Summarizing the state of the Catholic Youth Literature Project

This essay summarizes the purpose, process, and technical infrastructure behind the Catholic Youth Literature Project. In a few sentences, the purpose was two-fold: 1) to enable students to learn what it meant to be Catholic during the 19th century, and 2) to teach students the value of reading "closely" as well as from a "distance". The process of implementing the Project required the time and skills of a diverse set of individuals. The technical infrastructure is built on a large set of open source software, and the interface is far from perfect.

Purpose

The purpose of the project was two-fold: 1) to enable students to learn what it meant to be Catholic during the 19th century, and 2) to teach students the value of reading "closely" as well as from a "distance". To accomplish this goal a faculty member here at the University of Notre Dame (**Sean O'Brien**) sought to amass a corpus of materials written for Catholic youth during the 19th century. This corpus was expected to be accessible via tablet-based devices and provide a means for "reading" the texts in the traditional manner as well as through various text mining interfaces.

During the Spring Semester students in a survey class were lent Android-based tablet computers. For a few weeks of the semester these same students were expected to select one or two texts from the amassed corpus for study. Specifically, they were expected to read the texts in the traditional manner (but on the tablet computer), and they were expected to "read" the texts through a set of text mining interfaces. In the end the the students were to outline three things: 1) what did you learn by reading the text in the traditional way, 2) what did you learn by reading the text through text mining, and 3) what did you learn by using both interfaces at once and at the same time.

Alas, the Spring semester has yet to be completed, and consequently what the students learned has yet to be determined.

Process

The process of implementing the Project required the time and skills of a diverse set of individuals. These individuals included the instructor (Sean O'Brien), two collection development librarians (**Aedin Clements** and **Jean McManus**), and librarian who could write computer programs (myself, **Eric Lease Morgan**).

As outlined above, O'Brien outlined the overall scope of the Project.

Clements and McManus provided the means of amassing the Project's corpus. A couple of bibliographies of Catholic youth literature were identified. Searches were done against the University of Notre Dame's library catalog. O'Brien suggested a few titles. From these lists items were selected for inclusion for purchase, from the University library's collection, as well as from the Internet Archive. The items for purchase were acquired. The items from the local collection were retrieved. And both sets of these items were sent off for digitization and optical character recognition. The results of the digitization process were then saved on a local Web server. At the same time, the items identified from the Internet Archive were mirrored locally and saved in the same Web space. About one hundred items items were selected in all, and they can be seen as a set of PDF files. This process took about two months to complete.

Technical infrastructure

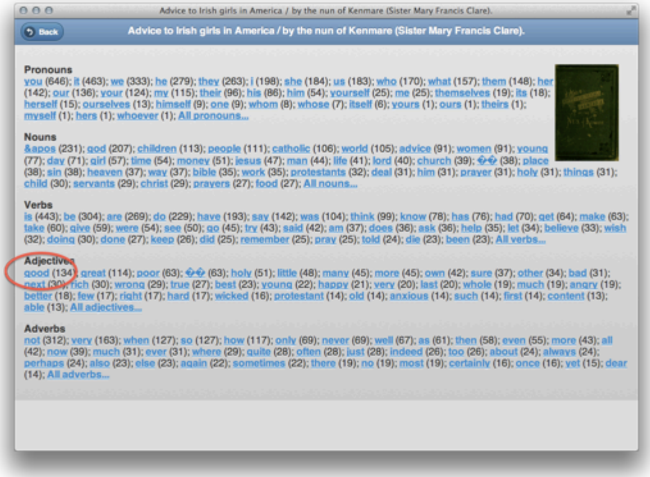
The Project's technical infrastructure enables "close" and "distant" reading, but the interface is far from perfect.

From the reader's (I don't use the word "user" anymore) point of view, the Project is implemented through a set of Web pages. Behind the scenes, the Project is implemented with an almost dizzying array of free and open source software. The most significant processes implementing the Project are listed and briefly described below:

- **mirroring** - Much of the text mining services require extensive analysis of the original item. To accomplish this local copies of the texts were mirrored locally. By feeding the venerable wget program with a list of URLs based on Internet Archive unique identifiers, mirroring content locally is trivial.

- **name-entity extraction** - There was a desire to list the underlying names, places, and organizations from each text. These things can put a text into a context for the reader. Are there a lot of Irish names? Is there a preponderance of place names from the United States? To accomplish this task and assist in answering these sorts of questions, a Perl script was written around the Stanford Named Entity Recognizer. This script extracts the entities, looks them up in DBedia, and saves metadata (abstracts, URLs to images, as well as latitudes & longitudes) describing the entities to a locally defined XML file for later processing. A CGI script was then written to provide a reader-interface to these files.

