Build Script (Team1)

Amith Ramanagar Chandrashekar Pooja Himanshu Oza
Rachel E Cates Ahmed M Alnazer

April 25, 2019

1 Dependency

Java version: Java8 or Higher

Build Tool: Maven

2 How to Run

Please perform the below steps

- sudo git clone https://gitlab.cs.unh.edu/cs953-2019/cs953-team1.git
- sudo chmod -R 777 cs953-team1
- cd cs953-team1
- ./install.sh (please run the ./install.sh in one of the tmux session as described below)

2.1 Tmux session creation

- tmux new -s team1 (create new session)
- Run the ./install.sh inside the tmux
- ctrl+b+d (Disconnect from the session)
- tmux a -t myname (To attach to the session)

2.2 Note

Running the ./install.sh creates all the runs files in the **result** directory.we have the Benchmark-Y1 test as the query file in the install script.

Also Self executable jar can be created using the maven and the program can run independently without the install script. Please follow below steps

- mvn clean compile
- mvn package
- java -jar cs953-team1-1.0-SNAPSHOT-jar-with-dependencies.jar sub-commands option

2.3 General Note

The below error can be ignored. This error is from the Nd4j.

SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.htmlStaticLoggerBinder for further details.

2.4 ranklib-normalizer

The python program is used to combine the run files into features file using the zscore normalization. All the query expansion methods described used this program to combine the scores.

When the –ranklib option is passed , it runs the ranklib to create the weight vector,this approach is during the training.

```
sudo python3 ranklib_normalizer.py
--qrelpath <qrel_path> -n --ranklib <path_to_jar> --dirpath <run1,run2,run3...>
```

When the –modelfile option is passed, it uses the modelfile , read the weight vector and uses the feature vector to generate the final score for query-document.

```
sudo python3 ranklib_normalizer.py
--qrelpath <qrel_path> -n --modelfile <model_file> --dirpath <run1,run2,run3...>
```

