TREC Genomics Baselines

1. Data (/shared/trecgenomics/data)

#download TREC Genomics data

```
mkdir –p /shared/trecgenomics/data
cd /shared/trecgenomics/data
wget http://skynet.ohsu.edu/trec-gen/data/2006/documents/ajepidem.zip
wget http://skynet.ohsu.edu/trec-gen/data/2006/documents/ajpcell.zip
…
```

(total 59 files to be downloaded)

#unzip TREC Genomics data

```
unzip '*.zip'
```

#convert data into trec format

```
cd ~/biocaddie/scripts
./trecgenomics2trec.sh
```

***#documents=**162259**

Output: **/shared/trecgenomics/data/trecText/trecgenomics_all.txt**

Also make a copy at **/data/trecgenomics/data/**

2. Indexes (/shared/trecgenomics/indexes/trecgenomics_all)

Index param file: **~/biocaddie/index/build_index.trecgenomics.params**

#Content

```
<parameters>
  <index>/shared/trecgenomics/indexes/trecgenomics_all</index>
  <indexType>indri</indexType>
  <corpus>
    <path>/shared/trecgenomics/data/trecText/trecgenomics_all.txt</path>
    <class>trectext</class>
  </corpus>
</parameters>
```

#Build index

```
mkdir -p /shared/trecgenomics/indexes/
cd ~/biocaddie
IndriBuildIndex index/build_index.trecgenomics.params
```

Output is saved at **/shared/trecgenomics/indexes/trecgenomics_all**

Also make a copy at **/data/trecgenomics/indexes/trecgenomics_all**

3. Queries

#download topics to /shared/trecgenomics/queries folder

```
wget http://skynet.ohsu.edu/trec-gen/data/2007/2007topics.txt
```

#convert query into trec format (use trecgentopics2trec.sh to create queries.combined.orig)

```
cd ~/biocaddie
scripts/trecgentopics2trec.sh      #for 2007 queries
scripts/trecgentopics2trec2006.sh  #for 2006 queries
```

Output is saved at **/shared/trecgenomics/queries**

Also make a copy of the query at **/data/trecgenomics/queries**

4. Qrels

2007 qrels:

#download qrels to /shared/trecgenomics/qrels folder

```
wget http://skynet.ohsu.edu/trec-gen/data/2007/trecgen2007.all.judgments.tsv.txt
```

#convert qrels into correct format for trec_eval (add in 0 in second column, replace NOT_RELEVANT with 0 and RELEVANT with 2, remove columns 4 and 5)

```
grep -v "#" /shared/trecgenomics/qrels/trecgen2007.all.judgments.tsv.txt | sed -e 's/\tRELEVANT/\t2/g' -e 's
/\tNOT_RELEVANT/\t0/g' -e 's/\t/\t0\t/1' | cut -f 1,2,3,6 > trecgenomics-qrels-2007.txt
```

2006 qrels:

#download qrels to /shared/trecgenomics/qrels folder

```
wget http://skynet.ohsu.edu/trec-gen/data/2006/topics/2006topics.txt
```

#convert qrels into correct format for trec_eval (add in 0 in second column, replace NOT with 0, POSSIBLY with 1 and DEFINITELY with 2, remove columns 4, 5 and 6)

```
grep -v "#" /shared/trecgenomics/qrels/trec2006.raw.relevance.tsv.txt | sed -e 's/\tDEFINITELY/\t2/g' -e 's
/\tPOSSIBLY/\t1/g' -e 's/\tNOT/\t0/g' -e 's/\t/\t0\t/1' | cut -f 1,2,3,7 > trecgenomics-qrels-2006.txt
```

*****Problem with TREC Genomics qrels.**

The relevant judgements generated above contain duplicate values such as a document for a query might have multiple judgements (RELEVENT/NON-RELEVANT) based on the document's maximum-length span.

Eg: In trecgen2007.all.judgments.tsv.txt file:

```
200 9063387 2059 1870 NOT_RELEVANT
200 9063387 7300 1702 RELEVANT
200 9063387 58122 4989 NOT_RELEVANT
200 9063387 82135 1426 RELEVANT
200 9063387 83588 3235 RELEVANT
200 9063387 97901 27036 NOT_RELEVANT
```

In trecgenomics-qrels.txt:

```
root@integration-1:/data/trecgenomics/qrels# grep 9063387  trecgenomics-qrels.txt
200     0       9063387 0
200     0       9063387 2
200     0       9063387 0
200     0       9063387 2
200     0       9063387 2
200     0       9063387 0
```

To fix this problem, use Rscript **trecgenqrels.R** (in ~/biocaddie/scripts), this script will group by query & document number and sum up the relevant number. If sum=0 -> document is non-relevant, its relevant number is kept 0; if sum>=2 -> document might include multiple relevant and non-relevant judgements, so we assign its relevant number to 2.

