COMP38120 Workshop 6: Enhancing search

1. Acronyms

Many people use acronyms when searching, which can lead to many highly irrelevant results. Using http://www.nactem.ac.uk/software/acromine/ enter in the Fullform text box the string:

substance abuse

and hit search. Note the acronym returned. Now clear the fullform text box and enter the acronym in the abbreviation textbox. Investigate a few results. Try searching for the following acronyms: AD, AMA, SOAP, APRIL. Then try searching for CRE, then CREB.
From your brief observations, what can you say in relation to the following:

- The role of acronym length
- The complexity of acronym formation
- What cues have been used to match full forms with their acronyms in this system.

Acromine was initially trained on MEDLINE. However, this version is essentially a dictionary for demo purposes. It has also been trained on the major life sciences archive of full text articles Europe PubMed Central, so you can see how it is used in an actual search engine. Connect to: http://labs.europepmc.org/
and try out the Acromine-based search from that page. Use the same example acronyms as before. Clicking on suggestions (right of window) will narrow the search. Note how variant forms are used to do query expansion (the query in the text box will change to include up to the 7 most frequent variants in an OR query).

2. Semantic search

2.1. History of Medicine

We recently worked with historians of medicine interested in what is known as the "modern epidemiological transition", and produced a semantic search system for them to allow search over London Medical Officer of Health reports (1848–1972) and the entire archive of the British Medical Journal (BMJ) (from 1840 onwards). This involved processing of OCR results for early material, identification of named entities (conditions (diseases), anatomical ,and biological entities, signs and symptoms, environmental entities, therapeutic or investigational entities, etc.) and also events for causality and affecting. All of this information was then indexed. The search interface allows a user to drill down and to refine a search.

Project URL: http://www.nactem.ac.uk/MHM/
Demo URL: http://nactem.ac.uk/hom/
Username: students
Pwd: specialOnes77

Important note: **do not** communicate these login details to anyone outside this class, as we must respect BMJ licensing restrictions.

You'll try a guided example or two first then the group can explore how best to find relevant items answering an information need. Try this (after logging in, as otherwise you will not be able to see the full articles):

Note that the collection contains >385 thousand full text documents. These cover a date range from 1840 to the present.

Click on "Refine query" and then on "Term" and input "tuberculosis" (without the quotes). In the term cloud, click on "phthisis", which is an old term for tuberculosis. Any mis-spellings you see are there in the original sources, however note that the term cloud has clustered them with 'proper' terms. How? By using distributional semantics (vectors containing terms in their contexts). Other processing has brought together historical variants. Why? Well, a user may not know the older, no-longer-in use terms, so would otherwise miss finding relevant documents. The graph at the top of the screen will show the occurrence of these two terms over time: you can see how "phthisis" dropped out of use and how "tuberculosis" replaced it.

Click on "New search". During 1918–1919, influenza killed more people (20–40 million) than the First World War. Scientists and doctors struggled to understand how people became infected and how they could be protected.

Refine your search to cover the publication span 01/01/1918 to 31/12/1919. Enter the Term "immunity", then select Entity/Condition = "influenza". Have a look at the document "Influenza among posion gas workers" (should be the 4th hit if you have followed the above). In the document, in the right-hand frame, you will see 4 "Affect" events have been found. Open these and inspect "Effect on influenza infection" where you will see the event trigger and its argument(s). Click on the event, then page down in the main pane to see it highlighted. When you read the sentence containing this event, you may be surprised to learn what can apparently protect you from influenza (Don't try this at home). We took a few steps to find that.

Now use the back button, and delete ('x') the term and the condition, to leave the date range. Refine the search and now click on "Event". Select "Affect" to get an event template. In the "Target" field, enter "influenza infection" and search. This is essentially asking "*What* affects influenza infection?" as you have left the cause field blank. This time, you will be taken straight to the single document with the instance of that event.

See if you can now in your group satisfy this information need via this system:

What was a) the main suspected cause of influenza at that time (1918–1919) and b) a cause that was just beginning to be investigated? (NB: the goal here is to find out what scientists publishing *at that time* thought, not what was subsequently discovered decades later.) Thinking in terms of conventional search will take you some time to plough through many results. ThInking in terms of a semantic search should get you a small number of relevant results in a few minutes.

2.2. Finding associations

Once you extract information, you can mine it for associations. FACTA+ allows you to do this, and lets you find direct associations (already known and recorded in the literature) but also, crucially, *indirect associations* (known by no-one, never explicitly recorded in the literature). Essentially, for the latter, it carries out inferencing of the type "if A is related to B and B is related to C then A is related to C", but does so in addition probabilistically.

Text based system (with link to detailed help): http://www.nactem.ac.uk/facta/
Visualizer (help only on visualizer itself): http://www.nactem.ac.uk/facta-visualizer/

Use the visualizer for now.

