CMPUT 401 Better Find an Expert Project (BFex)

These notes are to kick off our discussion. These will evolve as we talk in more detail and as we think about it more.

Deliverable 1 : **Client Interaction with BFEX**

1) a *web-based interface* mode for the BFEX software, like the current Find-an-Expert (FEX), with these enhancements
    a. interface with spell checker, unlike current version ("nanotecholy" → do you mean "nanotechnology")
    b. multi-word input handled as single keyword phrase, e.g., *software engineering* should not be interpreted "software OR engineering" or as "software AND engineering". Ditto input such as "climate change", "green energy". These are all single concepts. So a keyword is one or more words that the user enters, and if it's more than one word, it's interpreted as a *phrase* that must match exactly.
    c. A keyword is also a faculty member name, e.g., Eleni Stroulia ; Buriak
    d. Most users are not really going to care about Boolean search capability for this use, but in principle it would be nice if the user could type in
        - *software engineering and health*
        - *nanotechnology and energy*
        - *algebra or ring theory*

      But certainly nothing complex like ((a and b) or c)
    e. It operates like the current FEX: type in a keyword, get back all the faculty members plus their entire keyword set, where the typed in keyword matches something in the keyword set

2) a *batch mode* for the BFEX software for testing, which I as the client (and you) can use to demo the capabilities easily
    a. the input to batch-BFEX is a list of <keyword , keyword-building-approach-ID> pairs, where keyword is what a user would enter and **keyword-building-approach-ID** corresponds to one of several of keyword-building-approaches, which are sort of defined below.
    b. the output of batch-BFEX is the <keyword, method pair> followed by the faculty members that are returned, along with the *entire keyword set* that was developed for that faculty member.
    c. Basically, batch mode is just like you'd expect batch mode to work and it saves its output in some file for later review.

Deliverable 2: Implement Different keyword building approaches

A **keyword building approach** is one or more *data source* combined with one or more *manipulations* of whatever is found from that data source.

Here is a list of possible *data sources* for mining keywords about a faculty member:

1. OrcID web page
2. Researcher ID web page

Example: See upper right hand side of https://www.ualberta.ca/science/about-us/contact-us/faculty-directory/philip-currie

Not every faculty member has a researcherID or an orcID.

3. NSERC grant "lay paragraphs"
    - you have to ask Prof. Stroulia or David Turner for this data source
    - let me know if it's for all people in Science at the UofA, or just some departments.

4. Faculty member web pages
    - There are two 'types' of faculty web pages, generally speaking. The first type is some kind of standard web page that a department has for each faculty member. The second type is whatever they might link off this standardized web page. So Currie has a lab page. Most Chemistry profs will have a "lab group" page or very elaborate research pages with projects described.
    - However, there is not necessarily the same kind of standardized web page across departments
    - Some web pages might actually have "keywords" listed. Often called "interests" or "research interests"

5. How departments classify their own faculty members and how faculty members classify themselves
    - A department web site has invested some effort in classifying its own faculty members. Thus, the categories of research that appear on a department web page are themselves meaningful research keywords. See Computing Science's research page, and its various ways of grouping faculty members into categories. So these words or categories themselves have some core ground truth to them. See Chemistry's Research page, See Physics. Research page You'll find these faculty groupings under "research".

6. Other?

A few observations:
    - Some of these different sources serve to reinforce each other. So clearly, Stroulia would have "software engineering' on her standardized web page, on her ResearcherID web page, and this same term would appear in how the Department of Computing Science categorizes her

What are *manipulations* of data sources?
    - Well, for some of these data sources, the first and the only manipulation is to *extract* what is found there, .e.g, in an html field marked "keywords" on a web page.
    - After that, the manipulations are up to you to define and then we see how different ones behave and which ones seem to give a 'good' set of keywords. I suppose another example of a *manipulation* is a word-cloud building algorithm operating on some particular datasources, where the top 10 words identified by the word cloud algorithm are the ones that get selected for the keywords.
    - But coming up with different manipulations performed on these data sources would be useful.
    - Maybe there is a role for synonyms for words (e.g., if someone types in *bird* but what you have stored is avian, or if someone types in *insects* and what you have stored is *entomology*. Would it be smart or seem like an error to return. Maybe your search algorithm itself could decide to do something quasi intelligent, e.g., if I type in *insects* and the return set is null, the BFEB decides to find *related words* for insects. Maybe it just goes off and googles "study of insects" and "synonyms for insects" and uses what it finds to re-search again….

- The reason to implement different keyword building approaches is to play and to experiment. We don't know *a priori* what's going to end up being 'good' If a keyword building approach seems not so good rather than throw it away, name it, and save it. It might become better in combination with some other method. Just don't throw away any method that you define that seems plausible, on the face of. Someone is going to say "why didn't you just do X" and then you can say "well, we did do X, and here's what X produces"

Other specifications

- Range of faculty members to use
  - There are only about 300 faculty members. Might as well use them all.
  - David Turner can give you a listing of each faculty member and his/her department
  - After you get the faculty member list, it might be helpful to the following features for each one
    - ORCHID URL (null or the actual URL)
    - Researcher URL (null or actual URL)
    - Department research web page URL
    - Faculty member 'official' web page URL
    - Any other additional web pages the faculty member has
    - The lay paragraph from the NSERC database (david turner to supply)
    - Other additional data sources (maybe titles from researcher ID)
  - Seems if you build this, then any time a URL changes or a paragraph changes, this DB could be updated and then the algorithms could be re-run.

Thoughts re: quantity and quality of keywords that are developed

- Somehow, we want more than just "software engineering" to show up for Professor Stroulia, or just "nanotechnology" to show up for Prof Buriak. The issue will be how many and how are they decided.
- Spend time looking at a lot of the faculty members in different departments. In some disciplines, the use of a particular methodology is very important in describing their expertise. E.g., *NMR* is important to distinguish someone's area of expertise, just the way that *Bayesian* indicates a certain kind of approach in statistics
- The entire keyword set you return will, *collectively*, describe a person's expertise.
- The more esoteric the discipline, the more important it is to have "what it's good for" keywords in there. See the difference between Jillian Buriak's keywords vs. John Davis's keywords. Buriak keywords has some words in there that ground her research 'what it's good for', but Davis's doesn't…and it should, and could. So perhaps your *manipulation* of some data sources looks for high frequency non-technical terms as possible keywords.

Other (maybe?) useful resources

- Word frequency data:
  - See https://en.wikipedia.org/wiki/Word_lists_by_frequency, and within that, see
    The American Heritage Word Frequency Book (Carroll, Davies and Richman, 1971)
    The Brown (Francis and Kucera, 1982) LOB and related corpora
- Wordnet
  - http://wordnet.princeton.edu

- Prof Stroulia also told me she has URLs for 'official' indexing word sets and taxonomies, used by specific disciplines. Ask her about these.

Renée Elio, scigri@ualberta.ca