Practice exercises for the midterm

CS242 – UCR

Prepared by Vagelis Hristidis with the help of Huayue Gu

1. Consider collection with 1M documents and documents D1, D2

D1: Play basketball today

D2: Play soccer tomorrow

Query Q: “play football”

“Play” appears in 1000 documents, “soccer” in 100, “tomorrow” in 500, “today” in 400, “football” in 600.

In each question, make any assumptions necessary.

a. Compute BM25 scores of the documents.
b. Compute the vector space model cosine similarity between the documents, where the weight of a term is tf*idf, where tf=#occurences/documentlength and idf=#documentsincollection/#documentswithterm
c. Compute the unigram LM scores of D1 and D2 without smoothing
d. Compute the unigram LM scores of D1 and D2 with Dirichlet smoothing

Answer:

(a) BM25

1) $qf = 1$

2) No relevance information $r = R = 0$

3) $N = 1{,}000{,}000$

4) “play” occurs in 1000 doc $n_1 = 1000$

5) “football” occurs in 600 doc $n_2 = 600$

6) “play” occurs in doc $f_1 = 1$

7) “football” occurs in doc $f_2 = 0$

8) Set $\frac{dl}{avgdl} = 90\%, k_1 = 1.2, b = 0.75, k_2 = 100$, $K = 1.11$ set the base e here

$$BM25(Q, D_1) = log \frac{0.5/0.5}{(1000 + 0.5)/(1000000 - 1000 + 0.5)} * \frac{(1.2 + 1) * 1}{1.11 + 1} * \frac{(100 + 1) * 1}{100 + 1}$$

$$+ log \frac{\frac{0.5}{0.5}}{\frac{600 + 0.5}{1000000 - 600 + 0.5}} * \frac{(1.2 + 1) * 0}{1.11 + 0} * \frac{101 * 1}{100 + 1}$$

$$BM25(D, Q_1) = 7.2$$

(b) vector space model cosine similarity between the documents

step 1: idf of each term needs to be calculated (here the base is 10) Set "Basketball" appears in 1000 documents,

Play log(1000000/1000)= 3

football log(1000000/600) = 3.22

Basketball log(1000000/1000) = 3

today log(1000000/400) = 3.40

soccer log(1000000/100) = 4

tomorrow log(1000000/500) = 3.30

step 2: tf for each term in each document

	play	basketball	today	soccer	tomorrow
d1	1	1	1	0	0
d2	1	0	0	1	1

step 3: tf-idf

	play	basketball	today	soccer	tomorrow
d1	3.00	3.00	3.40	0	0
d2	3.00	0	0	4.00	3.30

step 4: query 'play football'

	play	basketball	today	soccer	tomorrow	football
Q	**3.00**	**0**	**0**	**0**	**0**	3.22
d1	3.00	3.00	3.40	0	0	0
d2	3.00	0	0	4.00	3.30	0

$$score(d1, Q) = \frac{Q * d1}{|Q||d1|} = \frac{9}{4.4* \sqrt{3^2 + 3^2 + 3.4^2}}$$

$$score(d2, Q) = \frac{Q * d2}{|Q||d2|} = \frac{9}{4.4* \sqrt{3^2 + 4^2 + 3.3^2}}$$

(c) the unigram LM scores of D1 and D2 without smoothing is 0 for both d1 and d2. The reason is that football does not exist in any documents

(d) Here I choose to use query-likelihood model and Dirichlet smoothing to avoid zero

(base =10)

$$\mu = 2000\,, |d| = 3$$

$$Number\ of\ word\ occurences\ in\ the\ collection = 1000000 * 3 = 3000000$$

Q: play football

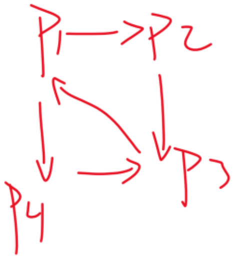
$$for\ the\ term\ \text{"play"}\quad f_{q_i,D} = 1, C_{q,i} = 1000$$

$$for\ the\ term\ \text{"football"}\quad f_{q_i,D} = 0, C_{q,i} = 600$$

$$Q_L(D_1, Q) = log\frac{1 + 2000 * (1000/3000000)}{3 + 2000} + log\frac{0 + 2000 * (\frac{600}{3000000})}{3 + 2000} = -6.77$$

$$Q_L(D_2, Q) = -6.77(same\ calculation\ as\ D1)$$

2. Consider the below 4 pages and hyperlinks.



a. Show first iteration of Pagerank computation

b. Show the updated formulas if only P2 and P3 are in the Base Set.

Answer:

(a)Set the number of web pages N = 4, the damping parameter d = 0.5

$$PR(p1) = \frac{1-d}{N} + d * \frac{PR(Tn)}{C(Tn)} = \frac{0.5}{4} + 0.5 * (\frac{PR(p3)}{1})$$

$$PR(p2) = \frac{0.5}{4} + 0.5 * (\frac{PR(p1)}{2})$$

$$PR(p3) = \frac{0.5}{4} + 0.5 * \left(\frac{PR(p2)}{1} + \frac{PR(p4)}{1}\right)$$

$$PR(p4) = \frac{0.5}{4} + 0.5 * \left(\frac{PR(p1)}{2}\right)$$

Iteration 0: PR(p1)= PR(p2)=PR(p3)=PR(p4)=1

Iteration 1: PR(p1) = 0.625, PR(p2) = 0.375, PR(p3)= 1.125, PR(p4)= 0.375

(b) Only P2 and P3 in the base set

$$PR(p1) = 0.5 * (\frac{PR(p3)}{1})$$

$$PR(p2) = \ \frac{0.5}{2} + 0.5 * (\frac{PR(p1)}{2})$$

$$PR(p3) = \ \frac{0.5}{2} + 0.5 * \left(\frac{PR(p2)}{1} + \frac{PR(p4)}{1}\right)$$

$$PR(p4) = 0.5 * \left(\frac{PR(p1)}{2}\right)$$

Iteration 0: PR(p1)= PR(p2)=PR(p3)=PR(p4)=1

Iteration 1: PR(p1) =0.5, PR(p2) = 0.5, PR(p3)= 0.75, PR(p4)= 0.25

3. Consider ranking for a query Q:

d1, d2, d3, d4, d5

Let the set of relevant documents for Q be {d3, d7}

a. Compute the Average Precision of the ranking
b. Compute the DCG at position 4. Assume Boolean relevance (0/1).
c. Compute Reciprocal Rank

Answer: r r x r r (x is a relevant result and r is not relevant)

(a) Average Precision = sum of precision/ the number of the relevant documents $= \frac{\frac{0}{1}*0 + \frac{0}{2}*0+\frac{1}{3}*1+ \frac{1}{4}*0+\frac{1}{5}*0}{2} = 0.167$

(b) DCG@4 = $0 + \frac{0}{log_2^2} + \frac{1}{log_2^3} + \frac{0}{log_2^4} = 0.631$

(c) MRR = $\frac{1}{3}$ first relevant document is retrieved at position 3

4. MapReduce

a. Write MapReduce pseudocode to find how many input documents contain more than 10 outgoing hyperlinks

Answer:

Algorithm1: MapReduceOfFinding

The mapper emits keypair and document values

```
The reducer show all results according to the keypair
procedure MapOutGoingCount(input)
  While not input.done() do:
        docurl ⟵ input.next()
        count <- 0
     For each hyperlink in outgoint_list do:
           Count + 1
   endwhile
    if (count > 10) do
        EMIT(docurl,1)
    endif
end procedure

procedure ReduceOutGoingCount(key, value):
   EMIT(key,value)
end procedure
```

This is an efficient implementation because only one pair is emitted per document, and the reducer has no work to do.

b. Write MapReduce pseudocode to find how many input documents contain more than 10 incoming hyperlinks

```
Algorithm2: MapReduceOfFinding
The mapper emits keypair and document values
The reducer show all results according to the keypair
procedure MapInComingCount(input)
  While not input.done() do:
        docurl ⟵ input.next()
     For each hyperlink of the form docurl→targetdocurl
in incoming_list do:
           Emit(targetdocurl,1)
   endwhile
end procedure

procedure ReduceInComingCount(key, value):
     count ⟵ 0
     document ⟵ key
    while not value.done() do:
          count ⟵ count +1
    end while

    if (count > 10) do
        EMIT(targetdocurl,count)
    endif
end procedure
```

Mapper:

for each document d, wit hyperlinks to d1, d2,…, we emit pairs (d1,1), (d2,1),….

Reducer

Count number of links and if >10 output