**DS 785: Capstone Course Project Proposal Outline**

Connie Sosa

**• Project Description**: a summary and overview of the type and topic of your proposed Capstone Project

This is an unpaid client-based project with the [Duke Clinical Research Institute](https://www.dcri.org/) (DCRI). Clinical Trials Transformation Initiative, co-founded by FDA and Duke, developed the Aggregate Analysis of Clinical Trials (AACT) dataset that aggregates the content of ClinicalTrials.gov and makes access of clinical studies information available to the public. I will be exploring the use of automated methods using text mining algorithms to analyze text data from the AACT dataset.

**• Rationale (list of reasons) for undertaking this project.** An explanation of how the project could be useful and provide value in terms of implementation

Traditionally, clinical trial studies are classified manually into MeSH term (Medical Subject Headings). The goal of this project is to increase the efficiency of categorization and find an efficient way to categorize clinical trials into clinical domains.

**• Proposed Project Title**: indicative of content, should look good on resume

Machine Learning Text Mining Methods for Categorizing Clinical Trials by Clinical Domain

**• Proposed Project Purpose:**

**Client-based project: include the client organization description and client contact (name and position) and purpose**

**Client organization:** DCRI, a research group within Duke University Medical Center, is the largest academic clinical research organization in the world. DCRI grew out of the Duke Databank for Cardiovascular Diseases, since its inception in 1969, they’ve evolved into an organization with major efforts in clinical trials, outcomes research, and health policy. The organization performs clinical research across the spectrum of diseases, ranging from Phase I through IV clinical trials. Their mission is to develop and share knowledge that improves the care of patients.

**Client contact:** [Sheri Tibbs](mailto:%20sheri.tibbs@duke.edu), an informaticist at DCRI, will be the main client contact for this project.

**Purpose:** Provide clinical researchers, clinicians, and healthcare partners with improved navigation of clinical domain information that can help improve clinical research and development.

**• Project Objectives:** List and describe 3-5 objectives on what you want to accomplish by undertaking the proposed capstone project. Please be specific.

* Propose and utilize machine learning methods to help facilitate the classification of studies in the pharmaceutical space.
* Present results to the organization and evaluate feedback
* Capture lessons learned and recommend future direction for the organization

**• Project Timeline with Activities:** List and describe activities you will be engaged in to achieve the objectives listed above and a timeline of completion of these activities.

* The project must be divided into 12 (twelve) weekly sections.
* There are 5 (five) written project submissions during the semester and each of these submissions is a separate section of your final product; there is a final draft due one week prior to the last day of the course. These submission due dates should be reflected on your timeline.
* Other timeline notes include peer review interaction and scheduled interviews**.**

Weekly client meeting (in-person and phone every other week) through the end of July.

|  |  |  |  |
| --- | --- | --- | --- |
| week | date | activities | Due date |
| 0  1 | pre-semester  5/30 – 6/03 | Formulate research questions and problem.   * research project ideas, read articles, papers, and publications * meet with client to discuss DS project requirement and project ideas   Read course content and complete the necessary forms |  |
| 2 | 6/04 – 6/10 | 6/08: client meeting (in person)  Identify and contact two+ people to interview  Formulate research questions and problem (finalize)  Select/evaluate tools and prepare project environment   * identify necessary resource, obtain, install, and configure query/analytic tools. * setup computing environment on Amazon AWS * establish AACT connection via R, Tableau, pSQL, pgAdmin | 6/05 project idea submission |
| 3 | 6/11 – 6/17 | 6/15: client meeting (phone)  Test recording equipment  Conduct interview 1  Analyze data constructs and information to refine solution approach   * understand AACT data schema diagram, table definitions, and data dictionary   Collect and manage data to devise solutions   * identify tables, fields, and data needed * obtain a list of clinical domain subcategories from the client | 6/12 project proposal form & timeline |
| 4 | 6/18 – 6/24 | 6/22: client meeting (in person)  Conduct interview 2  Select models and methods to address the research questions   * identify applicable algorithms to use | 6/19 activity update 1 |
| 5 | 6/25 – 7/01 | 6/29: client meeting (phone)  Prepare PowerPoint for interview 1 summary  Apply and evaluate models and methods to address the research questions   * unsupervised machine learning methods, part I | 6/30 interview 1 summary |
| 6 | 7/02 – 7/08 | 7/06: client meeting (in person)  Apply and evaluate models and methods to address the research questions   * unsupervised machine learning methods, part II | 7/03 activity update 2 |
| 7 | 7/09 – 7/15 | 7/13: client meeting (phone)  Apply and evaluate models and methods to address the research questions   * supervised machine learning methods, part I * Interpret and assess models results |  |
| 8 | 7/16 – 7/22 | 7/20: client meeting (in person)  Class interaction & peer review  Apply and evaluate models and methods to address the research questions   * supervised machine learning methods, part II * evaluate the limitations of the results | 7/17 activity update 3  7/19 peer review, class interaction |
| 9 | 7/23 – 7/29 | 7/27: client meeting (phone)  Prepare PowerPoint for interview 2 summary  Finalize the capstone project/product | 7/28 interview 2 summary |
| 10 | 7/30 – 8/05 | Capstone final paper (draft version)  Schedule client meeting (for August) to present the final project/ product | 7/31 activity update 4 |
| 11 | 8/06 – 8/12 | Capstone final paper (draft version)  Document source code | 8/11  **final draft** |
| 12 | 8/13 – 8/18 | Capstone final paper (final version)  Upload developed source code and document to Git version control repository | 8/14 activity update 5  8/18 **final capstone project** |

