# SoftDes Midterm Proposal Spring 2025

In the project proposal, we’ll ask you and your team to provide key information about your project. Your answers to the proposal questions serve several important purposes:

* Telling us what your project is about.
* Justifying why you and your team are well-suited to carry out this project.
* Showing us that you’ve done a few tasks to set up and plan your project.

## Project Overview

In one sentence, what is your project about?

**Analyzing trends in positive and negative reviews for restaurants.**

What are the research questions you hope to answer by doing this project?

* **How do positive and negative reviews vary in their speech style?**
* **Does the diversity in language (e.g. how many words were used, repetition of words, number of words) used correlate to the star rating the customer gave?**
* **Do people get more extreme (as in how strong their word selections are like using all capital letters in a sentence or cursing or using otherwise derogatory terms) in the 1 or 5 star reviews?**

## Repository

Please make sure that you have created a new repository from the template repository, https://github.com/olincollege/python-project-template. Make sure that the repository fulfills the following criteria:

* The repository is hosted in olincollege and not in a personal account.
* The name of the repository is properly formatted: it does not contain underscores (\_), the name of the year/semester, team members, instructors/CAs, course, or “midterm project”.
* All team members are collaborators on the repository.
* All team members have cloned the repository to their machines.
* The softdes-2025-01-assistant and softdes-2025-01-teaching teams have permissions on your repository.

## Data Retrieval

Data overview: In a few sentences, what data are you collecting? (This should be understandable by someone who doesn’t know Python.)

**We are collecting reviews and ratings for restaurants. Specifically, we are collecting the text for 1 and 5 star reviews and the rating for that review (so if we got a 5 star rating review, we would know it was from a 5 star rating).**

Justification: How will this data help you answer the questions above?

**This is the data we are going to analyze to extrapolate trends.**

Data Source: Where are you getting your data from?

**Trip Advisor Website**

Data retrieval method: Is your team retrieving data through API calls, Web scraping, or a Python library? (Then, answer the relevant detailed questions below, depending on your method)

**API Calls**

### API Calls

API key preparation: Please be prepared to show your API key (we’ll review this with you at the check-in; **do not include it here or commit this to your GitHub repository under any circumstances**).

Sample code: Please be prepared to show a Python function that uses that key to download some of your desired data (we’ll review this with you at the check-in)

Description: Please provide samples of **all** data you plan to retrieve (at a minimum, sample output from every API endpoint you plan to use), **OR** data types of **all** data you plan to retrieve, by field/column (e.g., saying “JSON” or “data frame” is not specific enough)

**From API Call:**

**Note: Values and Objects not used are omitted.**

**data [type: array]**

**Id [type: int32] - The Tripadvisor ID for the review.**

**Rating [type: int32] - Overall rating for this POI. Not applicable to geographic locations. Rating levels are defined as follows: 5: Excellent 4: Very good 3: Average 2: Poor 1: Terrible**

**Text [type: string] - The full text of the review.**

**Title [type: string] - The title of this review.**

**is\_machine\_translated [type: Boolean] - True or false depending on whether this is a machine-translated review. (Outputs only if partner configured for inclusion of machine-translated reviews)**

### Web Scraping

Pages: Please provide a list of the webpages that you plan to scrape from (if you are going to replace one part of the URL with different values, like page=1, page=2, etc., then feel free to describe those - as long as we know what the actual URL for one of those pages would be).

**https://api.content.tripadvisor.com/api/v1/location/[Location ID]/reviews?key=[API Key]&language=[Language]&limit=[Number of Results]**

**where:**

* **[Location ID] will be replaced by Location IDs that we will find.**
* **[API Key] will be replaced by an API Key that we will acquire.**
* **[Language] will be replaced by “en” to search for only English results.**
* **[Number of Results] will be replaced by an integer specifying the number of results we want. This is likely to be a big number.**

Webpage sample: Please provide a sample webpage downloaded with wget or curl: you can do this with wget <URL> or curl <URL> --output <file>, with <URL> being replaced by the URL you want to scrape and <file> being replaced by the file you want to download the webpage to (wget will choose the file name for you automatically).

