Applying text mining tools to Jewish community cookbooks

Anna M. Levia
Stanford University Libraries
amlevia@stanford.edu
2. Digital Humanities can be a difficult construct to wrap one’s mind around. It is, in the words of one practitioner of the field, “humanities scholars’ increasing use and exploration of information technology as both a scholastic tool and a cultural object in need of analysis.” Patrik Svensson, retrieved from http://digitalhumanities.org/dhq/vol/3/3/000065/000065.html
Digital Humanities:
Some friendly definitions

“Digital Humanities, loosely applied, refers to the use and application of computational tools and methods to humanist domains of study, and vice versa.”
Princeton University, https://digitalhumanities.princeton.edu/about/

“Digital Humanities, once referred to as "Humanities Computing", is the intersection of computing, research, and teaching in the fields of the humanities. It is the use of computing to engage with texts.”
Western Michigan University, http://libguides.wmich.edu/digitalhumanities

Thanks in large measure to the leadership and vision of director Michael Keller, the libraries at Stanford have been closely involved in Digital Humanities activities on campus for over twenty years. Stanford University Libraries, most recently through a newly formed administrative entity called the Center for Interdisciplinary Digital Research, partner with various departments to support faculty and students as they integrate Digital Humanities into their research and teaching.

Librarians at Stanford, even those who are not directly involved with digital humanities research, are encouraged to learn about ongoing projects.
It was in this spirit of learning and exploration that I took part earlier this year in a series of workshops for library staffers organized by Stanford textual research librarian Jacque Hettel. The series focused on skills and resources not only for helping patrons, but also for facilitating staff members' own digital research projects.
Cookbooks are great for text mining!

- Lots of elements to count and compare
- Community cookbooks provide cultural snapshots of time and place

When we got to the session on text-based model creation, I knew that I would have to do my own project involving my great love, Jewish community cookbooks. After all, Jacque made it sound so simple!
Here are the basic parameters for building a robust database. I devised a workflow for scanning and shared it with Levi Thompson, David Hirsch's hardworking student assistant at UCLA. We decided to start with one cookbook from each of four decades—1950s, 60s, 70s and 80s.
Step 2: Let the scanning begin!

- Takes about 45 minutes per book, section by section

We scanned the documents, including the title page at fairly high resolution—600 dpi. It’s important to keep the books as straight as possible on the scanner. We then saved each section as a pdf.
After trial and much error, we made a discovery: Google docs works better than other tools, including Adobe, for converting pdf files to machine readable text, and best of all, it’s free!

To do OCR conversion using Google Docs:

- Set the “upload setting” to “convert text from ...” and upload the pdf files.
- The documents will be converted and saved as a “gdoc” containing the original images and the OCR text.
- It’s important to limit the number of images to no more than 9 or 10 per file, since that’s all Google Docs can handle. If you try to convert files containing more than nine or ten page images, you end up with a lot of gibberish.
Here's a scan from the 1967 *Council Cookbook*, by the Nashville Chapter of the National Council of Jewish Women.
FRESH APPLE CAKE
4 cups apples, chopped
2 cups sugar
2 tsp. vanilla
Place two cups chopped apples in bowl. Cover with one cup sugar. Add balance of apples and cover with other cup of sugar. Let stand 45 minutes to form syrup. Beat eggs with oil at medium speed, then with 1 tsp. sugar, and add. Add flour, sifted with baking powder and soda, alternately with applesauce. Bake in large tube or loaf pans. Bake at 350 degrees for 45 minutes. Yields 16 servings.

APPLE CAKE
Cream butter, add sugar and beaten egg. Add alternately other dry ingredients and coffee. Add raisins and nuts. Bake at 350 degrees for 35 or 40 minutes. Frost while warm with 1 cup powdered sugar mixed with 2 tbsp. coffee.

APPLE SAUCE CAKE
1 tsp. each spices
2 tsp. soda
1% cups raisins
1% cups chopped pecans
5 cups sifted flour
Cream soft butter and margarine with sugar. Add one cup applesauce and whisk. Sift flour, mix in all spices with flour, adding raisins and nuts. Mix with the second cup of applesauce.Alternate adding the flour and applesauce to...
GEFILTE FISH

Broth:
Fish skin, bones and head
2 onions, sliced
1 or 2 stalks celery
½ tsp. sugar
salt and pepper to taste
1 carrot, sliced

Have fish cleaned, filleted and skinned. Rinse skin, bones and head and place in large kettle. Slice vegetables for broth and place over fish parts. Add 2 qts. cold water and seasonings. Bring to boil and let simmer 30 minutes.

