Topic Modeling for Online Product Review

maka.AI

BREAK THROUGH TECH

August 2024

More about maka.AI

Developing an AI driven learning software transforming data science and AI/ML education for college students and young professionals.

We’re excited to be your Challenge Advisors!

Amy Wang
maka.AI
Advisor
chenguangamywang@gmail.com

- One times entrepreneurs in the U.S.
- 9 years of experience in data science, tech AI/ML, and business intelligence in banking and e-commerce industries.
- Ph.D in Economics of Michigan State University, and MS in Economics at University of Tennessee

Richard Wang
maka.AI
Founder/CEO
Richard2024a@1maka.ai

- Two times entrepreneurs in the U.S.
- 8 years of experience supporting public school districts
- 15 years of experience in data science, tech AI/ML, and business management
- Visiting Fellow of MIT Sloan School of Management, MBA of University of Maryland, and MS in Computer Science at Southeastern University

Maka AI Company overview

- maka.AI — It’s innovated in MIT IHQ in June, 2023. It’s your gateway to the extraordinary world of computing science. Here, we revolutionize computing literacy education and cultivate leaders for the future computing system with a hint of quirkiness, a truckload of innovation, and heaps of AI tech!

One Broadway, Cambridge, MA 02142

AI Studio Challenge Project Overview

CHALLENGE SUMMARY

The goal is to build a topic model with a Large Language Model (LLM). Product reviews are textual, which can’t be used by traditional machine learning approaches. LLM which converts textual data into numeric representations, can be used to model a series of unstructured data.

YOUR TEAM’S OBJECTIVE

- Convert product reviews to LLM embeddings.
- Cluster product reviews based on their corresponding LLM embeddings.
- Generate labels for each cluster.

DESIRED OUTCOMES

Students gain industry experience of leveraging LLM to interpret contextual information such as product reviews

Business context

- Topic Modeling for Online Product Review Industry: Businesses in retail, hospitality, banking, and other consumer-service industries that collect online feedback on services and products, such as Amazon, Target, Expedia, Marriott, and Bank of America Machine Learning Problem Type: Clustering with Large Language Model (LLM)
- Context / Impact: The study of online product reviews provides a wealth of information that can be leveraged across various aspects of retail business operations. The product reviews reveal consumer’s preference on product attributes, helping the e-commerce in competitive analysis, product development, and design marketing strategies. Moreover, product reviews influence other customer’s purchase decisions, therefore, it is important for the retailer to learn insights to earn customer trust and loyalty and improve quality control.

Suggested ML approach

Topic Modeling

- Unsupervised learning, clustering, text generation
- Natural language processing (NLP), transformers, encoder, decoder
- Semantic similarity

Large Language Models

- **Pre-trained Models**
  Start with pre-trained models (e.g., GPT, BERT).
- **Preprocessing**
  Perform text preprocessing (e.g., tokenization, normalization).
- **Fine-Tuning**
  Fine-tune the model on your specific dataset.
- **Evaluation**
  Use metrics like BLEU, ROUGE, and human evaluation.
- **Deployment**
  Plan for integration and scalability.

Data overview

Online Product Review

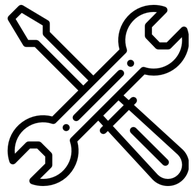
- The Amazon online product review dataset is 18MB, which includes 1,597 product review and 27 columns - id, asins, brand, date added, manufacturer, name, prices, review date, review rating, review text and etc.
- Pre-processing: NLP stopwords removal, lemmatization
- Data location: https://www.kaggle.com/datasets/yasserh/amazon-product-reviews-dataset

Python libraries

- pandas: data frame
- seaborn: plot
- scikit-learn: KMeans clustering
- sentence_transformer: sentence embeddings
- BERTopic: topic modeling

