

QTM-340: Data Science with Text
Homework 3
Due Thursday, 10/8 4:20 PM via Canvas

Use the *New York Times* obituaries corpus. Pose a question that can be answered with one of the methods we've learned in this unit (sentiment analysis, NER, POS-tagging, tf-idf, topic modeling) to arrive at a provisional answer.

The challenge of this homework is three-fold:

1. Posing the question itself. Part of the purpose of this assignment is to determine what it means to ask a good question for data science. Your question should not be already obvious to anyone already paying attention, nor should it be answerable without computation. And yet, it should not be merely pedantic, something only asked *because* we can answer it with computation, but lacking interest on its own. Pose an interesting question—interesting because it promises to teach us something worth learning—newly answerable because of our computational methods.
2. Implementation. The technical task at hand.
3. Interpretation. Can you provide an accurate, reasonable, and persuasive interpretation of your results in response to your initial question. Your interpretation need not bring us to a final answer, but might turn out to be a first step along the way. That's fine.

Submission details:

Your homework should consist of a Jupyter notebook with the following sections*:

1. A **text cell** that states the question you're exploring.
2. A **code cell** with any corpus pre-processing and/or segmentation. (Note that for some methods, such as topic modeling, the corpus segmentation comes **after** the implementation of the method; that is fine).
3. A **code cell** with any required tokenization / stopword removal
4. A **code cell** that implements the method you have chosen to answer the question
5. A **text cell** that briefly (i.e. in a short paragraph) interprets the results, and identifies some potential next steps

* Sections 2-4 are intended as a guide. If your approach requires a different format, you can diverge from the specified cells. But you must include cells #1 and 5 (the research question and the analysis) regardless of the structure of your project code.

Please make sure your last name is part of the file name of your Notebook.

Upload your Notebook (the file ending in .ipynb) to Canvas by the date specified above.