NC STATE UNIVERSITY

Data Science Academy

# Course Topic: Topic Modeling: Clustering Data through Machine Learning

## Final Capstone

### Proposal

Before you dive into building out your final capstone, we'd like you to craft a proposal for your project. This should allow you to clarify what you're looking to build, as well as get feedback on that objective before committing the full time it would take to build it.

The proposal should be approximately one page long and answer the following questions:

- What is the problem you are attempting to solve?
- How is your solution valuable?
- What is your data source and how will you access it?
- What techniques from the course do you anticipate using?
- What do you anticipate to be the biggest challenge you'll face?

When answering these questions, you should form a clear picture of the work you intend to do without having begun to build out the infrastructure to execute it yet. You may have written some code, done some initial scraping, or some initial analysis. Do not, however, start to actually build your product until it has been approved.

When you're ready to submit, send it to me by email so I will approve it and you can continue.

## Notebook

Now is your chance to show your understanding of the data science workflow when it comes to NLP Projects.

For this stage, we want you to build out a notebook that builds and demonstrates your NLP Product. Reference other scripts as needed but be sure to include those in the same GitHub repository. This notebook should demonstrate your technical prowess as well as visualization and narrative storytelling. As such it should include all stages of your process in a clean, easy to read form.

Specifically make sure to:

- Wrangle your data. Get it into the notebook in the best form possible for your analysis and model building.
- Explore your data. Make visualizations and conduct statistical analyses to explain what's happening with your data, why it's interesting, and what features you intend to take advantage of for your modeling.
- Build a modeling pipeline. Your model should be built in a coherent pipeline of linked stages that is efficient and easy to implement.
- Evaluate your models. You should have built multiple models, which you should thoroughly evaluate and compare via a robust analysis of residuals and failures.
- Present and thoroughly explain your product. Describe your model in detail: why you chose it, why it works, what problem it solves, how it will run in a production like environment. What would you need to do to maintain it going forward?

When you have the notebook finished send me the link to your repository for Grading.

### Presentation

And now you’ve come to the last step of your final capstone --- it’s time to present your product.

You can think of this presentation in several different ways. It could be like pitching your product to potential investors. It could be pitching it to a relevant company while interviewing. The key is this presentation should be engaging, informative, and beautifully put together. It should also run approximately 15 minutes and be technical but product focused. That means while you should explain the machine learning you used, focus more on explaining the product itself, what problem it solves, and why your solution is valuable.

NC STATE UNIVERSITY

Data Science Academy

## Final Capstone Rubric

| | Points | 0-50% | 50-75% | 75-100 |
|---|---|---|---|---|
| Questions | 14 | Questions overly simplistic, unrelated, or unmotivated | Questions appropriate, coherent, and motivated | Questions well motivated, interesting, insightful, and novel |
| Dataset & Analysis | 10 | Choice of analysis is overly simplistic or incomplete. Not available data | Analysis appropriate and relevant data used. | Analysis appropriate, complete, advanced, and informative.<br>Robust datasets found and used |
| Code review | 15 | Little evidence that group members are giving constructive feedback on each other's code. | Some evidence that group members are giving constructive feedback on each other's code, leading to better code. | Extensive evidence that group members are giving constructive feedback on each other's code, leading to better code. |
| Results | 15 | Conclusions are missing, incorrect, or not based on analysis<br>Inappropriate choice of plots; poorly labeled plots; plots missing | Conclusions relevant, but partially correct or partially complete<br>Plots convey information but lack context for interpretation | Relevant conclusions explicitly tied to analysis and to context<br>Plots convey information correctly with adequate and appropriate reference information |
| Presentation | 20 | Verbal presentation is illogical, incorrect, or incoherent.<br>Visual presentation is cluttered, disjoint, or illegible. The visual aesthetics are nonexistent<br>Verbal and visual presentation unrelated. Errors are present in the slides | Verbal presentation partially correct but incomplete or unconvincing<br>Visual presentation is readable and clear. The visuals aesthetics are appealing<br>Verbal and visual presentation related. Not many errors in slides | Verbal presentation is correct, complete, and convincing<br>Visual presentation is appealing, informative, and crisp. The visual aesthetics is pleasing and flows well with presentation<br>Verbal and visual presentation clearly related. No errors in slides. |
| Writing | 13 | Explanation is illogical, incorrect, or incoherent | Explanation is correct, complete, and convincing | Explanation is correct, complete, convincing, and elegant |
| Reproducibility | 13 | Code didn't run | Recipes in project directory correctly load data and generate all results and figures in report | Recipes additionally validate data against its source (such as URL or another download). The recipes generate all exploratory work and supplementary analysis |
| Total | 100 | | | |