המכללה האקדמית להנדסה סמי שמעון

מדור בחינות ומערכת שעות

הנדסת תוכנה

4/02/19
13:30-16:30

מבוא לאחזור מידע

מועד ב'

ליטבק מרינה

תשע"ט סמסטר א'

חומר עזר – אסור (מלבד דפי הנוסחאות המצורפים לטופס זה)

השאלון מכיל 11 עמודים (כולל דף נוסחאות וטיוטה).

==========

חומר עזר : נא סמן במשבצת המתאימה את המתאים

___ * ניתן להשתמש בכל מחשבון
_V_ * **לא** ניתן להשתמש במחשבון Casio FX-991EX
___ * לא ניתן להשתמש במחשבון

_V_ * **לא** ניתן להשתמש בחומר עזר
___ * מותר שימוש בדף נוסחאות, כמפורט: _____
___ * הבחינה בחומר פתוח – מותר להשתמש בכל חומר עזר מודפס או כתוב

הערות

_V_ יש לענות על **כל** השאלות במקומות המיועדים ע"ג טופס השאלון בלבד
___ יש להחזיר את השאלון ביחד עם הכריכה/מחברת.

אחר:
1. _____
2. _____
3. _____

השאלון מכיל 11 עמודים (כולל עמוד זה).

בהצלחה !

==========

שאלה 1 (25 נק')

- יש לענות לכל השאלות
- יש לסמן באופן ברור את התשובה הנכונה ביותר על גבי **שאלון הבחינה**
- סימון של יותר מתשובה אחת לאותה שאלה יקבל ציון של אפס

א. (5נק') ערך אפשרי של Euclidean distance ב-Vector Space Model נכלל בתחום:

1. [0, 1]
2. [0, ∞]
3. [-∞, 1]
4. [-∞,∞]

תשובה (1 /2 /3 /4) נימק:_____

ב. (5נק') תפקיד של-Name Entity Recognition הוא:

1. לחשב שכיחות של המילים במסמך
2. לצמצם כל מילה לשורש-מילה
3. לאתר ולסווג את שמות הישות לקטגוריות מוגדרות מראש כגון שמות אנשים, ארגונים, מיקומים, וכד'
4. להוציא מילות מפתח עבור מסמך

תשובה (1 /2 /3 /4) נימק:_____

ג. (5נק') אלגוריתם למידה בשם K-Nearest Neighbors מקבל כקלט מסמכים בצורה של:

1. ווקטור (Vector Space Model)
2. BOW (Bag of words)
3. גרף
4. טקסט ללא מילות עצירה

תשובה (1 /2 /3 /4) נימק:_____

ד. (5נק') אלגוריתם למידה בשם Rocchio אפשר לאפיין כ:

1. אלגוריתם סיווג לינארי (linear classifier)
2. אלגוריתם לומד ללא פיקוח (unsupervised learning)
3. אלגוריתם סיווג לא לינארי (nonlinear classifier)
4. אלגוריתם סיווג הסתברותי (probabilistic classifier)

תשובה (1 /2 /3 /4) נימק:_____

ה. (5נק') נתונה שאילתה *R*L*. ניתן לאחזר מסמכים שעונים על השאלתה בעזרת permuindex ושאילתה הבאה:

1. *RL*
2. *$RL*
3. *LR$*
4. *L$R**

תשובה (1 /2 /3 /4) נימק:_____

שאלה 2 (25 נק')

נתונים ששה מסמכים (a-h הם המילים):

D1: a h a b a h c b

D2: h b e b h

D3: d h b h d

D4: h d d a h a e

D5: d h d

D6: h c a h c a

מסווגים לשתי קטגוריות: P and N באופן הבא:

מסמך	קטגוריה
D1	P
D2	N
D3	N
D4	P
D5	N
D6	P

(10 נק') לא נתונה רשימת מילות מפתח (keywords), אך ישנה דרישה לחלץ אותם מהסמכים. כיצד ניתן לעשות זאת? תציע שיטה ותראה את היישום שלה על המאגר הנתון. יש להראות את מילות מפתח לכל מסמך.

D1: _____

D2: _____

D3: _____

D4: _____

D5: _____

D6: _____

(15 נק') יש לבצע בחירת מאפיינים (feature selection) בעזרת שיטה Mutual Information

שאלה 3 (30 נק')

נתון מאגר של ששה מסמכים (ראה שאלה 2).

א. (5 נק') יש לבנות אינדקס הפוך (inverted index) עבור כל המאגר, על בסיס מילים בודדות

ב. (5 נק') יש לבנות אינדקס הפוך (inverted index) עבור כל המאגר, על בסיס זוגות מילים (bi-grams)

ג. (5 נק') באיזה אינדקס (מסיף א' או ב') תשתמש עבור שאילתה a AND d AND c ? תאר תאליך אחזור ופלט שלו. האם תשובה תמיד תהיה מדוייקת?

ד. (5 נק') במה מוגבל זמן ריצה של תהליך אחזור עבור שאילתות מסוג a AND b AND c? האם ניתן לעייל את התהליך הזה? אם כן, כיצד? מה אינדקס אמור להכיל עבור זה?

ה. (6 נק') באיזה אינדקס (מסיף א' או ב') תשתמש עבור שאילתה a h c? תאר תהליך אחזור ופלט שלו. האם תשובה תמיד תהיה מדוייקת? נמק.

ו. (4 נק') אם תשובה בסעיף הקודם "אינה תמיד מדוייקת", תציע עדכונים לאינקס לתיקון הבעייה. אם התשובה בסעיף הקודם "תמיד מדוייקת" תראה מה באינקס מבטיח את זה.

שאלה 4 (20 נק')

נתון מאגר של ששה מסמכים (ראה שאלה 2). יש לדרג את המסמכים לפי שיטת TextRank באופן הבא:

א. (10 נק') יש לבנות גרף של מסמכים כך שקדקודים מייצגים מסמכים וקשתות מייצגות דמיון ביניים. יש להשתמש ב-Jaccard similarity

ב. (10 נק') יש לחשב PageRank עבור הדפים (חשבו שלוש איטרציות, מקדם השיכוך = 0.8)

Jaccard(A,B) = |A ∩ B| / |A ∪ B|

Cosine similarity: $$\cos(\vec{q},\vec{d}) = \frac{\vec{q}\bullet\vec{d}}{|\vec{q}||\vec{d}|} = \frac{\vec{q}}{|\vec{q}|}\bullet\frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

Term frequency:

$$\mathrm{tf}(t,d) = \frac{\mathrm{f}(t,d)}{\max\{\mathrm{f}(w,d) : w \in d\}}$$

Inverse document frequency (idf):

$$\mathrm{idf}(t,D) = \log\frac{N}{|\{d \in D : t \in d\}|}$$

Page Rank:

$$PR(A) = (1-d) + d\sum_{i=1}^{n}\frac{PR(T_i)}{C(T_i)}$$

Euclidean distance (in n-space):

$$d(\mathbf{p},\mathbf{q}) = d(\mathbf{q},\mathbf{p}) = \sqrt{(q_1-p_1)^2 + (q_2-p_2)^2 + \cdots + (q_n-p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}.$$

Mutual Information:

$$I(w,c) = \sum_{e_w\in\{0,1\}}\sum_{e_c\in\{0,1\}} p(e_w,e_c)\log\frac{p(e_w,e_c)}{p(e_w)p(e_c)}$$

טיוטה