Output file is **trecgenomics-qrels-nondup-<year>.txt** and saved at **/shared/trecgenomics/qrels**

Also make a copy of the qrels at **/data/trecgenomics/qrels**

5. IndriRunQuery - Output

```
cd ~/biocaddie/baselines/trecgenomics
./<model>.sh <topic> <collection> <year> |parallel -j 20 bash -c "{}"
```

Eg:

./jm.sh orig combined 2006| parallel -j 20 bash -c "{}"
./dir.sh orig combined 2006| parallel -j 20 bash -c "{}"
./tfidf.sh orig combined 2006| parallel -j 20 bash -c "{}"
./two.sh orig combined 2006| parallel -j 20 bash -c "{}"
./okapi.sh orig combined 2006| parallel -j 20 bash -c "{}"
./rm3.sh orig combined 2006| parallel -j 20 bash -c "{}"

IndriRunQuery outputs for different baselines are stored at:

/data/trecgenomics/output/<year>/tfidf/combined/orig

/data/trecgenomics/output/<year>/dir/combined/orig

/data/trecgenomics/output/<year>/okapi/combined/orig

/data/trecgenomics/output/<year>/jm/combined/orig

/data/trecgenomics/output/<year>/two/combined/orig

/data/trecgenomics/output/<year>/rm3/combined/orig

6. Cross-validation

```
cd ~/biocaddie
scripts/mkeval_trecgenomics.sh <model> <topics> <collection> <year>
```

Eg: scripts/mkeval_trecgenomics.sh tfidf orig combined

7. Compare models

```
cd ~/biocaddie
Rscript scripts/compare_trecgenomics.R <collection> <from model> <to model> <topic> <year>
```

Results (compared to tfidf baseline)

2007 data

Model	MAP	NDCG	P@20	NDCG@20	P@100	NDCG@100	Notes	Date
tfidf	0.2465	0.528	0.3361	0.4077	0.2081	0.3915	Sweep b and k1	06/23/17
Okapi	0.0666-	0.2568-	0.1389-	0.1393-	0.0953-	0.1415-	Sweep b, k1, k3	06/23/17
QL (JM)	0.2136-	0.4771-	0.3403	0.3951	0.1847-	0.3583-	Sweep lambda	06/23/17
QL (Dir)	0.2176	0.4772-	0.3514	0.4069	0.1881-	0.3576	Sweep mu	06/23/17
QL (TS)	0.2379	0.5128	0.3569	0.437	0.1986	0.399	Sweep mu and lambda	06/23/17
RM3	0.2536	0.5252	0.3653	0.4218	0.2164	0.3874	Sweep mu, fbDocs, fbTerms, and lambda	06/23/17

```
root@integration-1:~/biocaddie# Rscript scripts/compare_trecgenomics.R combined tfidf dir orig 2007
[1] "map 0.2465 0.2176 p= 0.9297"
[1] "ndcg 0.528 0.4772 p= 0.9838"
[1] "P_20 0.3361 0.3514 p= 0.2011"
[1] "ndcg_cut_20 0.4077 0.4069 p= 0.5111"
[1] "P_100 0.2081 0.1881 p= 0.9771"
[1] "ndcg_cut_100 0.3915 0.3576 p= 0.885"
root@integration-1:~/biocaddie# Rscript scripts/compare_trecgenomics.R combined tfidf two orig 2007
[1] "map 0.2465 0.2379 p= 0.7532"
[1] "ndcg 0.528 0.5128 p= 0.8973"
[1] "P_20 0.3361 0.3569 p= 0.1197"
[1] "ndcg_cut_20 0.4077 0.437 p= 0.1039"
[1] "P_100 0.2081 0.1986 p= 0.8416"
[1] "ndcg_cut_100 0.3915 0.399 p= 0.3308"
root@integration-1:~/biocaddie# Rscript scripts/compare_trecgenomics.R combined tfidf jm orig 2007
[1] "map 0.2465 0.2136 p= 0.996"
[1] "ndcg 0.528 0.4771 p= 1"
[1] "P_20 0.3361 0.3403 p= 0.4073"
[1] "ndcg_cut_20 0.4077 0.3951 p= 0.7083"
[1] "P_100 0.2081 0.1847 p= 0.9802"
[1] "ndcg_cut_100 0.3915 0.3583 p= 0.9727"
root@integration-1:~/biocaddie# Rscript scripts/compare_trecgenomics.R combined tfidf okapi orig 2007
[1] "map 0.2465 0.0666 p= 1"
[1] "ndcg 0.528 0.2568 p= 1"
[1] "P_20 0.3361 0.1389 p= 0.9999"
[1] "ndcg_cut_20 0.4077 0.1393 p= 1"
[1] "P_100 0.2081 0.0953 p= 0.9998"
[1] "ndcg_cut_100 0.3915 0.1415 p= 1"
root@integration-1:~/biocaddie# Rscript scripts/compare_trecgenomics.R combined tfidf rm3 orig 2007
[1] "map 0.2465 0.2536 p= 0.3791"
[1] "ndcg 0.528 0.5252 p= 0.5405"
[1] "P_20 0.3361 0.3653 p= 0.1006"
[1] "ndcg_cut_20 0.4077 0.4218 p= 0.3338"
[1] "P_100 0.2081 0.2164 p= 0.2333"
[1] "ndcg_cut_100 0.3915 0.3874 p= 0.5519"
```

