Run Lucene baselines

1. Variables and notations

Term	Meaning	Variable in script	Scripts
collection	name of the collection/dataset (biocaddie, ohsumed, treccds, trecgenomics)	col	all
subset	set of data used for running baselines (combined, test, train)	subset	all
topics	name of the topics file (short, orig, stopped, etc)	topics	all
year	year of the collection/dataset, available for few collections such as trecgenomics (2006, 2007)	year	all
model	retrieval model (dir, rm3, jm, pubmed, etc)	model	mkeval.sh, mkeval-lucene.sh
from model	retrieval model (dir, rm3, jm, pubmed, etc), for comparison (t-test)	from	compare.R
to model	retrieval model (dir, rm3, jm, pubmed, etc), for comparison (t-test)	to	compare.R
metric	evaluation metric (map, ndcg, P_20, ndcg_cut_20, etc)	metric	mkeval.sh, mkeval-lucene.sh, compare.R
run method	running method (indri, lucene)	run	compare.R

2. Files and their locations

Collections without year: **biocaddie, ohsumed**

Collections with year: **treccds (2015), trecgenomics (2006, 2007)**

Type	Location	Example
Indexes	**/data/<col>/lucene/<col>_all/shard0** **/data/<col>/lucene/<col><year>_all/shard0**	*/data/biocaddie/lucene/biocaddie_all/shard0* */data/trecgenomics/lucene/trecgenomics2006_all/shard0*
Queries	**/data/<col>/queries/queries.<subset>.<topics>** **/data/<col>/queries/queries.<subset>.<topics>.<year>**	*/data/biocaddie/queries/queries.test.short* */data/trecgenomics/queries/queries.combined.orig.2006*
Qrels	**/data/<col>/qrels/<col>.qrels.<subset>** **/data/<col>/qrels/<col>.qrels.<subset>.<year>**	*/data/biocaddie/qrels/biocaddie.qrels.test* */data/trecgenomics/qrels/trecgenomics.qrels.combined.2006*
Output	**/data/<col>/lucene-output/<model>/<subset>/<topics>** **/data/<col>/lucene-output/<year>/<model>/<subset>/<topics>**	*/data/biocaddie/lucene-output/dir/test/short* */data/trecgenomics/lucene-output/2006/dir/combined/orig*
Eval	**/data/<col>/lucene-eval/<model>/<subset>/<topics>** **/data/<col>/lucene-eval/<year>/<model>/<subset>/<topics>**	*/data/biocaddie/lucene-eval/dir/test/short* */data/trecgenomics/lucene-eval/2006/dir/combined/orig*
Loocv	**/data/<col>/loocv/<model>.<subset>.<topics>.<metric>.lucene.out** **/data/<col>/loocv/<year>/<model>.<subset>.<topics>.<metric>.lucene.out**	*/data/biocaddie/loocv/dir.test.short.ndcg.lucene.out* */data/trecgenomics/loocv/2006/dir.combined.orig.ndcg.lucene.out*

*** *Note: both Lucene and Indri's loocv results are saved in the same location for easy comparison across different runs.*

3. Run Lucene baselines

a) Lucene Run (lucene-output)

Using biocaddie_all indexes

```
cd ~/biocaddie
baselines/new/<model>-lucene.sh <topics> <subset> <col>| parallel -j 20 bash -c "{}"
baselines/new/<model>-lucene.sh <topics> <subset> <col> <year>| parallel -j 20 bash -c "{}"
```

Eg: baselines/new/dir-lucene.sh short test biocaddie| parallel -j 20 bash -c "{}"
baselines/new/tfidf-lucene.sh short test biocaddie| parallel -j 20 bash -c "{}"
baselines/new/jm-lucene.sh short test biocaddie| parallel -j 20 bash -c "{}"
baselines/new/bm25-lucene.sh short test biocaddie| parallel -j 20 bash -c "{}"
baselines/new/rocchio-lucene.sh short test biocaddie| parallel -j 20 bash -c "{}"

