Group 100
11-791 Project

Milestone 2 Progress Report:
ANALYZE THIS!

Joseph Chang, Nikolas Wolfe, Di Xu, Prajwal Yadapadithaya

Project Progress

- Implemented Web Cache
- Completion of Snippet extraction.
- Error analysis of tasks done so far
  - Adding Team 1's 'Year hack'...
- Crazy Stuff
  - Query Normalization
  - Query Expansion using PRF
- Milestone-3 discussion

Web Cache Proxy - Moar DP!

- Public interface has not changed for the API proxy, but there is now a *transparent cache* in between API clients and the actual web service
    - Clients have no idea whether they are hitting the Cache or the web...
    - But they have their suspicions! Runtime:
        - before: **~20.1 - ∞ s**
        - after: **~1.4 - 1.6 s**
- Proxy is now **composed** with “full snippet” API
    - Clients are using 2 separate APIs and transparent cache - *Shh!* They have. No. Idea.

Time For Error Analysis!

Because….

Results From Last Week

Query Type	**MAP**	**GMAP (ε=0.01)**
Documents	0.1038	0.0200
Concepts	0.1961	0.0572
Triples	0.000	0.0100

Results For THIS Week

(Plus '2012/2014 Year Hack' from Team 1 - *not actually a hack, btw…*)

Query Type	**MAP**	**GMAP (ε=0.01)**
Documents	0.1265 (+21.57)	0.0230 (+15%)
Concepts*	0.4056 (+106.8%)	0.2698 (+371.6%)
Triples**	0.0119 (From 0.00)	0.0134 (+34%)

* *year hack/trick*
** *considering **partial** matches*

Note the Citation…

We mentioned where we got the previous idea to hacktastically change the year in the ‘gold standard’ data…

We won’t call you by name, but
to the Team who forked our code…
Citation / Acknowledgements, please.

*After all, we did your work for you. *wink!**

Error Analysis

Concepts:

- Gold standard data is wrong (year 2012)
- Query Normalization
  - Punctuation, Spacing, Casing, Lemmatization
  - Difficult to evaluate when you can’t get anything back from the WebAPI
- Pseudo-Relevance Feedback
  - Take the results that are initially returned from a given query and to use information about whether or not those results are relevant to perform a new query
  - Recursively iterate query with PRF
  - Difficult to evaluate when you can’t get anything back from the WebAPI

Error Analysis

Concepts:

- Pseudo-Relevance Feedback

  Example (Initial Query):

  Is Rheumatoid Arthritis more common in men or women?

  Using Top-10 ranked Concept results, stemming, stop-word removal, lower-casing, and word frequencies we obtain:

  chain woman medium rheumatoid acid arthritis fatty syndrome synthase thioesterase serpin acyl juvenile man felty

  Iterating again in this way we get, oh wait, nothing... Web API fail.

Error Analysis

Triples

- The results retrieved from the API is not complete for any of the triples.
- Eg: <s>null</s><o>biological_process</o><p>http://linkedlifedata.com/resource/geneontology/namespace</p> => Here subject is null.
- This is the reason for MAP = 0.00 for triples
- Solution: Partial evaluation of triples.

Error Analysis

Snippets

Similarity Metric	F1-Score*
Cosine	0.03675
Jaccard	0.03387
Dice	0.03387
Average	0.03387

*excluding cases where the gold standard does not have a solution

Interesting UIMA bug

if you do

int[] a = new int[Integer.MAX_VALUE-Integer.
MIN_VALUE];

You will not get an exception. Instead, UIMA will simply let you return from the stack of the function call as nothing has ever happened!

Thanks!

Questions?