Suggest Dev and PM tools

- development tools: VS Code, Google Colab
- github repo: https://github.com/ichscheine/MIT_AI_Studio.git
- project management tools: Jira, Asana, Confluence
- Agile, Kanban, task delegation

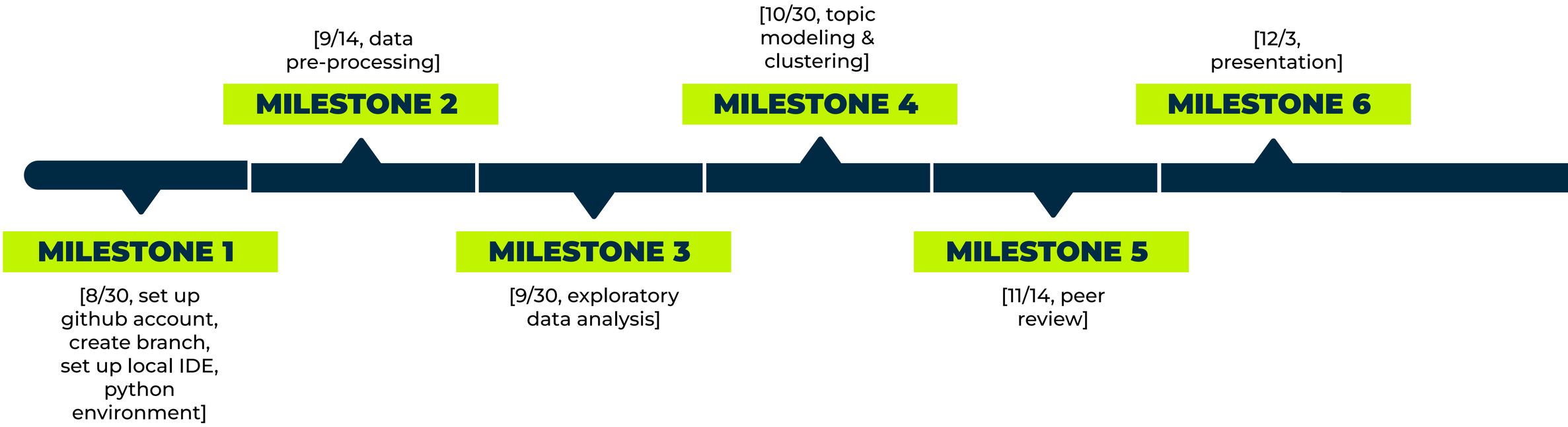


Helpful resources

- Stack Overflow
- Kaggle
- Medium
- Huggingface
- Blind

Project milestones and timeline

These are the milestones for your Challenge Project. They are roughly aligned to the CRISP-DM process you learned about in your ML Foundations course.



MILESTONE #1 TASKS

Set Up Github Account / Local Work Environment

- create a github account with your email address
- create a branch at https://github.com/ichscheine/MIT_AI_Studio.git
- download and install anaconda / miniconda on your computer
- download and install VS code / PyCharm / Intellij
- clone your remote branch to local repo
- set up work environment for local repo (conda create -n <topic_modeling> python=3.12)
- learn to use git for version control (git pull, git status, git commit, git push)

MILESTONE #2 TASKS

Data Preprocessing

- **Data cleaning**: handling missing values, outliers, duplicates, stopwords removal, lemmatization
- **Data transformation**: normalization and encoding
- **Data Reduction**: dimensionality reduction, e.g., PCA
- Suggested tools and libraries for this stage of the project (e.g., Pandas, NumPy, re, Scikit-learn)

MILESTONE #3 TASKS

Exploratory Data Analysis (EDA)

- **Data description**: number of columns and rows, type
- **Data distribution**: mean, variance, freq, mode
- Suggested tools and libraries for this stage of the project (e.g., Pandas, NumPy, re, Scikit-learn, Matplotlib, Seaborn, re, wordcloud, BERTopic, sentence-transformers)]

MILESTONE #4 TASKS

Topic Modeling / Clustering / Sentiment Analysis

- BERTopic: bert embedding, dimension reduction, clustersing, and topic representation (tf-idf).
- Clustering: K-means, fast clustering.
- Sentiment analysis.

