Mining digital libraries

The central objective of this tutorial is to show how researchers within the humanities can access and use a cloud-based corpus from within Jupyter Notebook, exemplified with the digital collections of the Norwegian national library. The tutorial is relevant for the Nordic humanistic community in several ways:

1) Show how copyrighted material can be utilized for corpus studies.
2) Different tools built as modules over an API for defining and accessing a corpus for analysis like character modelling, collocation analysis, clustering, growth diagrams and more.
3) Demonstrate the benefits of using Jupyter Notebook for researchers without programming background.
4) Study the connection between texts and library metadata

As for (1), a problem for many researchers is the use of copyrighted material. However, the researcher often has no need for the actual text as such; some features of it may suffice, like bag of words, a participle count or a character model. None of these features challenge the copyright holder. A centralized repository of copyrighted material can provide feature sets.

The points in (2) and (3) cover the need for programmers as well as researchers without programming background. While the former need a documentation of the low-level interface to the cloud, the actual API, the latter want an accessible interface for doing corpus analysis, and Jupyter Notebook provides such an interface via top level functions and commands expressed in a programming language, e.g. Python or R.

Item (4) shows how readymade library metadata can be integrated for building corpora based on those data, like Dewey decimal codes or topic words. We will show how metadata can be used to build, select and compare corpora.

The participants will experiment with the API, and get a hands on experience with the tools. The kind of material we go through will be distributed as a Github repository on format similar to this: https://github.com/Yoonsen/NB_API_Python.

Background

At the National Library of Norway, a mass digitization project was initiated in 2006, with the goal of digitizing the entire collection of books, newspapers, movies, radio- and television-broadcasts, music etc., in sum everything published in the public domain in Norway of all media types, the entire cultural heritage of Norway. For the books, the entire stock was finished digitized in 2017, a collection of about 500.000 books.

The corpus that is available via the API, consists of about 50 billion tokens, which is considered big for a rather small language like Norwegian (5 million speakers). In comparison, the Google Books corpus contains approximately 500 billion tokens for English.

The National Library cooperates with scholars of literary studies and linguistics in developing and applying methods of data mining to our material. We develop services that make our content available for quantitative research, without challenging intellectual property rights. One such service is NB N-gram for Norwegian, comparable to Google N-gram Viewer for English and other languages, as found here: http://www.nb.no/sp_tjenester/beta/ngram_1/..

Session Format

Full day tutorial. We start out with an introduction to Notebooks using a motivating example with commands expressed as Python functions. We then move on to show how to build corpora using metadata, either from the library itself or within notebooks, and exemplify a corpus investigation by finding concordances (KWIC). This part of the tutorial will also cover issues concerning OCR, and how they might be overcome in searching. The next topics are on comparing corpora with respect to statistics of terminology, for example, what are the key words of Dewey 641, or texts written in a certain period. Then we plan on covering collocation analysis and clustering. These topics are exemplified with notebooks in the Github repository mentioned above.

Target audience and number of participants

Humanities researchers, for example scholars of literary and media studies, corpus and computational linguists, as well as librarians and curators.

Number of participants should not exceed twenty.

Technical requirements

We will need projector and screen. Participants bring their own laptops.

Workshop leaders

- Lars G. B. Johnsen: Research librarian at the National Library of Norway, PhD in linguistics. Fields of interest: semantics, grammar, philosophy of language, probability theory and applications. Email: lars.johnsen@nb.no
- Yngvil Beyer: Research librarian at the National Library of Norway, Master in Media studies. Fields of interest: Manuscripts and media, handwritten text recognition, text encoding. Email: yngvil.beyer@nb.no