Using biocaddie_all.snowball indexes

```
cd ~/biocaddie
baselines/new/<model>-lucene-snowball.sh <topics> <subset> <col>| parallel -j 20 bash -c "{}"
baselines/new/<model>-lucene-snowball.sh <topics> <subset> <col> <year>| parallel -j 20 bash -c "{}"
```

Eg: baselines/new/dir-lucene-snowball.sh short test biocaddie| parallel -j 20 bash -c "{}"
baselines/new/tfidf-lucene-snowball.sh short test biocaddie| parallel -j 20 bash -c "{}"
baselines/new/jm-lucene-snowball.sh short test biocaddie| parallel -j 20 bash -c "{}"
baselines/new/bm25-lucene-snowball.sh short test biocaddie| parallel -j 20 bash -c "{}"
baselines/new/rocchio-lucene-snowball.sh short test biocaddie| parallel -j 20 bash -c "{}"

b) Evaluation and Cross-validation (lucene-eval, loocv)

```
cd ~/biocaddie
scripts/new/mkeval-lucene.sh <model> <topics> <subset> <col>
scripts/new/mkeval-lucene.sh <model> <topics> <subset> <col> <year>
```

Eg: scripts/new/mkeval-lucene.sh dir short test biocaddie
scripts/new/mkeval-lucene.sh tfidf short test biocaddie
scripts/new/mkeval-lucene.sh jm short test biocaddie
scripts/new/mkeval-lucene.sh bm25 short test biocaddie
scripts/new/mkeval-lucene.sh rocchio short test biocaddie

scripts/new/mkeval-lucene.sh dir-snowball short test biocaddie
scripts/new/mkeval-lucene.sh tfidf-snowball short test biocaddie
scripts/new/mkeval-lucene.sh jm-snowball short test biocaddie
scripts/new/mkeval-lucene.sh bm25-snowball short test biocaddie
scripts/new/mkeval-lucene.sh rocchio-snowball short test biocaddie

c) Compare models

We have to input running method for comparison:

0 - both from and to models are from Indri run

1 - both from and to models are from Lucene run

2 - from model is from Indri run, to model is from Lucene run

3 - from model is from Lucene run, to model is from Indri run

```
cd ~/biocaddie
Rscript scripts/new/compare.R <subset> <from> <to> <topics> <col>
Rscript scripts/new/compare.R <subset> <from> <to> <topics> <col> <year>
```

Eg: Rscript scripts/new/compare.R test tfidf dir short biocaddie

Rscript scripts/new/compare.R test tfidf-snowball dir-snowball short biocaddie

4. Results

Using biocaddie_all indexes.

Model	MAP	NDCG	P@20	NDCG@20	P@100	NDCG@100	Notes	Date
classic tfidf	0.3282	0.5824	0.6867	0.5478	0.5013	0.5018	No parameters	07/05/17
BM25	0.3543	0.6105+	0.7467+	0.5917+	0.506	0.5186	Sweep b, k1	07/05/17
QL (JM)	0.3382	0.6022	0.7233	0.571	0.5	0.4996	Sweep lambda	07/05/17
QL (Dir)	0.3675 (p-value=0.0548)	0.6163+ (p-value=0.0502)	0.6567	0.5664	0.5213	0.522	Sweep mu	07/05/17
Rocchio	0.4044+	0.6417 (p-value=0.0533)	0.6967	0.5403	0.492	0.4912	Sweep b, k1, fbTerms, fbDocs, fbOrigWeight	07/05/17

Using biocaddie_all.snowball indexes

Model	MAP	NDCG	P@20	NDCG@20	P@100	NDCG@100	Notes	Date
classic tfidf (tfidf-snowball)	0.3375	0.5944	0.6667	0.5256	0.4987	0.5002	No parameters	07/06/17
BM25 (bm25-snowball)	0.3764+	0.6239+	0.73+	0.6006+	0.5413+	0.539+	Sweep b, k1	07/06/17
QL (JM) (jm-snowball)	0.3448	0.6058	0.67	0.5813+	0.4987	0.5289+	Sweep lambda	07/06/17
QL (Dir) (dir-snowball)	0.3776+	0.6315+	0.7033	0.6006+	0.5307	0.5365+	Sweep mu	07/06/17
Rocchio (rocchio-snowball)	0.3959	0.6052	0.7267	0.598+	0.5453	0.525	Sweep b, k1, fbTerms, fbDocs, fbOrigWeight	07/06/17

