BioCADDIE Planning

7/25

- Garrick:
  - Testing plugin
  - Looking at document prior stuff
    - Query independent – researched and tested using client-side script (more complicated query)
      - Need to store priors ahead of time
      - Tested, works fine, is pretty easy
    - Query dependent priors
      - Requires an initial retrieval
      - Looking at plugins
      - "painless" hah.
      - Have code that's suggestive
  - Hasn't run document expansion stuff
- Mike:
  - Move elastichsearch repo
  - ES integration tests workingish
  - Testing finalization/documentation
  - Re-uploaded Indri/Maven artifacts and will push the final artifacts out
  - Snapshot ir-utils, es plugin, biocaddie
- Craig
  - Shutdown SDSC instances, move data
  - Release
  - Final documentation/report

7/18

- Contract ends 7/30
- What's left
  - ElasticSearch plugin – move repo (Mike)
  - Testing – at least a manual test plan, automated would be great (Mike)
  - PubMed ingest process (Craig)
  - biocaddie + plugin repo release (Craig)
  - Collect all data in place
  - Documentation/presentation
- Bonus
  - Parallel documentation
  - Kubernetes review
  - Publish data?
  - Doc expansion on OHSUMED + Genomics (Garrick)
    - Also PubMed expansion (Craig)
  - "Priors" – if we wanted to implement priors in Lucene/ElasticSearch, how would we?

7/11

- Mike at PEARC this week; Thuong's last week
- Final deliverables:
  - ElasticSearch plugin (NDS-868) and test process (NDS-956)
  - PubMed ingest process (new)
  - biocaddie repo release
  - Documentation/whitepaper
    - Results of comparative evaluation
    - Indri v Lucene
    - Baselines
    - BM25, BM25+Rocchio, BM25+PubMed Rocchio
- Others
  - Kubernetes + parallel
  - Publish data?
- Report/paper points (ECIR/10-16-17;
  - BioCADDIE
    - Baseline results
    - Query expansion and document expansion results
    - Indri > Lucene/ElasticSearch
      - Lucene's models aren't valid
      - No built-in query expansion
      - Limitations of the real-world search engine
    - Test collection
      - Train v test
      - Short v orig
    - Query characterization and QPP
  - Other collections
    - OHSUMED/TRECDS?/Genomics

    - Infrastructure
        - ir-tools/Maven
        - Cross-validation
        - Kubernetes/parellel

6/27

- Thuong's last day ~7/15; Garrick out next week; Craig out all next week
- Open discussion/status:
    - Garrick: focusing on query expansion/Rocchio; how to make a plugin
    - Mike: stress testing on Gluster/Kubernetes for BioCaddie; 4 large nodes;
    - Thoung: re-ran baselines with test queries only; updated results; ran TREC Genomics 2006/7 baselines; compared to official results; started looking at Lucene baselines; runquery/mkeval/compare generalization
    - Craig: merged LuceneRunQuery with 6.6 support; preliminary Rocchio implementation based on Garrick's work; QPP
- Revisit statement of work and task status (BioCADDIE)
    - What we've done:
        - Comparative evaluation of RM and Rocchio using BioCADDIE test collection
        - Comparative evaluation of SDM
        - Decided what to implement (ElasticSearch plugin, Rocchio expansion)
    - Still need to do
        - Implement actual plugin
        - Implement PubMed OA index and ingest process (ElasticSearch)
        - Testing (test plan, integration, performance, execution)
        - Release packaging (in progress)
        - Documentation
    - What we can't do
        - Analysis with respect to current pipeline (we never got it running)
    - What we did that wasn't on the SOW
        - Comparative evaluation with CDS, OHSUMED, Genomics
        - Document expansion
        - Train/test analysis
        - Query performance prediction
- Review "test" results + Genomics results
    - A few open questions (why OKAPI is so bad on 2007; why 2006 results are better for LM than 2007)
- Remaining priorities
    - From SOW
        - Create ES plugin (~~NDS-868~~ - Implement Rocchio ElasticSearch plugin RESOLVED)
            - Mike had an early prototype
              ~~NDS-840~~ - Prototype a simple ElasticSearch plugin to extend the REST API RESOLVED
            - Garrick implemented Rocchio/BM25 for Lucene (
              ~~NDS-829~~ - Investigate query expansion in ElasticSearch RESOLVED)
            - We have a rudimentary example, but now we need to implement.
        - Create ElasticSearch index for PubMed (NDS-876)
        - Lucene baseline runs: Use LuceneRunQuery to run baselines for biocaddie (NDS-949)
        - Lucene Rocchio runs: Once reviewed/merged, use LuceneRunQuery for Rocchio baselines for biocaddie
        - Testing (Mike?)
        - Release
        - Documentation
    - Other
        - Create ElasticSearch index for Wikipedia
        - Lucene baseline runs: Use LuceneRunQuery to run baselines for other collections
        - Lucene Rocchio runs: Once reviewed/merged, use LuceneRunQuery for Rocchio baselines for other collections
        - Audit/cleanup results: Review everything we've done, make sure we've run all models we want to
        - Finalize QPP analysis
        - Revisit repository priors