Fish Balls:
3 lbs. assorted fish, buffalo, pike and whitefish
2 onions
½ tsp. sugar
3 tbsp. matzo meal
salt and pepper generously
2 eggs

Grind on finest blade, fish alternated with onion. Mix ground mixture with eggs, matzo meal and seasonings. Add about 1/3 cup cold water or enough to make mixture soft. With wet hands form fish mixture into balls. Place gently into boiling broth. Cover kettle loosely and simmer about 2 hours. Add water if necessary to have enough to cover as liquid should be reduced to less than half when done. Turn off heat and let stand one hour before removing to cool. Strain the stock over fish balls and arrange carrots around it.

Note: Be sure to save bones, skin and head of fish to use in recipe.

It’s very important to do careful OCR cleanup, which can sometimes feel like typing the entire cookbook by hand.

- You open each file in Google Docs
- The document will be opened with the embedded editor
- You then review the document and make edits to the OCR. Changes are saved automatically.
- Be sure to keep the cookbook handy too--it’s easier to read the original text for reference.

OCR cleanup takes forever!
Step 5: Don’t forget to mark up the text!

Text Encoding Initiative (TEI) markup “allows texts to be marked up semantically at any level of granularity.”

http://www.tei-c.org

Free xml text editing software is available, but Oxygen works best.

The Text Encoding Initiative is an international organization founded in 1987 to develop guidelines for encoding machine-readable texts in the humanities and social sciences. ‘TEI’ is also used to refer to the Guidelines themselves, and to the set of schemas they describe. TEI markup supports a wide range of useful functions including text analysis. Although you can use any text editor to edit TEI documents, tools such as Oxygen allow you to validate your encoding against the TEI schema. XML, of course, stands for Extensible Markup Language and is used to structure, store, and transport information.

[Text retrieved and adapted from http://www.tei-c.org/About/faq.xml]

Without markup, your corpus is just a word soup.
Although the thought of making lots of mistakes is frightening, the best way to learn markup is to just dive in and do it. After the first page took me almost an hour to complete, I began using several templates, which I kept in an open Word doc. On the UCLA end, David’s assistant used this system as well.
...just paste in your data from the OCR'd Google doc—it takes only 6-8 minutes per page!

Our markup is pretty basic on this first pass—page breaks; headings; lists of ingredients; instructions; and names of recipe contributors, including whether or not the women referred to themselves by their husband’s name.
Here’s a screen shot of the Oxygen xml editor. The little green box in the upper right corner shows that each section of text is properly nested within opening and closing tags.
Levi Thompson at UCLA completed two cookbooks and I managed to complete one, the Council Cookbook. Of course, three cookbooks is not a large enough corpus from which to derive meaningful results, but it is enough to see examples of the types of information that text mining can reveal. Luckily, we ended up with one cookbook from each of three consecutive decades.
After the xml markup is complete, it is finally possible to do some textual analysis. On Jacque Hettel’s recommendation, I chose **WordSmith Tools**, “an integrated suite of programs for looking at how words behave in texts.”

**Wordlist** obviously produces a list of all the words in a text or group of texts, along with each word’s frequency.

**Concord** creates a concordance, which allows you to see any word or phrase in context.

**KeyWords** compares the words in the text with a reference set of words usually taken from a large corpus of text. Any word which is found to be outstanding in its frequency in the text is considered "key".

[Explanatory text retrieved and adapted from http://www.lexically.net/wordsmith/index.html]

At last, we can run our xml text files through WordSmith Tools, create lists, and see some results!
To create keyword files, I compared each cookbook wordlist, X, with a reference corpus made up of the other two, called Not X. This keyword sample shows some of the 209 words that occur with a notably higher frequency in the Council Cookbook than in the other two texts. RC frequency refers to Reference Corpus.
As you look over the Keyword and Word lists, questions begin to emerge...
What kinds of cooking equipment are mentioned, where, and when?
What kinds of typically Jewish recipes and ingredients appear?
How frequently do words associated with dietary trends, or words associated with kashrut, appear?
Are some non-kosher foods more acceptable than others?
There’s always room for Jell-O—in a **mold**!