Difference between unstemmed and stemmed indexes

Model	MAP	NDCG	P@20	NDCG@20	P@100	NDCG@100	Notes	Date
classic tfidf	0.3282	0.5824	0.6867	0.5478	0.5013	0.5018	No parameters	07/10/17
classic tfidf (tfidf-snowball)	0.3375	0.5944	0.6667	0.5256	0.4987	0.5002	No parameters	07/10/17
BM25	0.3543	0.6105	0.7467	0.5917	0.506	0.5186	Sweep b, k1	07/10/17
BM25 (bm25-snowball)	0.3764+	0.6239	0.73	0.6006	0.5413+	0.539+	Sweep b, k1	07/10/17
QL (JM)	0.3382	0.6022	0.7233	0.571	0.5	0.4996	Sweep lambda	07/10/17
QL (JM) (jm-snowball)	0.3448	0.6058	0.67	0.5813	0.4987	0.5289+	Sweep lambda	07/10/17
QL (Dir)	0.3675	0.6163	0.6567	0.5664	0.5213	0.522	Sweep mu	07/10/17
QL (Dir) (dir-snowball)	0.3776	0.6315 (p-value=0.0534)	0.7033+	0.6006+	0.5307	0.5365	Sweep mu	07/10/17
Rocchio	0.4044	0.6417	0.6967	0.5403	0.492	0.4912	Sweep b, k1, fbTerms, fbDocs, fbOrigWeight	07/11/17
Rocchio (rocchio-snowball)	0.3959	0.6052-	0.7267	0.598+	0.5453+	0.525	Sweep b, k1, fbTerms, fbDocs, fbOrigWeight	07/11/17

Verification

Using biocaddie_all indexes:

```
thphan@biocaddie-dev:/data/thphan/biocaddie$ Rscript scripts/new/compare.R test tfidf dir short biocaddie
Please enter run methods for comparison:
        0: both are Indri
        1: both are Lucene
        2: from is Indri, to is Lucene
        3: from is Lucene, to is Indri
1
[1] "map 0.3282 0.3675 p= 0.0548"
[1] "ndcg 0.5824 0.6163 p= 0.0502"
[1] "P_20 0.6867 0.6567 p= 0.9461"
[1] "ndcg_cut_20 0.5478 0.5664 p= 0.186"
[1] "P_100 0.5013 0.5213 p= 0.2168"
[1] "ndcg_cut_100 0.5018 0.522 p= 0.1401"


thphan@biocaddie-dev:/data/thphan/biocaddie$ Rscript scripts/new/compare.R test tfidf jm short biocaddie
Please enter run methods for comparison:
        0: both are Indri
        1: both are Lucene
        2: from is Indri, to is Lucene
        3: from is Lucene, to is Indri
1
[1] "map 0.3282 0.3382 p= 0.1719"
[1] "ndcg 0.5824 0.6022 p= 0.0932"
[1] "P_20 0.6867 0.7233 p= 0.0831"
[1] "ndcg_cut_20 0.5478 0.571 p= 0.145"
[1] "P_100 0.5013 0.5 p= 0.5301"
[1] "ndcg_cut_100 0.5018 0.4996 p= 0.5552"


thphan@biocaddie-dev:/data/thphan/biocaddie$ Rscript scripts/new/compare.R test tfidf bm25 short biocaddie
Please enter run methods for comparison:
        0: both are Indri
        1: both are Lucene
        2: from is Indri, to is Lucene
        3: from is Lucene, to is Indri
1
[1] "map 0.3282 0.3543 p= 0.0846"
[1] "ndcg 0.5824 0.6105 p= 0.0148"
[1] "P_20 0.6867 0.7467 p= 0.0491"
[1] "ndcg_cut_20 0.5478 0.5917 p= 0.0496"
[1] "P_100 0.5013 0.506 p= 0.428"
[1] "ndcg_cut_100 0.5018 0.5186 p= 0.2195"

thphan@biocaddie-dev:/data/thphan/biocaddie$ Rscript scripts/new/compare.R test tfidf rocchio short biocaddie
Please enter run methods for comparison:
        0: both are Indri
        1: both are Lucene
        2: from is Indri, to is Lucene
        3: from is Lucene, to is Indri
1
[1] "map 0.3282 0.4044 p= 0.0188"
[1] "ndcg 0.5824 0.6417 p= 0.0533"
[1] "P_20 0.6867 0.6967 p= 0.3785"
[1] "ndcg_cut_20 0.5478 0.5403 p= 0.6276"
[1] "P_100 0.5013 0.492 p= 0.6184"
[1] "ndcg_cut_100 0.5018 0.4912 p= 0.6071"
```