**• Interviewees.** Identify two people you plan to interview to fulfill the capstone project requirements. When possible, the interviewees should be related to your project, but you can also interview individuals with knowledge or experience related to the project’s content or your career goals.

1. Sheri Tibbs is an informaticist at the Duke Clinical Research Institute. She has been in the health care related field for many years. Her specialty includes healthcare information technology, public health, data analysis, informatics, healthcare management, and research.
2. Mr. Fisher works for a scientific research and engineering company. Although he will not be working with me on the capstone project, he is someone who has worked with data science related technologies. As part of the data science related projects that he has worked in, he has worked in his company’s decision support team developing statistical based Machine Learning prediction tools for the US Army Institute of Surgical Research.

**• Application of Data Science Concepts** Provide insight into the data science topics, theories, and concepts that will be applied. Discuss knowledge that you expect to use in doing the project, including specifying skills and methodologies pulled from previous coursework

*Statistical methods:* Apply statistical procedures, implement statistical analysis, and interpret results.

*Programming for DS:* Use R as the programming language to analyze clinical trials data. Follow best practices for documenting code. Use Git for source code version control.

*Data Warehousing:* Clean and prepare data, use PostgreSQL, utilize AACT database schema, data dictionary, and data definition of AACT.

*Big Data - High performance computing:* Although this project doesn’t apply distributed processing across computer clusters, but it will rely on accessing AACT database hosted in the AWS cloud.

*Communicating About Data:* Apply technical, informational, and/or persuasive communication techniques learned, and communicate the results of the analysis.

*Data Mining:* Explore supervised and unsupervised data mining methods and procedures, specifically, classification and clustering algorithms. Model validation for supervised machine learning methods will also be applied.

*Visualization and Unstructured Data Analysis:* Utilize text mining methods for analyzing text data of clinical studies stored in AACT. Latent Dirichlet Allocation (LDA), among other text mining algorithms will be explored. Visualization techniques will also be applied.

*Prescriptive Analytics:* Optimization problems lie at the heart of many machine learning methods. LDA, topic modeling algorithm, is a probabilistic analysis of text data that guides the keyword selection and content optimization strategies. Gibbs Sampling method, a form of Markov chain Monte Carlo (MCMC) algorithm, is a sampling based approximation algorithm used for fitting the LDA model.

**• Description of Final Document.** This description must reflect a project that takes up the full academic semester.

Major sections of the final document will contain the following.

Abstract: The abstract will be a summary of the capstone project that describes the content, scope, and objective of the project. It will also include an outline of the methods used and the result of the findings.

Introduction: The introduction will outline the current method of categorizing clinical studies. It will cover the explanation of the MeSH terminology in the AACT dataset and the MeSH condition terms hierarchy. Additionally, the AACT data source and data model will be described.

Methodology: The proposed methodology for categorization including the process of grouping clinical studies in AACT by clinical domain will be presented. Rationale will be given for the selection of supervised learning methods and/or unsupervised learning methods.

Results and Conclusions: If more than one type of algorithm or learning approach is explored then comparisons of the different approaches will be presented. Where possible an automated approach for quality assessment of clinical domain classification will be used, such as in the case of using a validation dataset. For other approaches a manual study classification may be utilized.

Summary and Discussion: Share the summary of findings and recommend potential future work.

The final document will also include the table of contents, references, supporting information such as tables, figures, diagrams, and/or appendix.

**References:**

1. Department of Medicine, Duke University School of Medicine:

<https://medicine.duke.edu/divisions/cardiology/research/clinical-research/duke-clinical-research-institute>

1. Visualization in Data Science at IEEE VIS 2016: <http://www.visualdatascience.org/2016/program/>
2. Probabilistic Models for Text Mining: <https://pdfs.semanticscholar.org/c6ee/67dfa4502e6f50de823f563563151f661f89.pdf>
3. The Interplay of Optimization and Machine Learning Research (Journal of Machine Learning Research):

<http://www.jmlr.org/papers/volume7/MLOPT-intro06a/MLOPT-intro06a.pdf>