

**MMAI 891**

**Natural Language Processing**

**Dr. Stephen W. Thomas**

**Individual Assignment**

**Click Here and Input Due Date**

**Click Here and Enter Individual or Team Name**

## Order of files:

|  |  |  |
| --- | --- | --- |
| **Filename** | **Pages** | **Comments and/or Instructions** |
|  |  |  |
|  |  |  |

**Additional Comments:**

# Sentiment Analysis, Two Ways

This question involves programming. Please submit your code, but please also answer the questions below.

1. Choose a dataset with text data labeled as either positive or negative. A few examples include the [Liu dataset list](https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#datasets) and [UCI](http://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences). There are more on [my GitHub page](https://github.com/stepthom/text_mining_resources/blob/master/README.md#datasets). You may choose any dataset in the world!
   1. Describe the dataset and its characteristics. How were the labels created? Are they trustworthy?
2. Perform sentiment analysis on the dataset twice: first, using the lexicon approach, and second, using the ML approach.
   1. For the lexicon approach, you may use any of several R or Python packages, e.g., qdap, syuzhet, tidytext, or TextBlob.
      1. Note that in the lexicon approach, you do not need to split your data into training and testing, and you do not need to perform cross validation or hyperparameter tuning.
      2. Describe which package/method you used, and briefly outline how it works under the hood.
   2. For the ML approach, build a text classifier using all the steps we’ve learned in class: preprocessing the text, extracting features from the text (i.e., vectorization using BoW or topics or doc2vec), using a ML algorithm to build a model. Don’t forget everything that we’ve learned in our ML course: cross validation, hyperparameter tuning, handling imbalanced data, etc. Make reasonable decisions and try to create the best-performing classifier that you possibly can.
      1. Briefly describe your ML process and the steps you took.
3. Compare the performance of the lexicon approach with the ML approach.
   1. Show the two confusion matrices.
   2. What is the accuracy, sensitivity, and specificity of each approach?
4. Show three example observations for each of the following scenarios:
   1. The lexicon approach was correct, but the ML approach was incorrect.
   2. The ML approach was correct, but the lexicon approach was incorrect.
   3. Both approaches were correct.
   4. Both approaches were incorrect.
5. After examining the performance and behavior of each approach, what conclusions and recommendations can you offer?

# Contributing to Open Source

Uncle Steve maintains a curated list of NLP resources on GitHub:

<https://github.com/stepthom/text_mining_resources>

The list has gained some traction on GitHub and reddit, and is becoming a useful resource for the general NLP and AI community.

For this part of the assignment, you will have a chance to make the list even better. In doing so, you will learn to use GitHub (a popular version control platform used in industry) and you will become a contributor to open source. (Both good for the resume!)

Your task is make at least five improvements/changes to the repository’s README.md file. Types of changes include:

* Adding a new link to a book, blog article, package, tool, dataset, course, etc. A very short, well-written description of the new link would be appreciated, but is not necessary.
* Removing a link that is broken.
* Adding a short, well-written description to an existing link that doesn’t currently have one.
* Fixing a typo or formatting error.

If you do not have a GitHub account, you will need to create one.

Submit your changes in the form of a pull request on GitHub. Briefly, the entire process will look like this:

* You will create your own fork (i.e., copy) of my *text\_mining\_resources* repository by clicking on the “Fork” button on the top-right of my repository’s page.
  + This will create a copy of my repository in your GitHub account. Your copy will also be named *text\_mining\_resources.*
* Working in your fork, you will edit the README.md file to add your five changes. Note that you can do this all within the GitHub interface by clicking the pencil icon on the top right of the README.md fileto edit the file directly in your web browser. There is no need to download the repository to your personal machine (although you can if you like).
* Once you are happy with your changes, you can create a pull request by clicking the “New Pull Request” button in your repository. Makes sure the base fork is set to stepthom/text\_mining\_resources.

For more information on how to submit a pull request, please see these resources:

* [https://help.github.com/articles/proposing-changes-to-your-work-with-pull-requests](https://help.github.com/articles/proposing-changes-to-your-work-with-pull-requests/)
* [https://thenewstack.io/getting-legit-with-git-and-github-your-first-pull-request](https://thenewstack.io/getting-legit-with-git-and-github-your-first-pull-request/)

For information about the Markdown file format, please see:

* [https://guides.github.com/features/mastering-markdown](https://guides.github.com/features/mastering-markdown/)

Some guidelines for this task are as follows: Please:

* Only submit links to *recent* and *high-quality* resources.
* Follow the formatting and writing style convention of the README.md.
* Make sure your English is correct and well-written.
* Use your best judgement and give me your best effort to help me to take good care of this important resource!

# Dialog Systems At Work

In this question, I want you to think about how to use dialog systems (i.e., chatbots and task-oriented agents) at your current workplace. If you are not currently employed, use your most recent employer as a proxy.

1. Describe your current employer: business model, organizational structure, objectives, revenues, etc.
2. Brainstorm at least three different opportunities in which dialog systems could be used at your organization to improve an existing process/workflow/interaction. For each opportunity, include:
   1. What kind of dialog system it would be?
   2. Who would build/maintain the dialog system?
   3. Who would the users be? How often would they use it?
   4. How would it improve the existing process/workflow/interaction? (I.e., what are the expected benefits?)
   5. What are some example user utterances and ideal system responses?
   6. What are some areas for concerns/issues to look out for? What could possibly go wrong, and what would the consequences be?
3. Select your favorite opportunity above and build a case for it.
   1. What would the ROI be? (Make some assumptions and list them.)
   2. What would the development timeline be? (Make some assumptions and list them.)