TREC CDS Baselines

1. TREC CDS 2015 collection (http://trec-cds.appspot.com/2015.html)

1.1. Data (/shared/treccds/data/2015)

#download TREC CDS 2015 data

```
mkdir –p /shared/treccds/data/2015
cd /shared/treccds/data/2015
wget http://ceb.nlm.nih.gov/~simpsonmatt/pmc-text-00.tar.gz
wget http://ceb.nlm.nih.gov/~simpsonmatt/pmc-text-01.tar.gz
wget http://ceb.nlm.nih.gov/~simpsonmatt/pmc-text-02.tar.gz
wget http://ceb.nlm.nih.gov/~simpsonmatt/pmc-text-03.tar.gz
```

#untar TREC CDS 2015 data

```
tar xvzf pmc-text-00.tar.gz --owner root --group root --no-same-owner 2>&1 >> 2015data.log
tar xvzf pmc-text-01.tar.gz --owner root --group root --no-same-owner 2>&1 >> 2015data.log
tar xvzf pmc-text-02.tar.gz --owner root --group root --no-same-owner 2>&1 >> 2015data.log
tar xvzf pmc-text-03.tar.gz --owner root --group root --no-same-owner 2>&1 >> 2015data.log
```

#convert data into trec format

```
cd ~/biocaddie/scripts
./xml2trec.sh 2015
```

Output: **/shared/treccds/data/2015/trecText/treccds2015_all.txt**

Also make a copy at **/data/treccds/data/2015**

1.2. Indexes (/shared/treccds/indexes/treccds2015_all)

Index param file: **~/biocaddie/index/build_index.treccds2015.params**

#Content

```
<parameters>
   <index>/shared/treccds/indexes/treccds2015_all</index>
   <indexType>indri</indexType>
   <corpus>
     <path>/shared/treccds/data/2015/trecText/treccds2015_all.txt</path>
     <class>trectext</class>
   </corpus>
</parameters>
```

#Build index

```
mkdir -p /shared/treccds/indexes/
cd ~/biocaddie IndriBuildIndex
index/build_index.treccds2015.params
```

Output is saved at **/shared/treccds/indexes/treccds2015_all**

Also make a copy at **/data/treccds/indexes/treccds2015_all**

1.3.Queries (/shared/treccds/queries/2015)

As topics2015A.xml and topics2015B.xml are quite similar (only few topics in topics2015B.xml have extra <diagnosis> tag, we use topics2015B.xml only and rename it to topics2015.xml

#download topics 2015

```
mkdir -p /shared/treccds/queries/2015
cd /shared/treccds/queries/2015
wget http://trec-cds.appspot.com/topics2015B.xml
mv topics2015B.xml topics2015.xml
```

#convert query into trec format (./topics2trec.sh <year>)

```
cd ~/biocaddie
scripts/topics2trec.sh 2015
```

Output is saved at **/shared/treccds/queries/2015/queries.combined.orig**

Also make a copy of the query at **/data/treccds/queries/2015**

1.4. Qrels (/shared/treccds/qrels/2015)

#download qrels 2015

```
mkdir -p /shared/treccds/qrels/2015
cd /shared/treccds/qrels/2015
wget http://trec-cds.appspot.com/qrels-treceval-2015.txt
```

Also make a copy at **/data/treccds/qrels/2015**

1.5. IndriRunQuery - Output (/shared/treccds/output/2015)

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

IndriRunQuery outputs for different baselines are stored at:

/data/treccds/output/2015/two/combined/orig

/data/treccds/output/2015/tfidf/combined/orig

/data/treccds/output/2015/dir/combined/orig

/data/treccds/output/2015/okapi/combined/orig

/data/treccds/output/2015/jm/combined/orig

/data/treccds/output/2015/rm3/combined/orig

1.6. Cross-validation

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

1.7. Compare models

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

Results (compared to tfidf baseline)

