Group Project

BAIS:6100 Text Analytics
Spring 2021

This is a group project. Please form a group of 2–3 people (no more than 3 and no fewer than 2) to complete the project. There will be an online spreadsheet that helps you actively search for teammates.

Summary of Deliverables and Deadlines

What?	When?	Where?
Project proposal (PDF)	6:00 PM on Thu, Mar 18	ICON Assignment
Project deliverables: three Jupyter notebook files, data file(s)	6:00 PM on Thu, Apr 8	IDAS
Presentation	6:00 PM on Thu, Apr 8	in class

There are three major components in the group project:

- Scraping of web articles
- Developing a sentiment classifier
- Performing a cluster and topic analysis

Note that the last two components are based on the data collected through the first component. Note also that everything should be done on IDAS. There will be a shared group folder on IDAS that can be used for your group project.

Scraping of Web Articles

The goal of this component is to experience large-scale web scraping. Choose a target web site that provides the web articles you are interested in. The types of web articles include news articles, blog articles, or whatever web documents you can call articles, excluding comments. Here, we assume that each web page corresponds to only one article. When choosing the target web site, it is recommended, but not required, to choose the one that is expected to have articles with many sentiments, rather than full of just neutral facts.

Scrape at least 3,000 articles from the web site. In order to scrape such a large number of web pages from a web site, you should first examine if there are any pages that list those many articles, e.g., the All Articles pages or the Featured Articles pages. You are not allowed to use any API that the web site may provide. Just use web scraping. Be careful not to overload the target web site. Each article page should be saved as an HTML file. Once you have collected all articles, parse the HTML code of each article and extract the following information from the article:

- Article title (required)
- Article body text (required)
- Posting date and time (if indicated)
- Author (if indicated)

All the extracted information from all the articles should be saved in a single CSV file, in which a row corresponds to an article.

Deliverables: +3,000 HTML files, a CSV file, a Jupyter notebook with code and output

Sentiment Classifier

The goal of this component is to learn the importance of labeled data in supervised machine learning and experience a classifier development process using the data you have collected yourselves. First, tokenize the body text in all the articles you have collected into sentences and merge all the identified sentences. Here, which article each sentence is part of does not matter. Second, create a random sample of 2,000 sentences out of all the sentences. Third, reading the 2,000 sentences one by one, label each sentence with one of the following three sentiments: positive, negative, and neutral. What you will have to decide includes, but not limited to,

- How to handle sentences with multiple sentiments
- How to handle sentences with sarcasm
- How to handle sentences with no clear sentiment or you cannot understand

Be advised that this labeling task could be time-consuming and labor-intensive. You are encouraged to begin by labeling some number of sentences all together and agree on a set of rules for labeling to guarantee consistency, and then each member takes their own part to make the labeling process parallel. Create a CSV file, in which each row has a sentence and its sentiment. Lastly, using the labeled text, develop a sentiment classifier that predicts the sentiment of a new sentence. Obviously, the sentiment should be one of the three labels: positive, negative, and neutral. Try to apply all the classification algorithms you have learned in class or even those not covered in class. Choose the classifier that yields the best performance. Make sure to show in your notebook all the stats from comparisons of different algorithms. After choosing the best-performing classifier, test it using at least ten example sentences to show how it responds to different types of sentences.

Deliverables: a CSV file, a Jupyter notebook with code and output

Cluster & Topic Analysis

The goal of this component is to experience a cluster and topic analysis on the documents you have collected yourselves. First, applying k-means clustering to all the +3,000 articles, find some clusters of similar articles. The decision of the number of clusters, or *k*, is up to you. Try different numbers of clusters until you find the clusters that you believe best explain the data. Second, applying LDA topic modeling to the same data, find some topics with contributing keywords. Again, the decision of the number of topics is up to you. Try different numbers of topics until you find the topics that you believe best explain the data.

Deliverables: a Jupyter notebook with code and output

Group Project Proposal

Prepare a one-page project proposal in PDF format containing the following items:

- Your group members
- Group name with no space, e.g., Spider, JumpingJack, etc.
- The target web site and the web articles you want to scrape (Make sure to do a proof-of-concept experiment on the target web site, so you can ensure that you will be able to scrape +3,000 web pages from the web site.)
- A short description of why you are interested in those articles and how you are planning to scrape the articles

Have one group member submit your project proposal to ICON by the deadline shown above and wait for the instructor to give feedback and approve the proposal.

Jupyter Notebooks

The Jupyter notebooks for the three components should contain not only Python code and output but also sufficient annotations in the form of mark-down cells, providing rich context to readers. The same notebooks will be used for presentations as well, so they should be self-contained.

Presentation

Prepare a 10-minute group presentation to present to the rest of the class on the date shown above. You will be using the Jupyter notebooks for presentation. During the presentation, all group members should present for an equal amount of time. Note that you do not need to write any report using Word or make presentation slides using PowerPoint. You will be presenting just the Jupyter notebooks.

Evaluation Criteria

Here are the evaluation criteria for the group project:

- Amount and depth of work
- Code management
- Presentation
- Peer evaluation (if necessary)