COL764 , Assignment2
Mihir Gupta
Language used: Python3

1)prob_rerank.py
Executable launching

python prob_rerank.py [query-file] [top-100-file] [collection-file][expansion-limit]

query-file: file containing the queries in the same tsv format as given in Table 1 for queries file
top-100-file: a file containing the top100 documents in the same format as train and dev top100 files given, which need to be reranked
collection-file: file containing the full document collection (in the same format as msmarcodocs file given)
expansion-limit: is a number ranging from 1—15 that specifies the limit on the number of additional terms in the expanded query

Note: bashfile submitted can be edited with required files for running all code files and installing nltk.
Libraries imported :

```
import operator
import sys
import csv
import nltk
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
import math
import pickle
```

In the code BM25 model has been used to re-rank the documents.
First entire document collection has been read line by line and memory address of each documents id has been stored as offset so that its data can be later loaded.
All the queries are then read from the queries file and tokenized using nltk regex tokenizer , stop words are also removed using porter stemmer and words are converted to lower case , data for each query is stored and then the top100 file is read and list of top 100 documents for each query are stored .
For each query , data for top100 documents is loaded and then pi and ui and eventually the BM25 weights are calculated for every term in the query using the term frequencies from the document.
Value of k1 taken is 1.2 and 0.75 for b to calculate the BM25 weights from RSJ weights.
In the range expansion limit , each time score of wi(pi-ui) (wi : RSJ weight)
Is calculated for the terms in the top documents assumed to be relevant and the word with the top score is added to the query data and pi , ui are updated using Bayesian inference result in which kappa is taken as 0.9 .
After expanding 1 term at a time expansion limit times, final weights are calculated and RSVd scores are used to rank the documents and store ranks in the output file in the TRECEVAL format .

Statistical Tests

Using the following :
https://www.socscistatistics.com/tests/signedranks/default2.aspx

Wilcoxon test for NDCG

Wilcoxon Signed-Rank Test Calculator

Success!

Explanation of results

We have calculated both a *W*-value and *z*-value. If the size of *N* is at least 20 - see the Results Details box - then the distribution of the Wilcoxon *W* statistic tends to form a normal distribution. This means you can use the*z*-value to evaluate your hypothesis. If, on the other hand, the size of *N* is low, and particularly if it's below 10, you should use the *W*-value to evaluate your hypothesis.

You should also note that if a subject's difference score is zero - that is, if a subject has the same score in both treatment conditions - then the test discards the individual from the analysis and reduces the sample size. If you have a lot of ties, this procedure will undermine the reliability of the test (and also suggests that the requirement that the data is continuous has not been met).

Treatment 1	Treatment 2	*Sign*	*Abs*	*R*	*Sign R*
0	0	n/a	0	n/a	n/a
0	0	n/a	0	n/a	n/a
0	0	n/a	0	n/a	n/a
1	0	1	1	6.5	6.5
0	0	n/a	0	n/a	n/a
0	0	n/a	0	n/a	n/a
0	0	n/a	0	n/a	n/a
0	0	n/a	0	n/a	n/a
0	0	n/a	0	n/a	n/a
0	0	n/a	0	n/a	n/a
0	0	n/a	0	n/a	n/a
0	0	n/a	0	n/a	n/a
1	0	1	1	6.5	6.5
0	0	n/a	0	n/a	n/a
1	0	1	1	6.5	6.5

0 0 1	0 0 0	n/a n/a 1	0 0 1	n/a n/a 6.5	n/a n/a 6.5

Significance Level:

- ○ .01
- ◉ .05

1 or 2-tailed hypothesis?:

- ○ One-tailed
- ◉ Two-tailed

Result Details

W-value: 0
Mean Difference: 1
Sum of pos. ranks: 78
Sum of neg. ranks: 0

Z-value: -3.0594
Mean (*W*): 39
Standard Deviation (*W*): 12.75

Sample Size (*N*): 12

Result 1 - *Z*-value

The value of *z* is-3.0594. The *p*-value is .00222.

The result is significant at $p < .05$.

Result 2 - *W*-value

The value of *W* is 0. The critical value for *W* at $N = 12$ ($p < .05$) is 13.

The result is significant at $p < .05$.

Paired t-test for MRR

T-Test Calculator for 2 Dependent Means

The value of *t* is -10.121046.

Explanation of results

The output of this calculator is pretty straightforward. The values of *t* and *p* appear at the bottom of the page. If the text is blue, your result is significant; if it's red, it's not. The only thing that might catch you out is the way that we've rounded the data. The data you see in front of you, apart from the *t* and *p* values, has been rounded to 2 significant figures. However, we did not round when actually calculating the values of *t* and *p*. This means that if you try to calculate these values on the basis of the summary data provided here, you're likely going to end up with a slightly different - and less accurate - result.

Treatment 1	Treatment 2	*Diff* (T2 - T1)	*Dev* (Diff - M)	*Sq. Dev*
0	0	0	0.16	0.03
0.0108	0.0102	0	0.16	0.02
0.0213	0.0149	-0.01	0.15	0.02
1	0.5	-0.5	-0.34	0.12
0	0	0	0.16	0.03
0	0	0	0.16	0.03
0.5	0.0167	-0.48	-0.33	0.11
0.0104	0.0169	0.01	0.16	0.03
0.1111	0.0333	-0.08	0.08	0.01
0.1429	0.0294	-0.11	0.04	0
0.1667	0.0714	-0.1	0.06	0
0.1429	0.0556	-0.09	0.07	0.01
1	0.0192	-0.98	-0.82	0.68
0	0	0	0.16	0.03
1	0.5	-0.5	-0.34	0.12
0.25	0.0164	-0.23	-0.08	0.01
0	0	0	0.16	0.03
0.012	0.0385	0.03	0.18	0.03
0	0	0	0.16	0.03
0.0256	0.0189	-0.01	0.15	0.02
1	0.0204	-0.98	-0.82	0.67

Significance Level:

- ○ 0.01
- ◉ 0.05
- ○ 0.10

One-tailed or two-tailed hypothesis?:

- ○ One-tailed
- ◉ Two-tailed

Difference Scores Calculations

Mean: -0.16
$\mu = 0$
$S^2 = SS/df = 39.02/(400-1) = 0.1$
$S^2_M = S^2/N = 0.1/400 = 0$
$S_M = \sqrt{S^2_M} = \sqrt{0} = 0.02$

T-value Calculation

$t = (M - \mu)/S_M = (-0.16 - 0)/0.02 = -10.12$

The value of *t* is -10.121046. The value of *p* is < .00001. The result is significant at p < .05.

Calculate Reset

2) lm_rerank.py

Executable launching

python lm_rerank.py [query-file] [top-100-file] [collection-file] [model=uni|bi]

query-file: file containing the queries in the same tsv format as given in Table 1 for queries file
top-100-file: a file containing the top100 documents in the same format as train and dev top100 files given, which need to be reranked
collection-file: file containing the full document collection (in the same format as msmarcodocs file given)
model=uni|bi: it specifies the unigram or the bigram language model that should be used for relevance language model.

NOTE : model has to be either “uni” or “bi” .

Libraries imported:

```
import operator
import sys
import nltk
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
import csv
import math
```

First entire document collection has been read line by line and memory address of each document id has been stored as offset so that its data can be later loaded.
All the queries are then read from the queries file and tokenized using nltk regex tokenizer, stop words are also removed using porter stemmer and words are converted to lower case and then data for each query is stored.
After that the top 100 file is traversed and data for top documents for each query is stored.
Now the documents are tokenized and data is stored similar to the way it was done for queries and vocabulary set is also built.
Frequencies of the words in the vocabulary is then recorded for each document for calculation used in Dirichlet smoothing.
100 documents for each query are reranked using the KL divergence score where relevance model probabilities P(w/R) are calculated using an estimate for P(w/q1,q2,q3..qk).
For the bigram model terms probabilities are considered pairwise as documents and queries are modelled as Markov processes where terms are considered pairwise and smoothing term for each word is also taken accordingly.
Value of μ is taken as 1.5 for Dirichlet smoothing in uni-gram and bi-gram models.

(In the bi-gram models value of lambda1, lambda2 and lambda3 are taken as μ/(|D|+μ) in the expression for P(w(i)/wi(i-1) ,D)) .

T-test for MRR for uni-model

T-Test Calculator for 2 Dependent Means

The value of *t* is -0.298753.

Explanation of results

The output of this calculator is pretty straightforward. The values of *t* and *p* appear at the bottom of the page. If the text is blue, your result is significant; if it's red, it's not. The only thing that might catch you out is the way that we've rounded the data. The data you see in front of you, apart from the *t* and *p* values, has been rounded to 2 significant figures. However, we did not round when actually calculating the values of *t* and *p*. This means that if you try to calculate these values on the basis of the summary data provided here, you're likely going to end up with a slightly different - and less accurate - result.

Treatment 1	Treatment 2	*Diff* (T2 - T1)	*Dev* (Diff - M)	*Sq. Dev*
0	0	0	0.02	0
0	0	0	0.02	0
0	0	0	0.02	0
1	0	-1	-0.98	0
0	0	0	0.02	0
0	0	0	0.02	0.96
0	1	1	1.02	0
0	1	1	1.02	0
0	0	0	0.02	0.96
0	0	0	0.02	0
0	0	0	0.02	0.96
0	0	0	0.02	1.04
1	0	-1	-0.98	0
0	0	0	0.02	0
1	0	-1	-0.98	0
0	0	0	0.02	0
0	0	0	0.02	0
0	0	0	0.02	0
0	0	0	0.02	0
0	0	0	0.02	1.04
1	1	0	0.02	0

Significance Level:

- ○ 0.01
- ◉ 0.05
- ○ 0.10

One-tailed or two-tailed hypothesis?:

- ○ One-tailed
- ◉ Two-tailed

Difference Scores Calculations

Mean: -0.02
$\mu = 0$
$S^2 = SS/df = 10.98/(50-1) = 0.22$
$S^2_M = S^2/N = 0.22/50 = 0$
$S_M = \sqrt{S^2_M} = \sqrt{0} = 0.07$

T-value Calculation

$t = (M - \mu)/S_M = (-0.02 - 0)/0.07 = -0.3$

The value of *t* is -0.298753. The value of *p* is .76639. The result is *not* significant at p < .05.

Calculate Reset

Ndcg for Wilcoxon for uni-model

Wilcoxon Signed-Rank Test Calculator

Success!

Explanation of results

We have calculated both a *W*-value and *z*-value. If the size of *N* is at least 20 - see the Results Details box - then the distribution of the Wilcoxon *W* statistic tends to form a normal distribution. This means you can use the*z*-value to evaluate your hypothesis. If, on the other hand, the size of *N* is low, and particularly if it's below 10, you should use the *W*-value to evaluate your hypothesis.

You should also note that if a subject's difference score is zero - that is, if a subject has the same score in both treatment conditions - then the test discards the individual from the analysis and reduces the sample size. If you have a lot of ties, this procedure will undermine the reliability of the test (and also suggests that the requirement that the data is continuous has not been met).

Treatment 1	Treatment 2	*Sign*	*Abs*	*R*	*Sign R*
0	0	n/a	0	n/a	n/a
0	0	n/a	0	n/a	n/a
0	0	n/a	0	n/a	n/a
0	1	-1	1	6	-6
0	0	n/a	0	n/a	n/a
0	0	n/a	0	n/a	n/a
1	0	1	1	6	6
1	0	1	1	6	6
0	0	n/a	0	n/a	n/a
0	0	n/a	0	n/a	n/a
0	0	n/a	0	n/a	n/a
0	0	n/a	0	n/a	n/a
0	1	-1	1	6	-6
0	0	n/a	0	n/a	n/a
0	1	-1	1	6	-6
0	0	n/a	0	n/a	n/a
0	0	n/a	0	n/a	n/a
0	0	n/a	0	n/a	n/a

1	1	n/a	0	n/a	n/a

Significance Level:

- ○ .01
- ◉ .05

1 or 2-tailed hypothesis?:

- ○ One-tailed
- ◉ Two-tailed

Result Details

W-value: 30
Mean Difference: 0.55
Sum of pos. ranks: 30
Sum of neg. ranks: 36

Z-value: -0.2667
Mean (*W*): 33
Standard Deviation (*W*): 11.25

Sample Size (*N*): 11

Result 1 - *Z*-value

The value of *z* is-0.2667. The *p*-value is .78716.

The result is *not* significant at $p < .05$.

Result 2 - *W*-value

The value of *W* is 30. The critical value for *W* at $N = 11$ ($p < .05$) is 10.

The result is *not* significant at $p < .05$.

Calculate Reset