The combined wordlist shows that the mold was a popular piece of equipment—there are 119 instances of the word appearing in all three cookbooks.
After highlighting the word, I computed a concordance for “mold.” We can see that the most molds by far, for both baking and jello, were used in the 1967 cookbook. In America from the 1920s through the 1960s, gelatin salads were promoted by women’s magazines as being the height of elegance.
From the 3-cookbook wordlist, we can see that the word “Processor” appears 15 times, but only in only 1 text

What about small appliances? Here’s the word “processor” as it appears in the combined wordlist.
Here’s the concordance for the word “processor.” We can see that it only appears in *Tasty Tidbits*, and that in several of the entries, other preparation methods are suggested in case the cook does not own a food processor. This makes perfect sense. The first home food processor, the Cuisinart, had only been available to home cooks in the US for less than a decade, so in 1980 they were not yet ubiquitous.
...but in 1967 the Nashville Council ladies really loved their blenders—30 out of 48 references

In the concordance for the word “blender” from the 3-cookbook wordlist, 5 of the recipes suggest alternative preparation methods; #25 cautions the cook not to use a blender...
Statistics, in no particular order, for some typical Jewish foods. When you consider that these three cookbooks contain thousands of ingredients and recipes, the numbers here seem especially low. On a regional note, the Dallas cookbook explains the word “kashe” as “buckwheat grits.”
We can all think of dietary trends and fads that have gone in and out of fashion over the decades. Six of the 15 references to “fat” in Tasty Tidbits are in the context of skimming it off the food before serving. In Our Cookery, two of the “fat” references are to fat removal; one is in a recipe called “No fat matzo balls.” Twelve out of 86 references in the Council Cookbook refer to skimming, pouring off, draining, or otherwise removing excess fat. The other references are to fat as an ingredient. The reference to “fattening,” from the Council Cookbook, says, “Revoltingly fattening, but delicious.” There are no mentions in any of the cookbooks of artificial sweeteners, carbohydrates, fiber, or gluten.
Is shellfish “more kosher” than pork?

The concordance for “shrimp”

None of these cookbooks are kosher, but non-kosher ingredients are more prevalent by far in the 1967 Nashville cookbook. The first time I looked at this cookbook, I did a double take and flipped back to the title page to make sure it was really published by a Jewish organization. Out of 127 references to shrimp, 107 are from the Council Cookbook.
Crabmeat was also a popular ingredient in the Nashville cookbook.
In the combined word list, “Bacon” occurs 27 times in 2 out of 3 texts.

*Our Cookery* from 1973 does not have any recipes that call for bacon.
There are 27 total references to bacon in the two other cookbooks. Two of the 9 *Tasty Tidbits* “bacon” occurrences, however, call for turkey bacon and imitation bacon.
It was startling to see the word “pork” in a Jewish cookbook, although it is the least common term in the treyf lexicon. According to these non-scientific results, the earlier the publication date, the less concern there is for kashrut. Would analysis of a larger corpus bear that out? These cookbooks are from the South and Western United States. Communities in other regions of the country may have made different choices regarding acceptability of so many non-kosher recipes.
Looking at the concordance for the word “dairy,” the two entries from the 1967 cookbook refer to dairy as opposed to meat; the others refer to dairy sour cream.
There are only 3 references to pareve foods, all from *Tasty Tidbits*. 
Although it’s not very surprising, none of these cookbooks mentions the word “kosher.”
I hope that I’ve succeeded in showing you the potential of text mining as a research tool. Once the heavy lifting is done, text mining makes deep lexical analysis much quicker and easier. It would be fascinating to analyze text from a much larger corpus in order to track which words occur in cookbooks from additional time periods, regions, and among communities of more diverse religious observance. This would require scanning, OCR conversion, and markup of about 57 additional community cookbooks from all parts of North America and spanning five or six decades.
Thank you! And thanks to...

Yonatan Levia (with Humphrey), our house: OCR cleanup

Jacque Hettel, PhD, Stanford University Libraries: Abundant and patient technical assistance

Levi Thompson, UCLA: Scanning, OCR cleanup, and markup—2 cookbooks!

For their invaluable advice and assistance, thanks to Glen Worthey of Stanford University Libraries; Elaine Levia and Oz Levia; and for being such a good sport thanks to David Hirsch.