Let's see what was lying around in the literature undiscovered concerning Parkinson's disease. An article published recently describes a discovery that an increase in an enzyme (PLK2) can be neuroprotective, as it reduces the effect of a protein (alpha-synuclein) that is implicated in Parkinson's disease. FACTA+ currently runs over the 2012 version of MEDLINE, so this discovery was not known then. In the visualizer, click on the "indirect associations" tab. Enter "PLK2" in the text box. Select as pivot concept "Gene/Protein" and as target concept "Disease", select "up to 200 targets" and click Search. Then click on "3" for 3rd set of results. You will see alpha-synuclein in the left frame. If you hover over it, you will then see a thick link to PARK1 (one of the forms of Parkinson's). The link thickness indicates importance. Hovering over PARK1 reveals a large difference between "expected information" and (new) "information" (this is the "information of surprise" in statistical terms), indicating this is a significant relationship. It does not tell us the nature of the association, but this is in effect telling you that there is something here worth investigating: this is hypothesis generation, in other words. Thus, this association was actually *knowable before the experiment reported* in the recent paper. Studies have shown that some associations could have been found up to 10 years before an experiment eventually found them, if text mining had been applied.

Use FACTA+ indirect associations search to find up to 10 interesting associations among the following:

Query: caffeine
Pivot: gene/protein
Target: disease

Decide what you would take as a significant difference in scores for Expected Info and Info.
Just concentrate on diseases you recognize. Report your findings as {caffeine, disease} pairs.

Keep the same pivot and target but now give
Query: nicotine
and repeat the task.

2.3. Fact finding

Let's see what else can be done if we apply named entity recognition and deep syntactic analysis of sentences. We've done this for Europe PubMed Central, with EvidenceFinder:
URL: http://labs.europepmc.org/evf

Query: interleukin-1

Ignore any "No citations" messages, this is due to recent ongoing and frequent changes in a Web service at the EBI that this application uses, we will update at our end when that WS becomes stable. In any case, we are interested here not in abstracts (to which the "No citations" refers) but in full text.

Explore the questions generated from the indexed facts by clicking on them in the right scrollable box.
Discuss to what extent the retrieved snippets appear to be good answers (given you are not biologists).
What kinds of mistakes are made in the analysis? How could they be prevented?

If time: try querying for: cyclic AMP

(Next page: if we have time or we decide this would be more interesting)

3. Named entity recognition

Work in your group to see what it would take to recognize some named entities.
Consider the following newswire text:

> Aberdeen, Scotland, 21st February, 2013. /GlobalOilNews Daily/
> Oil Rig Catches Fire In North Sea
> By T. Henry Stoneturner
> Ms. Senga MacBeth, a Dansk-ScotOil spokesperson, announced today at 07:00 GMT that its semi-submersible drilling rig, Forager-23, had caught fire at 20:15 GMT yesterday, some 200 km NE of Aberdeen. A 10-man fire and rescue team led by Capt. Magnus Thorhammer has arrived by Wave Hopper helicopter to tackle the blaze. Ms. MacBeth said that two Dansk-ScotOil employees, Olivier d'Harcourt-Bréville, Chief Geological Scientist and Martin "Muddy" Waters, the well-known drilling mud viscosity expert, had lost their lives in the initial explosion. According to MacBeth, a NorthSeaLines ferry, the Island Princess has taken 36 survivors on board and was last reported to be 80 km from Aberdeen Harbour. The share price of Dansk-ScotOil (LSE:DKSO) fell by 3% in early trading on the London Stock Exchange. A former Dansk-ScotOil employee, Max Grumpie, claimed the fire was caused by defects in the EzyFlowMaster device installed last year. Both Dansk-ScotOil and EzyFlow (UK) refused to comment. The Offshore Health and Safety Agency has launched an inquiry into the incident.

Instances of the following named entities are to be identified:

PERSON, LOCATION, DATE, TIME, ARTEFACT, TRANSPORT, COMPANY, ORGANISATION, DISTANCE, INTEGER_QUANTITY, LENGTH_UNIT, PERSON_TITLE, JOB_TITLE

i) For a ***rule-based analyser***, what entries would you add to a gazetteer (dictionary) to aid recognition of the desired named entities? Assume that a gazetteer entry specifies for a lookup token its named entity type and that this named entity type may or may not be different to those mentioned above.

ii) For a ***rule-based analyser***, what patterns (including contextual clues) would you use to help you write rules to identify the maximum number of instances of the above named entities in the text? For each pattern you specify, state which instances it would match. You may specify patterns informally (e.g., *"Lake" + one or more capitalised tokens = LOCATION*). If you find it useful to introduce other named entity types to aid your analysis, do so. Note any problematic aspects of the text that may cause your patterns to recognise too much, too little, or nothing, in certain cases.