- **parts-of-speech extraction** - Just as lists of named entities can be enlightening so can lists of a text's parts-of-speech. Are the pronouns generally speaking masculine or feminine? Over all, are the verbs active or passive? To what degree are color words used words in the text? To begin to answer these sorts of questions, a Perl script exploited a Perl module called Lingua::TreeTagger. The script extracts parts-of-speech from a text file and saves the result as a simple tab-delimited file for later use.


Advice to Irish girls in America / by the nun of Kenmare (Sister Mary Francis Clare).
Back

Advice to Irish girls in America / by the nun of Kenmare (Sister Mary Francis Clare).

Pronouns
you (646); it (463); we (333); he (279); they (263); i (198); she (184); us (183); who (170); what (157); them (148); her (142); our (136); your (124); my (115); their (96); his (86); him (54); yourself (25); me (25); themselves (19); its (18); herself (15); ourselves (13); himself (9); one (9); whom (8); whose (7); itself (6); yours (1); ours (1); theirs (1); myself (1); hers (1); whoever (1); All pronouns...

Nouns
&apos (231); god (207); children (113); people (111); catholic (106); world (105); advice (91); women (91); young (77); day (71); girl (57); time (54); money (51); jesus (47); man (44); life (41); lord (40); church (39); �� (38); place (38); sin (38); heaven (37); way (37); bible (35); work (35); protestants (32); deal (31); him (31); prayer (31); holy (31); things (31); child (30); servants (29); christ (29); prayers (27); food (27); All nouns...

Verbs
is (443); be (304); are (269); do (229); have (193); say (142); was (104); think (99); know (78); has (76); had (70); get (64); make (63); take (60); give (59); were (54); see (50); go (45); try (43); said (42); am (37); does (36); ask (36); help (35); let (34); believe (33); wish (32); doing (30); done (27); keep (26); did (25); remember (25); pray (25); told (24); die (23); been (23); All verbs...

Adjectives
good (134); great (114); poor (63); �� (63); holy (51); little (48); many (45); more (45); own (42); sure (37); other (34); bad (31); next (30); rich (30); wrong (29); true (27); best (23); young (22); happy (21); very (20); last (20); whole (19); much (19); angry (19); better (18); few (17); right (17); hard (17); wicked (16); protestant (14); old (14); anxious (14); such (14); first (14); content (13); able (13); All adjectives...

Adverbs
not (312); very (163); when (127); so (127); how (117); only (69); never (69); well (67); as (61); then (58); even (55); more (43); all (42); now (39); much (31); ever (31); where (29); quite (28); often (28); just (28); indeed (26); too (26); about (24); always (24); perhaps (24); also (23); else (23); again (22); sometimes (22); there (19); no (19); most (19); certainly (16); once (16); yet (15); dear (14); All adverbs...

The list of parts-of-speech provides the reader with some context. Notice how the word "good" is the most frequently used adjective.

- **word/phrase tabulation and concordancing** - To support rudimentary word and phrase tabulations, as well as a concordance interface, an Apache module was written around two more Perl modules. The first, Lingua::EN::Ngram, extracts word and phrase occurrences. The second, Lingua::Concordance, provides an object-oriented keyword-in-context interface.

- **metadata enhancement and storage** - A rudimentary catalog listing the items in the Project's corpus was implemented using a Perl module called MyLibrary. The MARC records describing each item in the corpus were first parsed. Desired metadata elements were mapped to MyLibrary fields, facets, and terms. Each item in the corpus was then analyzed in terms of word length as well as readability score through the use of yet another Perl module called Lingua::EN::Fathom. These additional metadata elements were then added to the underlying "catalog". To accomplish this set of tasks two additional Perl scripts were written.

- **HTML creation** - A final Perl script was written to bring all the parts together. By looping through the "catalog" this scrip generates HTML files designed for display on tablet devices. These HTML files make heavy use of JQuery Mobile, and since no graphic designer was a part of the Project, JQuery Mobile was a godsend.

The result is a system that enables the reader to view the texts online as well as do some analysis against them. The system functions in that it does not output invalid data, and it does provide enhanced access to the texts.

Summary

To summarize, the Catholic Youth Literature Project is far from complete. For example, it has yet to be determined whether or not the implementation has enabled students to accomplish the Project's stated goals. Does it really enhance the use and understanding of a text? Second, the process of selecting, acquiring, digitizing, and integrating the texts into the library's collection is not streamlined. Finally, usability of the implementation is still in question. On the other hand, the implementation is more than a prototype and does exemplify how the process of reading is evolving over time.

Eric Lease Morgan
University of Notre Dame
March 29, 2012

For more detail and links, see http://bit.ly/Hp1REt