Model	MAP	NDCG	P@20	NDCG@20	P@100	NDCG@100	Notes	Date

tfidf	0.0957	0.3017	0.3	0.2353	0.1857	0.194	Sweep b and k1	06/05/17
Okapi	0.0552-	0.1831-	0.23-	0.1824-	0.128-	0.1477-	Sweep b, k1, k3	06/05/17
QL (JM)	0.0971	0.2901	0.3167	0.2708	0.1827	0.212	Sweep lambda	06/05/17
QL (Dir)	0.1026	0.3002	0.3217	0.2705	0.1937	0.2197+	Sweep mu	06/05/17
QL (TS)	0.0976	0.3001	0.295	0.2555	0.1857	0.2145	Sweep mu and lambda	06/05/17
RM3	0.156+	0.3614+	0.375+	0.2926+	0.236+	0.2595+	Sweep mu, fbDocs, fbTerms, and lambda	06/06/17

```
root@integration-1:~/biocaddie# Rscript scripts/compare_treccds.R combined tfidf dir orig 2015
[1] "map 0.0957 0.1026 p= 0.2837"
[1] "ndcg 0.3017 0.3002 p= 0.5381"
[1] "P_20 0.3 0.3217 p= 0.2352"
[1] "ndcg_cut_20 0.2353 0.2705 p= 0.0834"
[1] "P_100 0.1857 0.1937 p= 0.2639"
[1] "ndcg_cut_100 0.194 0.2197 p= 0.0396"
root@integration-1:~/biocaddie# Rscript scripts/compare_treccds.R combined tfidf jm orig 2015
[1] "map 0.0957 0.0971 p= 0.4594"
[1] "ndcg 0.3017 0.2901 p= 0.7606"
[1] "P_20 0.3 0.3167 p= 0.3204"
[1] "ndcg_cut_20 0.2353 0.2708 p= 0.1038"
[1] "P_100 0.1857 0.1827 p= 0.5918"
[1] "ndcg_cut_100 0.194 0.212 p= 0.1283"
root@integration-1:~/biocaddie# Rscript scripts/compare_treccds.R combined tfidf okapi orig 2015
[1] "map 0.0957 0.0552 p= 0.9997"
[1] "ndcg 0.3017 0.1831 p= 1"
[1] "P_20 0.3 0.23 p= 0.9979"
[1] "ndcg_cut_20 0.2353 0.1824 p= 0.9991"
[1] "P_100 0.1857 0.128 p= 0.9999"
[1] "ndcg_cut_100 0.194 0.1477 p= 0.9993"
root@integration-1:~/biocaddie# Rscript scripts/compare_treccds.R combined tfidf two orig 2015
[1] "map 0.0957 0.0976 p= 0.432"
[1] "ndcg 0.3017 0.3001 p= 0.5467"
[1] "P_20 0.3 0.295 p= 0.5667"
[1] "ndcg_cut_20 0.2353 0.2555 p= 0.2152"
[1] "P_100 0.1857 0.1857 p= 0.5"
[1] "ndcg_cut_100 0.194 0.2145 p= 0.0652"
root@integration-1:~/biocaddie# Rscript scripts/compare_treccds.R combined tfidf rm3 orig 2015
[1] "map 0.0957 0.156 p= 0.0042"
[1] "ndcg 0.3017 0.3614 p= 0.005"
[1] "P_20 0.3 0.375 p= 0.0185"
[1] "ndcg_cut_20 0.2353 0.2926 p= 0.0348"
[1] "P_100 0.1857 0.236 p= 0.0155"
[1] "ndcg_cut_100 0.194 0.2595 p= 0.0052"
```

2. TREC CDS 2016 collection (http://trec-cds.appspot.com/2016.html)

Data and query format are same as TREC CDS 2015. However, **qrels are not available**.

3. TREC PM 2017 collection (http://trec-cds.appspot.com/2017.html)

Data and query format are different compared to TREC CDS. Also, **qrels are not available.**