Web Search Engines — Problem Set 3

Himaja Rachakonda
N14633788
hr970

March 28, 2016

1 Problem 1

The decisons of the judges are as follows –

		Judge 2		
		Yes	No	Total
Judge 1	Yes	2	4	6
	No	4	2	6
	Total	6	6	12

$$P(A) = \frac{4}{12} = 0.333$$
$$P(Relevant) = \frac{6}{12} = 0.5$$
$$P(Non - relevant) = \frac{6}{12} = 0.5$$

$$\begin{aligned} P(E) =& P(NR)^2 + P(R)^2 \\ =& 0.5^2 + 0.5^2 \\ =& 0.5 \end{aligned}$$

1.1 Calculation of Kappa κ

$$\begin{aligned} \kappa =& \frac{P(A) - P((E)}{1 - P(E)} \\ =& \frac{0.33 - 0.5}{1 - 0.5} \\ =& -0.36 \end{aligned}$$

1.2 Relevant if both judges agree

	Relevant	Non Relevant
Retrieved	1(tp)	4(fp)
Not Retrieved	3(fn)	4(fn)

1.2.1 Precision

$$\begin{aligned} Precision =& \frac{tp}{tp + fp} \\ =& \frac{1}{5} \\ =& 0.2 \end{aligned}$$

1.2.2 Recall

$$\begin{aligned} Recall =& \frac{tp}{tp+fn} \\ =& \frac{1}{4} \\ =& 0.25 \end{aligned}$$

1.2.3 F_1–Score

$$\begin{aligned} F_1 =& \frac{2PR}{P+R} \\ =& \frac{2*0.2*0.25}{0.2+0.25} \\ =& 0.222 \end{aligned}$$

1.3 Relevant if either of the judges agree

	Relevant	Non Relevant
Retrieved	5(tp)	0(fp)
Not Retrieved	5(fn)	2(fn)

1.3.1 Precision

$$\begin{aligned} Precision =& \frac{tp}{tp+fp} \\ =& \frac{5}{5} \\ =& 1 \end{aligned}$$

1.3.2 Recall

$$\begin{aligned} Recall =& \frac{tp}{tp+fn} \\ =& \frac{1}{2} \\ =& 0.5 \end{aligned}$$

1.3.3 F_1–Score

$$\begin{aligned} F_1 =& \frac{2PR}{P+R} \\ =& \frac{2*1*0.5}{1+0.5} \\ =& 0.666 \end{aligned}$$

2 Problem 2

2.1 Image search based on Text

Image search based on text should rank images based on the number of objects matching the semantic meaning of the query terms in the input. For Example: Query "A man in front of Taj Mahal" should have all attributes - man, Taj Mahal, and his placement before the monument.

2.2 Image search based on Input Images

An Image search with an input image should retrieve all visually similar images in terms of objects' placement, location, color, contrast and brightness. This does not require the images to be similar in a logical sense i.e, relevancy is only with respect to the features of the image and not what the image actually signifies.

Example: An input image with a library can be a library in NYU which has a similar image from a game which has a library scene with similar features.

2.3 How does text search differ from an image search

Text search looks for the semantic match of the query within the image, whereas image search mostly focuses on the visual similarity between the images.

3 Problem 3

	A	B	c
X	10	40	20
Y	50	10	20

3.1 Purity of cluster

The purity of a cluster is given by

$$purity(\Omega, \mathbb{C}) = \frac{1}{N} \sum_k \max_j |\omega_k \cap C_j|$$

Here –

$$\Omega = X, Y$$
$$\mathbb{C} = A, B, C$$

ω_k is the set of documents and c_j is the set of documents in c_j.

$$\begin{aligned} purity(\Omega, \mathbb{C}) &= \frac{1}{150}(40 + 50) \\ &= \frac{90}{150} \\ &= 0.6 \end{aligned}$$

3.2 Computation for TP, TN, FP, FN, P, R, F_β

3.2.1 TP — True Positives

Pairs $\alpha - \beta$ which belong to same cluster in both the gold standard and algorithmic clustering.

$$\begin{aligned} TP &= \frac{10 * 9}{2} + \frac{40 * 39}{2} + \frac{20 * 19}{2} + \frac{50 * 49}{2} + \frac{10 * 9}{2} + \frac{20 * 19}{2} \\ &= 20 * 39 + 25 * 49 + 20 * 19 + 10 * 9 \\ &= 2475 \end{aligned}$$

3.2.2 FP — False Positives

Pairs $\alpha - \beta$ which belong to same cluster in the algorithmic cluster but different clusters in the gold standard.

$$\begin{aligned} FP &= 10 * 40 + 10 * 20 + 40 * 20 + 50 * 10 + 50 * 20 + 10 * 20 \\ &= 400 + 200 + 800 + 500 + 1000 + 200 \\ &= 3100 \end{aligned}$$

3.2.3 FN — False Negatives

Pairs $\alpha - \beta$ which are in the same cluster in the gold standard, but different algorithmic clusters.

$$\begin{aligned} FN &= 10 * 50 + 40 * 10 + 20 * 20 \\ &= 500 + 400 + 400 \\ &= 1300 \end{aligned}$$

3.2.4 TN — True Negatives

Pairs $\alpha - \beta$ that are in different clusters in both the gold standard and the algorithmic clustering.

$$\begin{aligned} TN =& 10 * 10 + 10 * 20 + 40 * 50 + 40 * 20 + 20 * 50 + 20 * 10 \\ =& 100 + 200 + 2000 + 800 + 1000 + 200 \\ =& 4300 \end{aligned}$$

3.3 Rand index

$$\begin{aligned} RI =& \frac{TP + TN}{TP + FP + FN + TN} \\ =& \frac{6775}{11175} \\ =& 0.606 \end{aligned}$$

3.4 Precision

$$\begin{aligned} Precision =& \frac{TP}{TP + FP} \\ =& \frac{2475}{2475 + 3100} \\ =& \frac{2475}{5575} \\ =& 0.4439 \end{aligned}$$

3.5 Recall

$$\begin{aligned} Recall =& \frac{TP}{TP + FN} \\ =& \frac{2475}{2475 + 1300} \\ =& \frac{2475}{3775} \\ =& 0.6556 \end{aligned}$$

3.6 F score when β=1

$$\begin{aligned} F_\beta =& \frac{(\beta^2+1)*P*R}{\beta^2 P+R} \\ =& \frac{2PR}{P+R} \\ =& \frac{2*0.4439*0.6556}{0.4439+0.6556} \\ =& \frac{0.5820}{1.0995} \\ =& 0.5293 \end{aligned}$$