**Webpage: "https://api.content.tripadvisor.com/api/v1/location/[Location ID]/reviews?key=[API Key]d&language=en&limit=2"**

**{**

**"data": [**

**{**

**"id": 996881670,**

**"lang": "en",**

**"location\_id": 111418,**

**"published\_date": "2025-03-06T10:05:18Z",**

**"rating": 5,**

**"helpful\_votes": 0,**

**"rating\_image\_url": "**[**https://www.tripadvisor.com/img/cdsi/img2/ratings/traveler/s5.0-66827-5.svg**](https://www.tripadvisor.com/img/cdsi/img2/ratings/traveler/s5.0-66827-5.svg)**",**

**"url": "**[**https://www.tripadvisor.com/ShowUserReviews-g60745-d111418-r996881670-Reviews-Hilton\_Boston\_Park\_Plaza-Boston\_Massachusetts.html?m=66827#review996881670**](https://www.tripadvisor.com/ShowUserReviews-g60745-d111418-r996881670-Reviews-Hilton_Boston_Park_Plaza-Boston_Massachusetts.html?m=66827#review996881670)**",**

**"text": "Overall a wonderful place! Richard and Diego were SO accommodating and kind at the front desk! I travel a lot and they were some of the best front desks folks i’ve had at a hotel! They made the experience, along with the beautiful property and comfortable room.",**

**"title": "Richard and Diego make the experience along with the beautiful property",**

**"trip\_type": "Business",**

**"travel\_date": "2025-03-31",**

**"user": {**

**"username": "zm989",**

**"user\_location": {**

**"id": "50207",**

**"name": "Cleveland, Ohio"**

**}**

**},**

**"subratings": {}**

**},**

**{**

**"id": 996880167,**

**"lang": "en",**

**"location\_id": 17452622,**

**"published\_date": "2025-03-06T09:51:13Z",**

**"rating": 5,**

**"helpful\_votes": 0,**

**"rating\_image\_url": "**[**https://www.tripadvisor.com/img/cdsi/img2/ratings/traveler/s5.0-66827-5.svg**](https://www.tripadvisor.com/img/cdsi/img2/ratings/traveler/s5.0-66827-5.svg)**",**

**"url": "**[**https://www.tripadvisor.com/ShowUserReviews-g60745-d17452622-r996880167-Reviews-CitizenM\_Boston\_North\_Station\_Hotel-Boston\_Massachusetts.html?m=66827#review996880167**](https://www.tripadvisor.com/ShowUserReviews-g60745-d17452622-r996880167-Reviews-CitizenM_Boston_North_Station_Hotel-Boston_Massachusetts.html?m=66827#review996880167)**",**

**"text": "This was a first time for me so I was a little surprised when I saw the size of the room. Ten minutes into my stay the size didn't matter becuase the efficicancy was terrific. The ammenities outweighed the real estate. I would definitly revisit.\n",**

**"title": "Jay Lee",**

**"trip\_type": "NONE",**

**"travel\_date": "2025-03-31",**

**"user": {**

**"username": "jklee2025",**

**"user\_location": {**

**"id": "41817",**

**"name": "Somerset, Massachusetts"**

**},**

**"avatar": {**

**"thumbnail": "**[**https://media-cdn.tripadvisor.com/media/photo-t/1a/f6/f3/e4/default-avatar-2020-30.jpg**](https://media-cdn.tripadvisor.com/media/photo-t/1a/f6/f3/e4/default-avatar-2020-30.jpg)**",**

**"small": "**[**https://media-cdn.tripadvisor.com/media/photo-l/1a/f6/f3/e4/default-avatar-2020-30.jpg**](https://media-cdn.tripadvisor.com/media/photo-l/1a/f6/f3/e4/default-avatar-2020-30.jpg)**",**

**"medium": "**[**https://media-cdn.tripadvisor.com/media/photo-f/1a/f6/f3/e4/default-avatar-2020-30.jpg**](https://media-cdn.tripadvisor.com/media/photo-f/1a/f6/f3/e4/default-avatar-2020-30.jpg)**",**

**"large": "**[**https://media-cdn.tripadvisor.com/media/photo-p/1a/f6/f3/e4/default-avatar-2020-30.jpg**](https://media-cdn.tripadvisor.com/media/photo-p/1a/f6/f3/e4/default-avatar-2020-30.jpg)**",**

**"original": "**[**https://media-cdn.tripadvisor.com/media/photo-o/1a/f6/f3/e4/default-avatar-2020-30.jpg**](https://media-cdn.tripadvisor.com/media/photo-o/1a/f6/f3/e4/default-avatar-2020-30.jpg)**"**