3 Command Line Option

These are the command line options we have implemented.

```
Usage: <main class> [command] [command options]
  Commands:
    index      Command to Index the Corpus
      Usage: index [options]
```

```
    Options:
      --abstract-index
        Perform Entity Abstract Index
        Default: false
    * -i, --corpus-file
        Corpus file to index. In case of Entity Abstract, please specify
        Entity Index location
      -d, --dest-location
        Location to save the index file
        Default: C:\Users\amith\IdeaProjects\cs953-team1\indexed_file
      --entity-index
        Perform Entity Index
        Default: false
      --help

      --para-index
        Perform Paragraph Index
        Default: false

search      Command to search
  Usage: search [options]
    Options:
      --bias-fact
        Bias factor to get the document representation
        Default: 1
      -k, --candidate-set-val
        How many candidate set to retrieve using BM25
        Default: 100
      --cluster
        Cluster Ranking
        Default: false
      --cosine-sim
        Rerank the document based on the cosine similarity between two
        strings
        Default: false
      -dbpcontain, --dbpedia-contain
        Change the search to contain
        Default: false
      -bm25, --default-bm25
        Rerank the initial retrieved cluster using document similarity
        Default: false
      --dice-sim
        Rerank the document based on the Sorensen Dice coefficient
        similarity between two strings
        Default: false
      -ecm-qe-num, --ecm-query-expansion-terms-num
```

```
  ECM  Query Expansion Terms Number
  Default: 20
-ecm, --ecm-run
  ECM Entity run file
--entity-centroid
  Rerank the passages using entity average centroid
  Default: false
--entity-default-freq
  Rerank the initial retrieved document using entity frequency
  Default: false
--entity-degree
  Rerank the initial retrieved document using entity degree
  Default: false
--entity-doc-sim
  Rerank the initial retrieved document using entity abstract
  similarity
  Default: false
--entity-ecm-expand
  Rerank the initial retrieved document using ecm entities to expand
  query
  Default: false
--entity-expand
  Rerank the initial retrieved document using expanded query
  Default: false
-f, --entity-feature
  Entity feature vector file
--entity-index
  Pass the index location of entity index
-qrel, --entity-qrel
  Entity qrel file
--entity-ranklib
  Rerank the passages using entity ranklib
  Default: false
--entity-relation
  Generate the feature vectors and ranklib model
  Default: false
--entity-sim
  Rerank the initial retrieved document using entity abstract
  similarity
  Default: false
-dbpedia, --exist-dbpedia
  Find out if an entity exist in dbpedia
  Default: false
--ham-loc
  Directory to ham train file
--help
```

```
* -i, --index-loc
    Indexed directory to search
  --jaccard-sim
    Rerank the document based on the Jaccard similarity between two
    strings
    Default: false
  --jaro-sim
    Rerank the document based on the Jaro Winkler similarity between
    two strings
    Default: false
  --leven-sim
    Rerank the document based on the NormalizedLevenshtein similarity
    between two strings
    Default: false
  --mrf
    Uses the parallel stream for the reranker methods
    Default: false
  --parallel
    Uses the parallel stream for the reranker methods
    Default: false
  --prf-val
    Top k documents to consider as Pseudo relevance feedback
    Default: 5
  --prf-val-k
    Top k terms to consider for query expansion
    Default: 50
  --prf-val-term
    Top k terms to consider per query term
    Default: 10
  --qe-entity-degree
    Query expansion based on the entity degree and reranking
    Default: false
  --qe-exp-df
    Query expansion using PRF and the terms selected using DF
    Default: false
  --qe-exp-entity
    Query expansion using PRF, also considering entity abstract
    Default: false
  --qe-exp-idf
    Query expansion using PRF and the terms selected using IDF
    Default: false
  --qe-exp-rm3
    Relevance model 3 query expansion
    Default: false
  --qe-reranking
```

```
    Query expansion method based on the Entities and reranking
    Default: false
  --qrel-path
    Pass the absolute path of the Qrel
* -q, --query-cbor
    Query file (CBOR file)
  -qe, --query-expansion
    Rerank the document using Query expansion
    Default: false
  -qe-type, --query-expansion-type
    Select type of Query expansion (entityText, entityID ,
    entityTextID)
    Default: entityText
    Possible Values: [entityText, entityID, entityTextID, entityIDInEntityField]
  --rank-lib
    Provide path to the Ranklib
  -model, --ranklib-model
    Pass the file location of ranklib model file
  --rerank
    Rerank the initial retrieved document using document similarity
    Default: false
  --rerank-df
    Rerank the document based on the DF
    Default: false
  --rerank-idf
    Rerank the document based on the IDF
    Default: false
  --spam-filter
    Uses the spam filter before performing the re-rank
    Default: false
  --spam-filter2
    Uses the spam filter before performing the re-rank
    Default: false
  --spam-loc
    Directory to spam train file
  --test
    Only for testing purposes
    Default: false
  -V, --verbose
    Print out  some of the results  into stdout
    Default: false
  -dim, --word-dimension
    Dimension of the Word embeddings
    Default: 0
  -we, --word-embedding
    Pass the word embedding file GloVe/ Word2Vec
```

```
      -top
        specify the top number of selected entity to used in the Query
        expansion
        Default: 3
      article
        Article level retrieval
        Default: false
      section
        Section level retrieval
        Default: false

indexHamSpam      Command to create training and test data for the spam
        classifier
  Usage: indexHamSpam [options]
    Options:
      --help

      -p, --paragraphs-file
        paragraph corpus directory
        Default: C:\Users\amith\IdeaProjects\cs953-team1\indexed_file
      -q, --qrels-file
        qrels file
        Default: C:\Users\amith\IdeaProjects\cs953-team1\indexed_file
      -hamSpamTest
        Location to save the ham and spam test data
        Default: C:\Users\amith\IdeaProjects\cs953-team1\indexed_file
      -hamTest
        Location to save the ham test data
        Default: C:\Users\amith\IdeaProjects\cs953-team1\indexed_file
      -hamTrain
        Location to save the ham training data
        Default: C:\Users\amith\IdeaProjects\cs953-team1\indexed_file
      -spamTest
        Location to save the spam test data
        Default: C:\Users\amith\IdeaProjects\cs953-team1\indexed_file
      -spamTrain
        Location to save the spam training data
        Default: C:\Users\amith\IdeaProjects\cs953-team1\indexed_file

--help      Help Information
  Usage: --help

ranker      Ranker
  Usage: ranker [options]
    Options:
      --mname
```

```
    Method name suffix
    Default: mrfupdated
* --model-file
    Location of the model file
* --run-file
    Location of the run file
```