Using biocaddie_all.snowball indexes

```
thphan@biocaddie-dev:/data/thphan/biocaddie$ Rscript scripts/new/compare.R test tfidf-snowball dir-snowball
short biocaddie
Please enter run methods for comparison:
        0: both are Indri
        1: both are Lucene
        2: from is Indri, to is Lucene
        3: from is Lucene, to is Indri
1
[1] "map 0.3375 0.3776 p= 0.0387"
[1] "ndcg 0.5944 0.6315 p= 0.0072"
[1] "P_20 0.6667 0.7033 p= 0.1042"
[1] "ndcg_cut_20 0.5256 0.6006 p= 0.0046"
[1] "P_100 0.4987 0.5307 p= 0.0652"
[1] "ndcg_cut_100 0.5002 0.5365 p= 0.0207"

thphan@biocaddie-dev:/data/thphan/biocaddie$ Rscript scripts/new/compare.R test tfidf-snowball jm-snowball
short biocaddie
Please enter run methods for comparison:
        0: both are Indri
        1: both are Lucene
        2: from is Indri, to is Lucene
        3: from is Lucene, to is Indri
1
[1] "map 0.3375 0.3448 p= 0.2782"
[1] "ndcg 0.5944 0.6058 p= 0.2069"
[1] "P_20 0.6667 0.67 p= 0.475"
[1] "ndcg_cut_20 0.5256 0.5813 p= 0.0161"
[1] "P_100 0.4987 0.4987 p= 0.5"
[1] "ndcg_cut_100 0.5002 0.5289 p= 0.0117"

thphan@biocaddie-dev:/data/thphan/biocaddie$ Rscript scripts/new/compare.R test tfidf-snowball bm25-snowball
short biocaddie
Please enter run methods for comparison:
        0: both are Indri
        1: both are Lucene
        2: from is Indri, to is Lucene
        3: from is Lucene, to is Indri
1
[1] "map 0.3375 0.3764 p= 0.0284"
[1] "ndcg 0.5944 0.6239 p= 0.011"
[1] "P_20 0.6667 0.73 p= 0.0331"
[1] "ndcg_cut_20 0.5256 0.6006 p= 0.0045"
[1] "P_100 0.4987 0.5413 p= 0.0326"
[1] "ndcg_cut_100 0.5002 0.539 p= 0.0149"


thphan@biocaddie-dev:/data/thphan/biocaddie$ Rscript scripts/new/compare.R test tfidf rocchio-snowball short
biocaddie
Please enter run methods for comparison:
        0: both are Indri
        1: both are Lucene
        2: from is Indri, to is Lucene
        3: from is Lucene, to is Indri
1
[1] "map 0.3282 0.3959 p= 0.0427"
[1] "ndcg 0.5824 0.6052 p= 0.2869"
[1] "P_20 0.6867 0.7267 p= 0.1189"
[1] "ndcg_cut_20 0.5478 0.598 p= 0.0424"
[1] "P_100 0.5013 0.5453 p= 0.1152"
[1] "ndcg_cut_100 0.5018 0.525 p= 0.2733"
```

