Get test data

Here is the link and setup process for all the test data.

- 5 million sales records - http://eforexcel.com/wp/downloads-18-sample-csv-files-data-sets-for-testing-sales/
- 23 GB of author, book and publishers data - http://openlibrary.org/data/ol_cdump_latest.txt.gz
- NYC Yellow taxi data - https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
- Wiki pages - https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2
- Wiki pages - https://www.kdnuggets.com/2017/11/building-wikipedia-text-corpus-nlp.html
- Twitter data - You can download your own twitter data. Example is in the git location