Selecting Dataset:

- Finding a good dataset requires a good amount of research into clean text datasets
- Using websites like Kaggle, Project Gutenberg, 20 Newsgroups, and others, clean text datasets like the ones shown below can be found

A Presidents	A Party Political Affiliation	A transcripts Transcripts
44 unique values	Republican 43% Democratic 34% Other (10) 23%	**44** unique values
George Washington	Unaffiliated	Fellow Citizens of the Senate and the House of Representatives: Among the vicissitudes incident to l...
John Adams	Federalist	When it was first perceived, in early times, that no middle course for America remained between unli...

- As we aim to semantically cluster long text documents, datasets that encompass a large variety of ideas is required for optimal system performance
    - Examples of datasets that meet the length requirements are: non fiction books, long speeches, non dialogue format text documents.
    - Datasets to avoid: definitions, vague and shallow text.