6/20

- Sprint 27 extended until June 23
- ElasticSearch 1.7.5: plugin framework not working, will implement with newer ElasticSearch version for BioCADDIE deliverable.
- Train/test query analysis, rerunning test queries only (NDS-939)
- Rocchio expansion with Lucene
- Query performance prediction/adaptive feedback
- TREC Genomics baseline

6/13

- Sprint 28 extended until June 23
- Craig in Seattle
- Dirichlet scorer
    - Lucene does not support true language modeling. Index structure is designed for TFIDF/BM25
    - We will abandon LM in Lucene and focus on Rocchio expansion

- CDS/OHSUMED analysis

6/8/2017

- Mike is on vacation
- Craig in Seattle next week
- Dirichlet scorer (NDS-914)
    - Dense to get through
- Boolean retrieval (NDS-912)
    - Surprising result: RM3 did reasonably well
    - Not pursue
- TREC-CDS (NDS-917)
    - Why does OKAPI do so poorly?
    - RM3 is just as expected
    - Conclusion:
- OHSUMED (NDS-929)
    - Surprising that LM is lower
    - RM3 is better
    - No judged non-relevant
    - Why is TFIDF so much better?
- Query performance prediction
    - Craig to send QPP papers
- Query characterization
    - Garrick:
        - There are a couple of queries that are really similar – look at query pairs
    - Error analysis
- Sprint 27 tasks
    - Differences in Qrels for example/test queries, we haven't looked at it
        - Analysis of variance of scores for example/test
    - Error analysis
    - More on query characterization
    - More on QPP
    - More on Lucene

5/25/2017

Notes from NDS/BioCADDIE team meeting. This meeting is primarily to plan for the next sprint. The following are up for discussion:

- Evaluation framework -- where should we go from here?
    - Clean-up/prune ir-utils
    - Lucene-centric evaluation (lucene4ir)
    - Improving the shell-script approach (balance understandability/simplicity with scale)
    - Possible tasks:
        - Tie breaking
        - Retrieval models without rescoring
            - Hack Indri or extend Lucene
        - Extend Lucene
            - Dirichlet + TwoStage
            - RM/RM3
            - Is it KL
            - PLM
            - LDA
            - Kmeans
            - Handling priors
            - CER
- Distributed evaluation (Kubernetes)
    - Mike has a prototype working with hyperkube
    - Comment about missing Okapi expansion
    - Possible tasks:
        - Test on a real cluster via deploy-tools (NDS-hackathon project)
        - Provision attached storage for each node (already done with deploy-tools?)
        - How can we get data to and from all of the nodes (for prototype, manual is fine). Ideally, something similar to hdfsput hdfs get from hadoop.
        - Garrick: qrels/topics?
        - Explore AWS/GCE/Azure?
- ES RM plugin
    - Possible tasks:
        - 1.7.5 support!! (NDS-897)
        - Actually implement the plugin (NDS-868)
        - Custom scoring exploration (Garrick)
- Stemming in ES (NDS-885)
    - Create index both stemmed (Snowball) and unstemmed
- VM resources:
    - SDSC vs NCSA
    - Shared data directories
- Performance characterization (recommended by Kirk)
- New ideas?

    - Boolean/"sufficient" query - (Garrick)
        - Boolean queries in Indri Queries: scoreif
    - Structured search (using the document structure somehow)
    - Try other collections (UMLS/MeSH, medical subsets)
    - Analyze relevance judgments
    - Compare baselines against medical collections
        - TREC CDS – uh, this is the PubMed Open Access collection...
        - CLEF eHealth
        - OHSUMED
    - Cluster-based expansion models
    - Query performance prediction

Sprint 27 tasks

- Thuong:
    - Finalize stemming work
    - TREC-CDS baseline runs
    - Boolean/sufficient-query runs
- Garrick
    - Boolean/sufficient-query runs
    - Lucene Dirichlet implementation
    - Custom scoring exploration
    - QPP
    - ir-utils cleanup
- Craig
    - LOOCV tie-breaking
    - Output performance characterization
    - ir-utils evaluation framework
- Mike
    - 1.7.5 plugin support (NDS-897)
    - Implement RM plugin (NSD-868)
    - Distributed evaluation on real cluster (NDS-hackathon)
    - Define process for copying index data to nodes. Ideally, similar to hadoop fs put
    - Explore running on AWS/GCE or Azure

5/23/2017

Notes from BioCADDIE core developer meeting

- Presented status update
- BioCADDIE is running ES 1.7.5 in production, but more recent versions in development
- Xiaoling emailed results from DataMed system for full test collection in TREC format.
- Kirk suggested that we look at a fallback strategy – use one model for higher precision, another for long tail
    - When does it work? What queries does it work for?
    - Better characterization of what's working
    - DataMed is a P@20 system, mainly
- Gerard? has installed the current pipeline and will document. Maybe we can do the same.