0007 sheet 1 učo 409729 points

Perform the boolean query VIKINGS AND SPAM given the following inverted index:

- VIKINGS: 2, 3, 5, 7, 11, 13, 17, 23, 29, 31, 37, 41, 47, 53, 59, 61
- SPAM: 5, 13, 17, 23, 29, 31, 43, 47, 59

$|P| = 16 \quad \lfloor\sqrt{|P|}\rfloor = 4$

$|P| = 9 \quad \lfloor\sqrt{|P|}\rfloor = 3$

Present the result of the query [1 point] and the number of comparisons with [2 points], and without [2 points] a *skip list*. Assume the skip list has frequency $\lfloor\sqrt{|P|}\rfloor$, where $|P|$ is the cardinality of a posting P and $\lfloor . \rfloor$ is the *floor function*. *(Example:* $P = \{1, 2, 3\}, |P| = 3, \sqrt{|P|} \approx 1.73, \lfloor\sqrt{|P|}\rfloor = 1$*)*

Results of the query: 5, 13, 17, 23, 29, 31, 47, 59

Number of comparisons:

- with skip lists: 18

  (2, 5), (11, 5), (3, 5), (5, 5), (7, 13), (11, 13),
  (29, 13), (13, 13), (17, 17), (23, 23), (29, 29), (31, 31),
  (37, 43), (41, 43), (47, 43), (47, 47), (53, 59), (59, 59)

- without skip lists: 16

  (2, 5), (3, 5), (5, 5), (7, 13), (11, 13), (13, 13), (17, 17),
  (23, 23), (29, 29), (31, 31), (37, 43), (41, 43), (47, 43),
  (47, 47), (53, 59), (59, 59)

Write your solution only on this side of the sheet!

0007 *sheet* 2 *učo* 409729 *points*

Compute an unbiased estimate of a text retrieval system's *precision*, *recall*, and *the F_1 measure* on the first five results [3 points], and the *precision at 40% recall* [2 points] given the following lists of results for queries q_1, and q_2, where `R` is a relevant result, and `N` is a non-relevant result:

- Results for q_1 : `RNNRRNR` (10 relevant results for q_1 exist in the collection.)
- Results for q_2 : `NRNRRRRN` (5 relevant results for q_2 exist in the collection.)

First five results:

- Results for q_1: RNNRR
- Results for q_2: NRNRR

$P_{q_1}@5 = \frac{3}{5}$, $P_{q_2}@5 = \frac{3}{5}$

$R_{q_1}@5 = \frac{3}{10}$, $R_{q_2}@5 = \frac{3}{5}$

Macro-averaging P, R

$P@5 = (P_{q_1}@5 + P_{q_2}@5)/2 = \frac{3}{5} = 0,6$

$R@5 = (R_{q_1}@5 + R_{q_2}@5)/2 = \frac{9}{20} = 0,45$

Micro-averaging P, R

$P@5 = \frac{3+3}{5+5} = \frac{3}{5} = 0,6$

$R@5 = \frac{3+3}{10+5} = \frac{6}{15} = \frac{2}{5} = 0,4$

Macro-averaging F_1

$F_1@5 = \frac{2 \cdot R@5 \cdot P@5}{R@5 + P@5} = \frac{0,54}{1,05} \doteq 0,5143$

Micro-averaging F_1

$F_{1q_1}@5 = \frac{2 \cdot R_{q_1}@5 \cdot P_{q_1}@5}{R_{q_1}@5 + P_{q_1}@5} = \frac{0,36}{0,9} = 0,4$

$F_{1q_2}@5 = \frac{2 \cdot R_{q_1}@5 \cdot P_{q_1}@5}{R_{q_1}@5 + P_{q_1}@5} = \frac{0,72}{1,2} = 0,6$

$F_1@5 = (F_{1q_1}@5 + F_{1q_2}@5)/2 = 0,5$

Macro-averaging F_1

$F_1@5 = \frac{2 \cdot R@5 \cdot P@5}{R@5 + P@5} = \frac{0,48}{1} = 0,48$

Micro-averaging F_1

$F_{1q_1}@5 = \frac{2 \cdot R_{q_1}@5 \cdot P_{q_1}@5}{R_{q_1}@5 + P_{q_1}@5} = \frac{0,36}{0,9} = 0,4$

$F_{1q_2}@5 = \frac{2 \cdot R_{q_1}@5 \cdot P_{q_1}@5}{R_{q_1}@5 + P_{q_1}@5} = \frac{0,72}{1,2} = 0,6$

$F_1@5 = (F_{1q_1}@5 + F_{1q_2}@5)/2 = 0,5$

Results with 40% recall:

- Results for q_1: RNNRRNR
- Results for q_2: NRNR

$P_{q_1} = \frac{4}{7}$, $P_{q_2} = \frac{1}{2}$

$R_{q_1} = \frac{4}{10}$, $R_{q_2} = \frac{2}{5}$

Macro-averaging P

$P = (P_{q_1} + P_{q_2})/2 = \frac{15}{28} \doteq 0,5357$

Micro-averaging P

$P = \frac{4+1}{7+2} = \frac{5}{9} = 0,\overline{5}$

Write your solution only on this side of the sheet!