**}**

**},**

**"subratings": {**

**"0": {**

**"name": "RATE\_VALUE",**

**"rating\_image\_url": "**[**https://www.tripadvisor.com/img/cdsi/img2/ratings/traveler/s5.0-66827-5.svg**](https://www.tripadvisor.com/img/cdsi/img2/ratings/traveler/s5.0-66827-5.svg)**",**

**"value": 5,**

**"localized\_name": "Value"**

**},**

**"1": {**

**"name": "RATE\_ROOM",**

**"rating\_image\_url": "**[**https://www.tripadvisor.com/img/cdsi/img2/ratings/traveler/s4.0-66827-5.svg**](https://www.tripadvisor.com/img/cdsi/img2/ratings/traveler/s4.0-66827-5.svg)**",**

**"value": 4,**

**"localized\_name": "Rooms"**

**},**

**"2": {**

**"name": "RATE\_LOCATION",**

**"rating\_image\_url": "**[**https://www.tripadvisor.com/img/cdsi/img2/ratings/traveler/s5.0-66827-5.svg**](https://www.tripadvisor.com/img/cdsi/img2/ratings/traveler/s5.0-66827-5.svg)**",**

**"value": 5,**

**"localized\_name": "Location"**

**},**

**"3": {**

**"name": "RATE\_CLEANLINESS",**

**"rating\_image\_url": "**[**https://www.tripadvisor.com/img/cdsi/img2/ratings/traveler/s5.0-66827-5.svg**](https://www.tripadvisor.com/img/cdsi/img2/ratings/traveler/s5.0-66827-5.svg)**",**

**"value": 5,**

**"localized\_name": "Cleanliness"**

**},**

**"4": {**

**"name": "RATE\_SLEEP",**

**"rating\_image\_url": "**[**https://www.tripadvisor.com/img/cdsi/img2/ratings/traveler/s5.0-66827-5.svg**](https://www.tripadvisor.com/img/cdsi/img2/ratings/traveler/s5.0-66827-5.svg)**",**

**"value": 5,**

**"localized\_name": "Sleep Quality"**

**}**

**}**

**}**

**]**

**}**

Data sample: Please provide a sample of the data you would extract from the webpage and where it is located in the page.

**See section for API Calls.**

### Python Library

Library: What library do you plan to use?

**Pandas and Requests**

API key preparation: If the library uses an API key, please be prepared to show your API key (we’ll review this with you at the check-in; **do not include it here or commit this to your GitHub repository under any circumstances**).

Sample code: Please be prepared to show a Python function that uses that key to download some of your desired data (we’ll review this with you at the check-in)

Description: Please provide samples of **all** data you plan to retrieve (at a minimum, sample output from every API endpoint you plan to use), **OR** data types of **all** data you plan to retrieve, by field/column (e.g., saying “JSON” or “data frame” is not specific enough)

**See section for API Calls.**

## Analysis and Visualization

Data processing: what computations will you do with your data that will NOT appear in visualizations? (e.g., taking the average of something, filtering out invalid values or outliers)

**We will be sorting through all the responses to get only the review ID, Review Rating, Review Title, Review Text, and bool stating if the review was translated by a machine or not.**

Visualization descriptions: What data, or things computed from your data will you show in visualizations?

**Word cloud for most common words in 1 star and 5 star reviews (2 figures);**

Sample visualizations: Please provide examples of your three visualizations below.

A close up of words

AI-generated content may be incorrect.

**Figures 1 and 2 would be like the one above. The third figure would just be a bar chart.**

Narrative: What points will these visuals make?

**The visualizations will show which words are most used for both 1 and 5 star reviews. The third visualization will show the top 10 words across both 1 and 5 star reviews and it will compare how prevalent the words are in 1 star reviews vs. 5 star reviews.**

Justification: How will this visualization help answer your research questions in ways that the data by itself will not?

**The visualization will compare both how prevalent the words are in the two rating sets and which words are most commonly used in each set.**

## Planning

Collaboration style: How are you dividing the work for this project? (Will you write code together, or split the work up and merge your parts later?)

**We will split up the work and merge the parts later.**

Task allocation: Who is doing what between now and the next deadline?

**Mira will work on scraping data from TripAdvisor while Henry will be working on the code to process the data that was scraped from the internet to extract the text and ratings and other information listed above.**

Backup plans: How will you change your plan for the project if you are not able to do all that you had planned?

**If we are not able to scrape data from TripAdvisor, we will pivot to scraping data from Wikipedia and create a question that we can feasibly answer quickly.**