**Project Proposal**

1. Name, Contact info (e.g. email/phone).

   Kate Gallagher
   Regis email: kgallagher@regis.edu
   Personal email: mail@katherinemgallagher.com
   Phone number: 646 285 3325

2. Title of the project

   **An ETL-Driven NLP Analysis for Crafting Actionable Insights**

3. High level description of the project: what question or problem are you addressing?

   Question: How can we use customer review texts, which are rich in information but disorganized and difficult to analyze, to gain actionable insights into the positive and negative aspects of a product, and subsequently improve customer satisfaction and brand equity?

   I want to create a customer review analysis process in order to provide deeper insight into online product reviews. This would involve creating an ETL pipeline to acquire the data via web scraping, and applying NLP, keyword extraction, and sentiment analysis techniques to the data in order to identify common themes and issues in the reviews. Finally I will build an interactive integrated dashboard to visualize sentiment distribution, trends over time, and key insights.

   I envision this to be a 16-week project in totality, covering the span of my Practicum II project as well.

4. What type of data science task is it?

   ETL process (Extract, Transform, Load)
   Web Scraping
   Data Processing
   Natural Language Processing
   Database System Design, Implementation and Management
   Data Visualization

5. Data: Brief description of data. How big do you expect the data will be? Is amount of your data too big or too small? If you're web-scraping or collecting data, how long do you expect to collect the data?

   The data will consist of the text of customer reviews scraped from three to five webpages using a web-scraping program. Because I am keeping the number of products small, I do not anticipate the dataset to be too large. I plan to spend up to two weeks establishing and implementing a web-scraping program to acquire the data before moving on to data cleaning, transformation and analysis.

6. How will you analyze the data? What machine learning methods do you plan to use, and/or what business intelligence aspect do you plan on incorporating?

   Once I establish my ETL pipeline and extract my data, I plan to implement NLP analysis techniques such as sentiment analysis, named entity recognition, topic modeling, keyword extraction, and frequency analysis. I will also design an interactive dashboard in Tableau to incorporate a business intelligence aspect.

7. Describe any anticipated difficulties and problems. Discuss how you may overcome the problems.
   - Timeline: as discussed, I will be unable to do onscreen work for 1 week near the beginning of the project. I plan to mitigate this by ensuring my project is approved and in progress before that week begins, and creating a project schedule and adhering to it as closely as possible, so that I don't fall too far behind.
   - I am unfamiliar with some of the libraries, packages, and tools I plan to use, which may mean that I need some trial and error before I understand how to use them

8. Suggest a timeline for the project. This should be a weekly breakdown of what you plan on doing each week.

| Week | Week Plan | Tasks |
|---|---|---|
| 1 | Project Planning and Kickoff | - Define project goals, objectives, success criteria<br>- Identify target review pages<br>- Plan architecture of ETL pipeline and database |
| 2 | Data Collection and Database Setup | - Set up web scraping to collect customer review data<br>- Create script to fetch and store raw text in database<br>- Set up database system |
| 3 | Data Preprocessing | - Clean and preprocess raw text data (e.g. stripping and lowercasing, html tag removal, tokenization)<br>- Implement stopword removal and lemmatization |
| 4 | Feature Extraction and Loading | - Extract relevant features (keywords, phrases etc.) from text<br>- Load preprocessed data into database |
| 5 | Sentiment Analysis | - Determine sentiment analysis technique (e.g. VADER, TextBlob) and prepare models |
| 6 | Sentiment Analysis | - Apply sentiment analysis to preprocessed review data<br>- Explore results with EDA |
| 7 | Dashboard Prototyping | - Develop dashboard prototype with current sentiment analysis results and placeholders for future analysis |
| 8 | Presentation Preparation | - Create presentation materials<br>- Deliver presentation |

   (see Appendix I for potential practicum II project plan)

9. Create GitHub repository for your Practicum project. Add this proposal, begin a ReadMe document, and begin adding your data to your repository. Add a link to your GitHub repository to this document.

   https://github.com/kmg-regis/kmg_regis_msde_practicum_1

## **Appendix I:** Possible Practicum II project outline

| Week | Week Plan | Tasks |
|---|---|---|
| 9 | Issue Identification | - Implement techniques to identify common issues in reviews (e.g. keyword extraction, frequency analysis) |
| 10 | Topic Modeling | - Apply topic modeling techniques to identify common review topics (e.g. LDA) |
| 11 | Aspect-Based Sentiment Analysis | - Apply aspect-based sentiment analysis to identify sentiments toward specific aspects of the product (e.g. battery life, user interface) |
| 12 | Named Entity Recognition | - Implement NER techniques to identify key entities in reviews (e.g. NLTK)<br>- Integrate Part-of-Speech tagging to improve accuracy |
| 13 | Dashboard Prototyping | - Further develop dashboard with NLP results<br>- Integrate dashboard with database for real-time updates |
| 14 | Integration Testing and Database Validation | - Perform integration testing on entire ETL pipeline<br>- Validate database entries and confirm data integrity |
| 15 | Finalization and Cleanup | - Address any remaining project issues<br>- Finalize project documentation<br>- Summarize results of analysis |
| 16 | Presentation Preparation | - Create presentation materials<br>- Deliver presentation |