MILESTONE #5 TASKS

Peer Review / Code Review

- Readability
- Efficiency
- Coverage
- Security

MILESTONE #6 TASKS

Final Presentation

- Introduction: problem statement, goals
- Literature review: previous works
- Data: pre-processing, EDA
- Methodology: topic modeling, clustering
- Results: e.g., what features should be improved for which products
- Recommendations / future work

How we'll work together this semester

Check-in meetings	• Milestones are arranged to match meeting schedules (2-week Sprint) • Try to follow the provided notebook and reference resources • Try to integrate what you learned from previous classes • I will give feedback at each milestone
Reporting	• 2-week Sprint
Communication	• Email: chenguang.amy.wang@gmail.com, expect reply in 48 hours
Tools and platforms	• Coding platform: Google Colab, GitHub • IDE: VS Code, PyCharm, Intellij • Project management: Jira, Asana
Other project norms	• Sprint

How to get started

Here’s what I suggest for your immediate next steps. I’ll follow up on your progress and help address any challenges in our next check-in meeting:

Review these slides and note down questions

I’ll email you a copy of this deck to store in your Google Drive project folder. Review it as a team and note down any questions you’d like to discuss in our next meeting.

Complete your “Project Scope and Deliverables”

As a team, work on this required Break Through Tech assignment before and during your Sept 7th Maker Day. We’ll review it in our next meeting.

[TBD next step]

Set up work environment and browse the dataset

Questions?

What questions do you have?

Anything I can help clarify?

What are you most excited about?

Anything you’re unsure about?

BEFORE YOU BEGIN

Guidance for filling in this template

You'll be meeting virtually with your assigned student team on or around **Wednesday, August 14th**. This template deck is what **we ask all Challenge Advisors to fill in and be ready to present during this first meeting.** There are notes below each slide to help guide you. Feel free to make this deck your own (e.g., change slide titles, add slides). The resources below will help ensure that your Challenge Project meets program requirements and is a good fit for our students. Please review them before beginning to populate this template. It's ok and encouraged to rescope your Challenge Project (vs. what you originally submitted to us) for the sake of alignment with the project guidelines, dataset guidance, etc., linked below.

THANK YOU!

AI Studio Challenge Project Guidelines
Contains essential guidelines related to scoping your Challenge Project to meet the requirements for inclusion in AI Studio.

Dataset Guidance for AI Studio Challenge Projects
Specifies when and how to provide your project data, what the dataset should look like, and other essential guidance.

Machine Learning Foundations Course Outline
Will help you understand what the students learned in their 9-week eCornell summer course. It can help you "level" your Challenge Project appropriately.

Review your original project submission
You can use what you submitted to us in the Challenge Project submission form as the foundation for this Project Overview deck. However, please make sure to add in the additional information requested in this template. And remember that you can and should rescope if your original submission is not well-aligned with the guidance shared above. If you do not have a copy of your original project submission and would like to review it, please reach out and we can email you a reference copy.

Advice from former students: How to scope out and present your project

- Give us **structure, clear goals, and a timeline** to work toward at the outset of the project. Open-ended prompts can be hard for us since we're used to college course assignments.
- Provide **guidance on the initial steps** for us to take. Don't assume we know what to do, even if it's obvious to you! Some of us are new to Python and data science.
- Provide us with a **useable dataset** that (a) includes proper documentation and (b) doesn't require so much cleaning or pre-processing that it prevents us from getting started.
- **Help us anticipate potential challenges** during later phases of the project (and, when the time comes, help us tackle them by pointing us in the right direction).
- Suggest **resources** for us to better understand the problem space and possible approaches. What would **you** have found helpful as an undergrad?