0007 *sheet* 3 *učo* 409729 *points*

Define the two assumptions the *Naive Bayes* classifier makes [2 points]. Explain the advantage of computing a product of probability estimates as a sum in the logarithmic space [1 point]. Given an observation $\mathbf{x}$, and the classes c_1, and c_2, is the knowledge of $P(\mathbf{x} \mid c_1)P(c_1) > P(\mathbf{x} \mid c_2)P(c_2)$ sufficient to decide whether $P(c_1 \mid \mathbf{x}) > P(c_2 \mid \mathbf{x})$? Why or why not? [2 points]

Given the following list of observations, use the Naive Bayes classifier to decide whether to play golf when it is sunny, hot, windy, and the humidity is normal. [5 points]

Outlook	Temperature	Humidity	Windy	Play golf
Sunny	Mild	High	False	Yes
Sunny	Mild	Normal	False	Yes
Overcast	Hot	High	False	Yes
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	True	No
Rainy	Hot	High	False	No
Sunny	Cool	Normal	True	No
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Hot	High	True	No
Rainy	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	True	Yes

$P(x_i \mid x_{i+1}, x_{i+2}, \dots, x_n, c_k) = P(x_i \mid c_k)$

Numerical stability of multiplying small real numbers.

By Bayes' theorem, $P(c_1 \mid \vec{x}) = \frac{P(\vec{x} \mid c_1)\, P(c_1)}{P(\vec{x})}$, and $P(c_2 \mid \vec{x}) = \frac{P(\vec{x} \mid c_2)\, P(c_2)}{P(\vec{x})}$. Clearly, $P(\vec{x} \mid c_1)\, P(c_1) > P(\vec{x} \mid c_2)\, P(c_2)$ implies $P(c_1 \mid \vec{x}) > P(c_2 \mid \vec{x})$, and vise versa.

$P(\text{Yes} \mid \text{Sunny}, \text{Hot}, \text{Normal}, \text{True}) = P(\text{Sunny} \mid \text{Yes}) \cdot P(\text{Hot} \mid \text{Yes}) \cdot P(\text{Normal} \mid \text{Yes}) \cdot P(\text{True} \mid \text{Yes}) = \frac{3}{9} \cdot \frac{2}{9} \cdot \frac{6}{9} \cdot \frac{3}{9} \cdot \frac{9}{14} = \frac{972}{91854}$

$P(\text{No} \mid \text{Sunny}, \text{Hot}, \text{Normal}, \text{True}) = P(\text{Sunny} \mid \text{No}) \cdot P(\text{Hot} \mid \text{No}) \cdot P(\text{Normal} \mid \text{No}) \cdot P(\text{True} \mid \text{No}) = \frac{2}{5} \cdot \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{3}{5} \cdot \frac{5}{14} = \frac{60}{17500}$

$P(\text{Yes} \mid \text{Sunny}, \text{Hot}, \text{Normal}, \text{True}) > P(\text{No} \mid \text{Sunny}, \text{Hot}, \text{Normal}, \text{True})$. Therefore, play golf.

Write your solution only on this side of the sheet!

0007 *sheet* 4 *učo* 4097129 *points*

Given a directed graph G that represents three Web pages $V(G) = \{a, b, c\}$, and the links $E(G) = \{(b, a), (c, a), (c, b), (b, c)\}$ between these three pages, draw G [1 point] and produce *the adjacency matrix* (also known as *the link matrix*) $\mathbf{A}$ [1 point], and *the Markov transition matrix* $\mathbf{P}$ [2 points].

Describe the intuition behind *the PageRank algorithm* [1 point]. Compute the *PageRank* of the pages a, b, and c using a single iteration of the PageRank algorithm [2 points].

Describe what we mean, when we call a page a *hub*, or an *authority* [1 point]. Compute the *hub*, and *authority scores* of the pages a, b, and c [2 points].

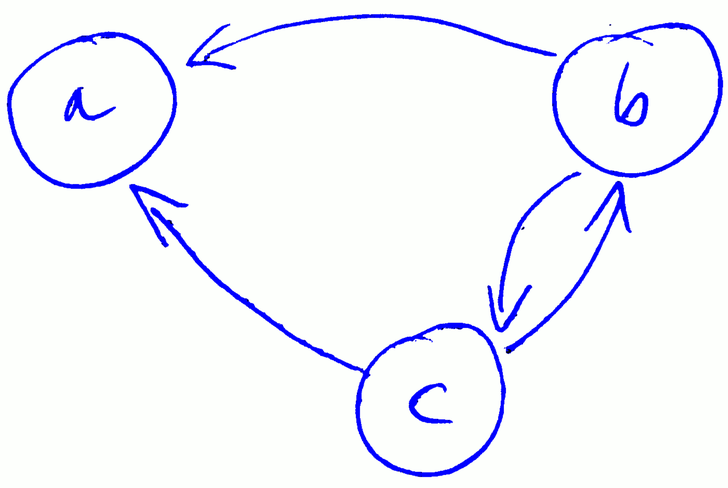

Figure: graph G

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

$$P = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \circ \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/2 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1/2 \end{bmatrix} \cdot (1-\alpha) + \alpha \cdot \frac{1}{3},$$

where $\circ$ is the Hadamard product.

The PageRank algorithm computes the probability that a hypothetical random surfer will end up at a given web page.

$$\vec{x_0} = (1\ 0\ 0) \qquad \vec{x_1} = \vec{x_0} \cdot P = [1\ 0\ 0] \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ \cdot & \cdot & \cdot \end{bmatrix}$$

$$= [1/3 \quad 1/3 \quad 1/3]$$

A hub is a web page pointing to many authorities.

An authority is a web page that many hubs point to.

$$\vec{h_0} = [1\ 1\ 1]^T$$

$$\vec{a_0} = [1\ 1\ 1]^T$$

$$AA^T = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix}$$

$$A^TA = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

$$\vec{h_1} = (AA^T \vec{h_0} = [0\ 3\ 3]^T) / 3 = [0\ 1\ 1]^T$$

$$a_1 = (A^TA\ \vec{a_0} = [4\ 2\ 2]^T) / 4 = [1\ 1/2\ 1/2]^T$$

Write your solution only on this side of the sheet!