Compare results between unstemmed and stemmed indexes:

```
thphan@biocaddie-dev:/data/thphan/biocaddie$ Rscript scripts/new/compare.R test tfidf tfidf-snowball short
biocaddie
Please enter run methods for comparison:
        0: both are Indri
        1: both are Lucene
```

```
        2: from is Indri, to is Lucene
        3: from is Lucene, to is Indri
1
[1] "map 0.3282 0.3375 p= 0.1463"
[1] "ndcg 0.5824 0.5944 p= 0.1454"
[1] "P_20 0.6867 0.6667 p= 0.808"
[1] "ndcg_cut_20 0.5478 0.5256 p= 0.8715"
[1] "P_100 0.5013 0.4987 p= 0.5819"
[1] "ndcg_cut_100 0.5018 0.5002 p= 0.5652"

thphan@biocaddie-dev:/data/thphan/biocaddie$ Rscript scripts/new/compare.R test dir dir-snowball short biocaddie
Please enter run methods for comparison:
        0: both are Indri
        1: both are Lucene
        2: from is Indri, to is Lucene
        3: from is Lucene, to is Indri
1
[1] "map 0.3675 0.3776 p= 0.0842"
[1] "ndcg 0.6163 0.6315 p= 0.0534"
[1] "P_20 0.6567 0.7033 p= 0.0011"
[1] "ndcg_cut_20 0.5664 0.6006 p= 0.0222"
[1] "P_100 0.5213 0.5307 p= 0.1942"
[1] "ndcg_cut_100 0.522 0.5365 p= 0.0645"

thphan@biocaddie-dev:/data/thphan/biocaddie$ Rscript scripts/new/compare.R test jm jm-snowball short biocaddie
Please enter run methods for comparison:
        0: both are Indri
        1: both are Lucene
        2: from is Indri, to is Lucene
        3: from is Lucene, to is Indri
1
[1] "map 0.3382 0.3448 p= 0.2603"
[1] "ndcg 0.6022 0.6058 p= 0.3358"
[1] "P_20 0.7233 0.67 p= 0.8885"
[1] "ndcg_cut_20 0.571 0.5813 p= 0.2551"
[1] "P_100 0.5 0.4987 p= 0.55"
[1] "ndcg_cut_100 0.4996 0.5289 p= 0.0026"

thphan@biocaddie-dev:/data/thphan/biocaddie$ Rscript scripts/new/compare.R test bm25 bm25-snowball short
biocaddie
Please enter run methods for comparison:
        0: both are Indri
        1: both are Lucene
        2: from is Indri, to is Lucene
        3: from is Lucene, to is Indri
1
[1] "map 0.3543 0.3764 p= 0.0317"
[1] "ndcg 0.6105 0.6239 p= 0.0945"
[1] "P_20 0.7467 0.73 p= 0.8548"
[1] "ndcg_cut_20 0.5917 0.6006 p= 0.2775"
[1] "P_100 0.506 0.5413 p= 0.0209"
[1] "ndcg_cut_100 0.5186 0.539 p= 0.0441"

thphan@biocaddie-dev:/data/thphan/biocaddie$ Rscript scripts/new/compare.R test rocchio rocchio-snowball short
biocaddie
Please enter run methods for comparison:
        0: both are Indri
        1: both are Lucene
        2: from is Indri, to is Lucene
        3: from is Lucene, to is Indri
1
[1] "map 0.4044 0.3959 p= 0.6841"
[1] "ndcg 0.6417 0.6052 p= 0.9667"
[1] "P_20 0.6967 0.7267 p= 0.2037"
[1] "ndcg_cut_20 0.5403 0.598 p= 0.0465"
[1] "P_100 0.492 0.5453 p= 0.0035"
[1] "ndcg_cut_100 0.4912 0.525 p= 0.0625"
```