

What do Presidents talk about?

Obtain the *Presidential Inaugural Addresses* corpus from NLTK or another source and explore it with LDA topic modeling methods (You may use either the NLTK gensim (GenSimTopicModeling.py) or any R implementation of LDA).

If you use gensim you will have to 'pip install' it. I recommend reading this documentation for starters:

https://rstudio-pubs-static.s3.amazonaws.com/79360_850b2a69980c4488b1db95987a24867a.html

If you want to try R, the Quanteda overview on the syllabus is a good place to start. I did not have much success with the code at this link.

<https://cran.r-project.org/web/packages/quanteda/vignettes/quickstart.html>

Experiment and find the model that does the best job of capturing what you are interested in with respect to the addresses. Do not run just one model (for the reasons discussed in class and in the readings).

When you have found that model,

- 1) Describe how you arrived at it in a Word doc (e.g. wilkerson-topics.doc)
- 2) Provide a couple of examples of how you validated your topics
- 3) Make a .ppt slide that displays your top substantive topics in 10 columns.
 - a. The topic label you have assigned at the top
 - b. The top 10 terms of each topic below that

I am NOT asking you to produce a 10 topic model. Your top 10 list might be drawn from a 10, 15, 20 or 75 topic model.

Please ask if you have questions about any part of this process

Please submit to dropbox by Saturday evening!