2006 data

Model	MAP	NDCG	P@20	NDCG@20	P@100	NDCG@100	Notes	Date
tfidf	0.2714	0.4897	0.3375	0.3922	0.1725	0.3884	Sweep b and k1	06/27/17
Okapi	0.2372	0.3963-	0.2393-	0.3176-	0.1343-	0.3216	Sweep b, k1, k3	06/27/17
QL (JM)	0.2774	0.5003	0.3268	0.4354+	0.1764	0.4307+	Sweep lambda	06/27/17
QL (Dir)	0.2939	0.5196	0.3375	0.4524+	0.1836	0.4554+	Sweep mu	06/27/17
QL (TS)	0.2884	0.5153	0.3268	0.4509+	0.1861	0.4479+	Sweep mu and lambda	06/27/17
RM3	0.3415+	0.5093	0.3768	0.4521+	0.2168+	0.448+	Sweep mu, fbDocs, fbTerms, and lambda	06/27/17

```
root@integration-1:~/biocaddie#  Rscript scripts/compare_trecgenomics.R combined tfidf dir orig 2006
[1] "map 0.2714 0.2939 p= 0.1046"
[1] "ndcg 0.4897 0.5196 p= 0.0642"
[1] "P_20 0.3375 0.3375 p= 0.5"
[1] "ndcg_cut_20 0.3922 0.4524 p= 0.006"
[1] "P_100 0.1725 0.1836 p= 0.1524"
[1] "ndcg_cut_100 0.3884 0.4554 p= 0.0023"
root@integration-1:~/biocaddie#  Rscript scripts/compare_trecgenomics.R combined tfidf jm orig 2006
[1] "map 0.2714 0.2774 p= 0.3499"
[1] "ndcg 0.4897 0.5003 p= 0.2605"
[1] "P_20 0.3375 0.3268 p= 0.6859"
[1] "ndcg_cut_20 0.3922 0.4354 p= 0.0334"
[1] "P_100 0.1725 0.1764 p= 0.3434"
[1] "ndcg_cut_100 0.3884 0.4307 p= 0.0157"
root@integration-1:~/biocaddie# Rscript scripts/compare_trecgenomics.R combined tfidf two orig 2006
[1] "map 0.2714 0.2884 p= 0.1384"
[1] "ndcg 0.4897 0.5153 p= 0.0823"
[1] "P_20 0.3375 0.3268 p= 0.6934"
[1] "ndcg_cut_20 0.3922 0.4509 p= 0.0062"
[1] "P_100 0.1725 0.1861 p= 0.1425"
[1] "ndcg_cut_100 0.3884 0.4479 p= 0.0032"
root@integration-1:~/biocaddie# Rscript scripts/compare_trecgenomics.R combined tfidf okapi orig 2006
[1] "map 0.2714 0.2372 p= 0.8888"
[1] "ndcg 0.4897 0.3963 p= 0.9909"
[1] "P_20 0.3375 0.2393 p= 0.9909"
[1] "ndcg_cut_20 0.3922 0.3176 p= 0.953"
[1] "P_100 0.1725 0.1343 p= 0.9559"
[1] "ndcg_cut_100 0.3884 0.3216 p= 0.9465"
root@integration-1:~/biocaddie# Rscript scripts/compare_trecgenomics.R combined tfidf rm3 orig 2006
[1] "map 0.2714 0.3415 p= 0.0136"
[1] "ndcg 0.4897 0.5093 p= 0.2485"
[1] "P_20 0.3375 0.3768 p= 0.0741"
[1] "ndcg_cut_20 0.3922 0.4521 p= 0.0199"
[1] "P_100 0.1725 0.2168 p= 0.0189"
[1] "ndcg_cut_100 0.3884 0.448 p= 0.0206"
```

Comments:

- TREC Genomics collection is a full-text collection (different from bioCaddie which is descriptive metadata collection). It consists of full-text HTML documents from 49 journals published via Highwire Press. Hence, each document's text is much longer.
- Topics used in TREC Genomics collection are common queries and quite similar to bioCaddie original queries.
  Eg:
  <200>What serum [PROTEINS] change expression in association with high disease activity in lupus?
  <201>What [MUTATIONS] in the Raf gene are associated with cancer?
- Relevant judgements contain judgements for different passages of a document (RELEVANT or NON-RELEVANT). Some documents can be divided into multiple passages of different length and can have different judgement for each passage. However, in our baselines run, we use the judgement for the whole document; hence if a document has one or more relevant passages, it is considered RELEVANT.
  Eg:
  200 10090921 10160 2221 RELEVANT
  200 10090921 12404 720 NOT_RELEVANT
  200 10090921 13147 1084 NOT_RELEVANT
  200 10090921 59180 515 RELEVANT
  200 10090921 101717 349 RELEVANT
  Document number **10090921** is considered **RELEVANT** as it has at least 1 RELEVANT passage.

- Baselines run results:
  2007:
  - Okapi significantly performed worse than tfidf in all metrics (no queries failed to run)
  - Query Likelihood baselines did not show significant improvements compared to TFIDF (few metrics were even worse)
  - RM3 did yield better results but the data did not provide a significant improvement compared to TFIDF baselines.
  - MAP results are above mean and median of official results of TREC 2007 Genomics Track (whose MAP with mean 0.1862, median 0.1897)
  2006:
  - Okapi significantly performed worse than tfidf in most of the metrics (but not such huge difference like 2007).
  - Query Likelihood baselines did show significant improvements compared to TFIDF for NDCG@20 and NDCG@100.
  - RM3 provided a significant improvement compared to TFIDF baselines (MAP, P@100, NDCG@20, NDCG@100)
  - MAP results are lower than mean and median of official results of TREC 2006 Genomics Track (whose MAP with mean 0.2887, median 0.3083)
  Verify the methods with high MAP metric from official results.
  UICGenRun1,2,3(~0.52-0.54): *two dimensional ranking, query expansion.*
  NLMinter (~0.47) *refined queries using combination of topic terms, query expansion (using Entrez Gene, GeneCard, MeSH)*
  THU1,2,3 (~0.43): *best result, shorter passages return, longer passage return*
  iitx1,2,3 (~0.41-0.42): *customized function*
  UIUCinter (~0.41): *interactive run*
  PCPsgRescore,Clean,Aspect (~0.42): *Combine multiple types of resources for constructing queries; Hierarchical language model smoothing; Post result filter*
  uchsc2,1,3 (~0.40 - 0.41): *Expanded queries are sent to the search engine Lemur. Results undergo zone filtering, and top remaining Lemur results are sent to a singular value decomposition algorithm to expand the results pool by selecting similar paragraphs based on a latent semantic Dirichlet similarity score. Results of the SVD are filtered using Naive Bayes with lexical and conceptual features with training data dervied from manual evaluation of Lemer output*
  NLMfusion (~0.37): *equally-weight fusion of the results of 4 automatic methods.*
  UniNE1,2,3 (~0.34-0.36): *Data fusion of two IR systems (based on normalized RSV values (Z-score), baserun for comparisons) IR system*
  UofG0,1 (~0.35-0.36): *Retrieval based on the language modeling approach. The results are further filtered based on document coverage.*

  Compare our results with similar methods used in official runs.

Method	2006 Official Run	Our run
tfidf weighting	0.2755 (**UniGe** - a vector-space with tf.idf weightings and a modified version of pivoted normalization) 0.129 (**kyoto2** - probabilistic model of term occurrence)	0.2714
Language Modelling	0.3459 (**EMCUT1,2** - Document retrieval is performed using a language-modelling approach. Passage selection is based on identification of concepts from the UMLS metathesaurus and a gene thesaurus in both the query and the documents) 0.3517 (**UofG0** - Retrieval based on the language modeling approach)	0.2774 (JM) 0.2939 (Dir) 0.2884 (TS)
Riochio relevance feedback	0.3634 (**DUTgen1, DUTgen2** - using Rochio relevance feedback based on 2005's gold standard, Two levels of indexes, BM25, reranking) 0.2964 (**UMassCIIR1** - Query-biased pseudo relevance feedback. 250 word passages with overlap removed)	0.3415 (RM3)
Okapi	0.3365 (**york06ga1** - 1. Use Okapi BM25 for concept-based structured query 2. Use the blind feedback with term selection technique 3. Use a dual index model for passage retrieval 4. No aspect-level retrieval)	0.2372