המכללה האקדמית להנדסה סמי שמעון

מדור בחינות ומערכת שעות

הנדסת תוכנה

14/01/19
09:00-12:00

מבוא לאחזור מידע

מועד א'

ליטבק מרינה

תשע"ט סמסטר א'

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

השאלון מכיל 13 עמודים (כולל נספח 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 9 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 Jaccard similarity נכלל בתחום הבא:
1. [0, 1]
2. [0, ∞]
3. [-∞, 1]
4. [-∞,∞]

תשובה (1 /2 /3 /4) נימק:_____

ב. (5נק') תפקיד של-clustering הוא:
1. לחשב שכיחות של המילים במסמך
2. לצמצם כל מילה לשורש-מילה
3. לזהות קבוצות של מסמכים דומים
4. להוציא מילות מפתח עבור מסמך

תשובה (1 /2 /3 /4) נימק:_____

ג. (5נק') אלגוריתם למידה בשם NaiveBayes 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') מילות עצירה (stop words) אלא מילים שניתן לאפיין על-ידי:
1. tf (term frequency) גבוה ו- idf (inverse document frequency) נמוך
2. tf (term frequency) גבוה ו- idf (inverse document frequency) גבוה
3. tf (term frequency) נמוך ו- idf (inverse document frequency) נמוך
4. tf (term frequency) נמוך ו- idf (inverse document frequency) גבוה

תשובה (1 /2 /3 /4) נימק:_____

ה. (5נק') נתון ביטוי "Natural Language Processing". ניתן לאחזר מסמכים שמכילים אותו בעזרת:
1. אינדקס אפוך (inverted index) בנוי על מילים בודדות ושאילטה "Natural AND Language AND Processing"
2. אינדקס אפוך בנוי על זוגות המילים (bi-grams) ושאילטה "Natural Language AND Language Processing"
3. אינדקס אפוך בנוי על מילים בודדות עם מיקומם בתוך מסמך ושאילטה " Natural AND Language AND Processing"
4. מדיידת דמיון כ-cosine similarity בין ווקטורים של מסמכים ושאילטה

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 לשלוש קטגוריות: P, B and S באופן הבא:

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

א. (10 נק') לא נתונה רשימת מילות עצירה (stop words), אך ישנה דרישה לסנן אותם מהסמכים. כיצד ניתן לעשות זאת? תציע שיטה ותראה את היישום שלה על המאגר הנתון. יש להראות את התוכן המסמכים לאחר סינון.

D1: _____

D2: _____

D3: _____

D4: _____

D5: _____

D6: _____

ב. (15 נק') יש לבצע בחירת מאפיינים (feature selection) בעזרת שיטה χ^2

שאלה 3 (30 נק')

נתונים ששה מסמכים

א. (15 נק') יש לבנות מודל NaiveBayes עבור סיווג מסמכים

ב. (5 נק') נמק למה בחרת בסוג האלגוריתם (סוג של-NB) שהשתמשת בו? אילו עוד סוגים יש?

ג. (5נק') סווגו את שלושת מסמכי המבחן לפי מודל הנלמד בסעיף ב'

i. D7: b h c d e

ii. D8: b h d

iii. D9: a c d a e

ד. (5 נק') חשבו דיוק המבחן (test accuracy) על שלושת מסמכי המבחן אם ידוע סיווג אמיתי (gold standard) שלהם: D7 מסווג ל-N, D8 מסווג ל-N, D9 מסווג ל-P

שאלה 4 (20 נק')

נתון מאגר של ששה מסמכים המתוארים בשאלה מסי 2 ושאילתה Q: a h d h a.
יש לדרג שלושה מסמכים הראשונים (D1, D2, D3) ביחס לשאילתה תוך שימוש ב- tf-idf ו- cosine similarity. יש להתחשב בכל המאגר לחישוב idf.

NaiveBayes Classifier:

$$1 = \arg\max_c \hat{P}(c) \prod_i \hat{P}(x_i \mid c)$$

$$\hat{P}(c_j) = \frac{N(C = c_j)}{N} \qquad \hat{P}(x_i \mid c_j) = \frac{N(X_i = x_i, C = c_j) + 1}{N(C = c_j) + k}$$

Multinomial NB = One feature X_i for each word <u>position</u> in document

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):

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

Chi-square (χ^2):

$$X^2(\mathbb{D}, t, c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}}$$

טיוטה: