Information Retrieval & Web Search SS 2023

Ex4: Machine Learning and IR, Evaluation, Web Search & Link Analysis

Benedikt Ebing, Fabian David Schmidt

partially based on “An Introduction to Information Retrieval” by Manning, Raghavan and Schütze

Machine Learning and IR

Task 1

Consider the document-term matrix computed in the previous task and assume a binary logistic regression model has been trained to classify whether a document is about Lord of the Rings (=relevant). The learned weights correspond to $w_{Frodo} = 0.74, w_{Sam} = 0.986, w_{beast} = 0.3, w_{orc} = 0.625, w_{blue} = 0.124, w_0 = b = 0$

$$\text{document-by-term-matrix: } = \begin{array}{cc} & \begin{array}{ccccc} \text{Frodo} & \text{Sam} & \text{beast} & \text{orc} & \text{blue} \end{array} \\ & \begin{pmatrix} 0.1249 & 0.2499 & 0 & 0 & 0 \\ 0.2499 & 0.1249 & 0.6021 & 0 & 0 \\ 0.2499 & 0 & 0 & 0 & 0.6021 \\ 0 & 0.1249 & 0 & 0 & 0.3010 \end{pmatrix} & \begin{array}{c} \text{doc 1} \\ \text{doc 2} \\ \text{doc 3} \\ \text{doc 4} \end{array} \end{array}$$

For each of the documents, what is the probability of it being relevant?

$$P(relevant|x) = \frac{1}{1+e^{-z}}, z = \langle w, x \rangle$$

$$P(relevant|x = doc1) = \frac{1}{1+e^{-0.3388}} = 0.5839 \qquad z_{doc1} = 0.74 \cdot 0.1249 + 0.986 \cdot 0.2499 + 0.3 \cdot 0 + 0.625 \cdot 0 + 0.123 \cdot 0 + 0 = 0.3388$$

$$P(relevant|doc2) = 0.6198$$

$$P(relevant|doc3) = 0.5645$$

$$P(relevant|doc4) = 0.54$$

Task 2

Recall the word embedding task from the previous exercise: We want to train CBOW word embeddings over the following vocabulary:

$$[\text{“}Frodo\text{”}, \text{“}followed\text{”}, \text{“}Sam\text{”}, \text{“}into\text{”}, \text{“}the\text{”}, \text{“}dark\text{”}, \text{“}Mordor\text{”}, \text{“}Ring\text{”}]$$

Our current instance consists of the center word Sam and the context words Frodo, followed, into, the. Our model has predicted the following prediction vector:

$$\hat{y} = [0.4 \quad 0.32 \quad 0.12 \quad 0.1 \quad 0.050.09 \quad 0.12 \quad 0.15]$$

What is the cross-entropy of this example?

$$J(\theta) = -\sum_i y^{(i)} \cdot \log(h(x^{(i)})|\theta) + (1 - y^{(i)}) \cdot \log(1 - h(x^{(i)}|\theta))$$ (binary classification)

$$J(\theta) = -\sum_i \sum_j y_j^{(i)} \cdot \log(h(x_j^{(i)})|\theta)$$ (general)

$$y = [0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]$$

$$J([W, W']) = -\log(0.12) = 0.921$$

Evaluation

Task 1

An IR system returns 8 relevant documents, and 10 nonrelevant documents. There are a total of 20 relevant documents in the collection. What is the precision of the system on this search, what is its recall?

Positive (document):	Document in result list	**True Positive (TP):**
		False Positive (FP):
			returned document is relevant
			returned document is non-relevant

Negative (document):	Document not in result list	**True Negative (TN):**
		False Negative (FN):
			correct decision to not return the document
			document should have been in result list

$$Precision = \frac{TP}{TP+FP} = \frac{8}{8+10} = 0.44, \qquad Recall = \frac{TP}{TP+FN} = \frac{8}{8+12} = 0.4$$

Task 2 (a)

Consider an information need for which there are 4 relevant documents in the collection. Contrast two systems run on this collection. Their top 10 results are judged for relevance as follows (the leftmost item is the top ranked search result):

System 1 R N R N N N N N R R
System 2 N R N N R R R N N N

What is the MAP of each system? Which has a higher MAP?

$$\text{MAP}\left(system1\right) = 1/4 \cdot \left(1 + \frac{2}{3} + \frac{3}{9} + \frac{4}{10}\right) = 0.6 \qquad \text{MAP}\left(system2\right) = 1/4 \cdot \left(\frac{1}{2} + \frac{2}{5} + \frac{3}{6} + \frac{4}{7}\right) = 0.49286$$

System 1 has a higher mean average precision.

Task 2 (b)

Does this result intuitively make sense? What does it say about what is important in getting a good MAP score?

Both systems return the same number of relevant documents. MAP values system 1 higher because it ranks the first five documents better than system 2. In general, the documents appearing near the top determine the MAP score the most, later documents impose only minor differences in MAP.

Task 2 (optional)

What is the R-precision of each system? Does it rank the systems the same as MAP?

$$R\text{-}precision\,(system1) = \frac{1}{2}, \qquad R\text{-}precision\,(system2) = \frac{1}{4}$$

R-precision ranks both system the same as MAP.

Task 3 (a) - (c)

The following list of numbers represent the relevance scores of a returned ranked list of 20 documents retrieved in response to a query from a collection of 10,000 documents. The numbers represent a range from nonrelevant (=0) to highly relevant (=3). The top of the ranked list (the document the system thinks is most likely to be relevant) is on the left on the list. This list shows 6 relevant documents. Assume that there are 8 relevant documents in total in the collection.

3 2 0 0 0 0 0 0 1 0 3 0 0 0 2 0 0 0 0 1

(a) What is the precision of the system on the top 20?
(b) What is the F1-Score and what is the Recall on the top 20?
(c) Assume that these 20 documents are the complete result set of the system. What is the MAP for the query?

(a)

3 2 0 0 0 0 0 0 1 0 3 0 0 0 2 0 0 0 0 1
R R N N N N N N R N R N N N R N N N N R

Precision = 0.3

(b) Recall = 0.75, F1-Score = 0.42857

(c) $MAP(q) = \frac{1}{6} = \left(1 + 1 + \frac{3}{9} + \frac{4}{11} + \frac{5}{15} + \frac{6}{20}\right) = 0.555$

Task 3 (optional)

How does the F1-Score compare to MAP in the context of Information Retrieval? Which measure is more suitable? Why?

F-Measure views the result as a set, two systems returning the exact same documents, but one with the relevant documents on top and the others on the bottom will have F-Measure as other systems.

Task 3 (e) - (f)

Assume, now, instead, that the system returned the entire 10,000 documents in a ranked list, and these are the first 20 results returned.

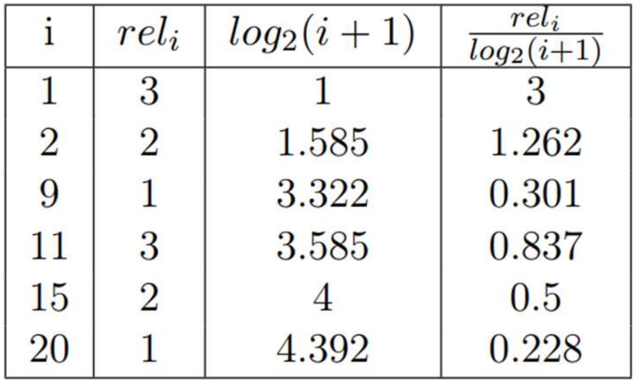
(e) What is the largest possible MAP that this system could have?
(f) What is the smallest possible MAP that this system could have?

(e) $$MAP_{largest}(q) = \frac{1}{8} \cdot \left(1 + 1 + \frac{3}{9} + \frac{4}{11} + \frac{5}{15} + \frac{6}{20} + \frac{6}{20} + \frac{7}{21} + \frac{8}{22}\right) = 0.503$$

(f) $$MAP_{smallest}(q) = \frac{1}{8} \cdot \left(1 + 1 + \frac{3}{9} + \frac{4}{11} + \frac{5}{15} + \frac{6}{20} + \frac{6}{20} + \frac{7}{9999} + \frac{8}{10000}\right) = 0.416$$

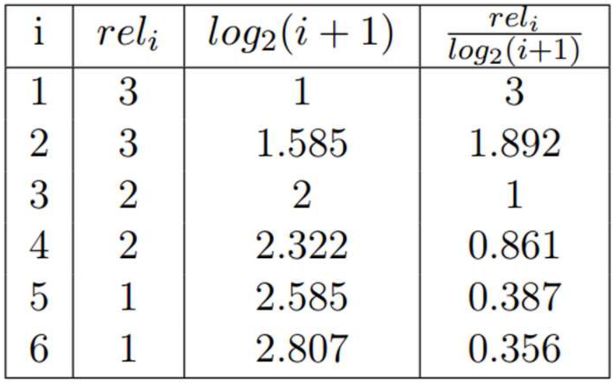
Task 3 (h)

What is the nDCG of the system? Assume this time that the returned documents comprises our corpus.



i	rel_i	$log_2(i+1)$	$\frac{rel_i}{log_2(i+1)}$
1	3	1	3
2	2	1.585	1.262
9	1	3.322	0.301
11	3	3.585	0.837
15	2	4	0.5
20	1	4.392	0.228

DCG = 3 + 1.262 + … + 0.228 = 6.128



i	rel_i	$log_2(i+1)$	$\frac{rel_i}{log_2(i+1)}$
1	3	1	3
2	3	1.585	1.892
3	2	2	1
4	2	2.322	0.861
5	1	2.585	0.387
6	1	2.807	0.356

IDCG = 3 + 1.892 + … + 0.356 = 7.496

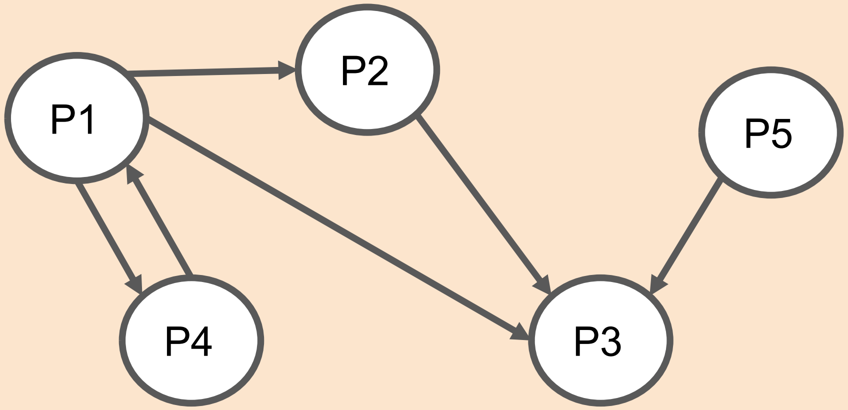
$$DCG(k) = \sum_{i=5}^{k} \frac{rel_i}{\log_2(i+1)}$$

$$nDCG = \frac{DCG}{IDCG} = 0.818$$

PageRank

Setup

Consider the following webgraoh consisting of five websites.



$$r(q, P_1) = 0.43$$
$$r(q, P_2) = 0.31$$
$$r(q, P_3) = 0.05$$
$$r(q, P_4) = 0.12$$
$$r(q, P_5) = 0.84$$

Additionally we have the following content-based relevance value informations (these could be, for example, similarity values from a vector space model) for our query.

Our search engine follows the random surfer model and computes the final relevance values with the PageRank-Algorithm. The probability of jumping to a random page is 1 − d = 0.1 and correspondingly the probability of following a link is d = 0.9.

Task 1

Which pages will be on top after three iterations?

Transition Probability Matrix:

$$\text{link matrix/adjacency matrix} = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

$$\text{row-normalized adjacency matrix} = \mathrm{S} = \begin{bmatrix} 0 & 1/3 & 1/3 & 1/3 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

$$G_{ij} = d \cdot S_{ij} + (1-d)\tfrac{1}{n}$$

$$\text{transition probability matrix} = \mathrm{G} = \begin{bmatrix} 0.02 & 0.32 & 0.32 & 0.32 & 0.02 \\ 0.02 & 0.02 & 0.92 & 0.02 & 0.02 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.92 & 0.02 & 0.02 & 0.02 & 0.02 \\ 0.02 & 0.02 & 0.92 & 0.02 & 0.02 \end{bmatrix}$$

$G_{2,3} = 0.9 * 1 + 0.1 * 1/5 = 0.92$

$G_{1,1} = 0.9 * 0 + 0.1 * 1/5 = 0.02$

$G_{i,j}$ describes the transition probability from page i to page j

Iterative PageRank:

$$\pi_0 = \begin{bmatrix} 0.43 \\ 0.31 \\ 0.05 \\ 0.12 \\ 0.84 \end{bmatrix}$$

$$\pi_1^T = \pi_0^T G = [0.152 \quad 0.173 \quad 1.208 \quad 0.173 \quad 0.044]$$

$$\pi_2^T = \pi_1^T G = [0.408 \quad 0.298 \quad 0.493 \quad 0.298 \quad 0.252]$$

$$\pi_3^T = \pi_2^T G = [0.392 \quad 0.246 \quad \boxed{0.742} \quad 0.246 \quad 0.124]$$

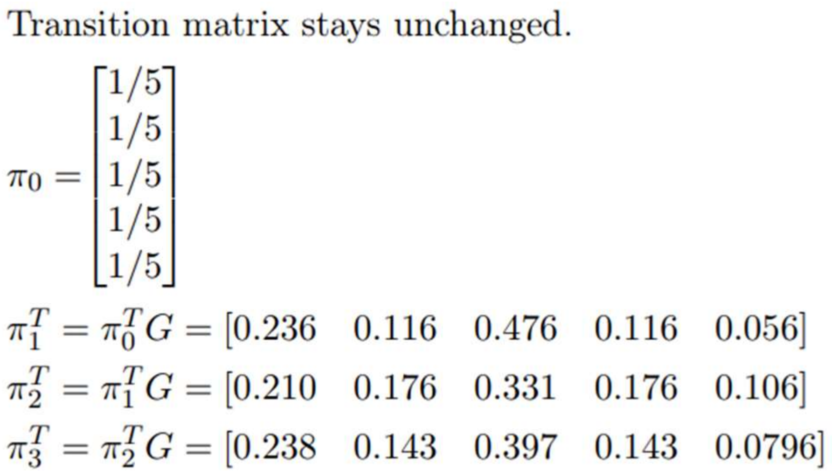
Page 3 will rank on top
with the highest score.

(other options: power-method, simulating random walk, principal eigenvector)

Task 2

Now consider all pages have the same initial relevance value, i.e., r(q, P_i) = 1/5, i ∈ {1, . . . , 5}. Re-compute the PageRank values. Is the result list still the same?

Iterative PageRank:



Transition matrix stays unchanged.

$$\pi_0 = \begin{bmatrix} 1/5 \\ 1/5 \\ 1/5 \\ 1/5 \\ 1/5 \end{bmatrix}$$

$$\pi_1^T = \pi_0^T G = [0.236 \quad 0.116 \quad 0.476 \quad 0.116 \quad 0.056]$$

$$\pi_2^T = \pi_1^T G = [0.210 \quad 0.176 \quad 0.331 \quad 0.176 \quad 0.106]$$

$$\pi_3^T = \pi_2^T G = [0.238 \quad 0.143 \quad 0.397 \quad 0.143 \quad 0.0796]$$

Page 3 ranks on top again.