — Tool for installing and managing Python packages
NumPy — Python package for scientific computing
SciPy Library — Python package for mathematics, science, and engineering
Matplotlib — Python package for 2D plotting
pandas — Python package for high-performance, easy-to-use data structures and data analysis tools
IPython — Architecture for interactive computing with Python
scikit-learn — Python package for machine learning