המחלקה להנדסת תוכנה
22/01/2021
9:00-12:00

מבוא לאחזור מידע
מועד א'
ד"ר מרינה ליטבק

תשפ"א סמסטר א'

חומר עזר : נא סמן במשבצת המתאימה את המתאים
_V_ * ניתן להשתמש בכל מחשבון
___ * 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
___ יש לענות על **כל** השאלות במקומות המיועדים ע"ג טופס השאלון בלבד
___ יש להחזיר את השאלון ביחד עם הכריכה/מחברת.

אחר:
בדיקת המבחן לא תביא בחשבון את דפי הטיוטה או תוספות בגב העמוד.

השאלון מכיל __5_ עמודים (כולל עמוד זה ונספח).

הצהרת סטודנט:
עם חתימתי מטה, הנני מצהיר בזאת כי פתרתי את הבחינה בעצמי ,ללא סיוע אסור.
ידוע לי כי למרצה שמורה האפשרות לבדוק את ידיעותיי ,במידה ותתגלה חריגה מההצהרה,
החריגה תחשב כעבירת משמעת חמורה על כל המשתמע מכך_____

בהצלחה !
==========

שאלה 1 (35 נק') – נכון או לא נכון ונימוק

- יש לענות על כל השאלות (כל אחת במשקל של 5 נקודות)
- יש לסמן באופן ברור את התשובות הנכונות ביותר ולנמק בקצרה
- יתכן שיש כמה תשובות נכונות בחלק מהשאלות

א. (5 נק') ערך אפשרי של מרחק מנהטן (Manhattan distance) בין שני מסמכים מיוצגים כווקטורים של

משקלות (term frequency) tf נכלל בתחום הבא:

a. [0, 1]
b. [0, ∞]
c. [-∞,∞]
d. [-1, 1]

נימוק: _____

_____

ב. (5 נק') clustering בשילוב עם מנוע חיפוש יכול:

a. לשפר precision
b. לשפר recall
c. לקצר רשימת המסמכים בתוצאת חיפוש
d. להחזר מסמכים עם מילים נרדפים (synonyms) למילים של שאילתה

נימוק: _____

_____

ג. (5 נק') שטח מתחת לעקומת ROC מודד:

a. איכות הדירוג המסמכים של מנוע חיפוש\אחזור מידע:
b. ציון הפרופורציונאלי ל- recall
c. ציון הפרופורציונאלי ל - precision
d. ציון הפרופורציונאלי ל - accuracy

נימוק: _____

_____

ד. (5 נק') מילה שלא מופיעה בהרבה מסמכים (כלומר עם document frequency נמוך) היא:

a. לא חשובה וכדאי לסנן אותה לפני בניית הייצוג הווקטורי עבור כל מסמך
b. חשובה כי היא לא מופיעה בהרבה מסמכים ולכן כדאי להכיל אותה בייצוג הווקטורי של כל מסמך
c. חשובה למסמך שבו היא מופיעה הרבה מסמכים ולכן כדאי להכיל אותה בייצוג הווקטורי של אותו המסמך
d. אי-אפשר לדעת, כי חשיבות המילה לא קשורה לתדירות שלה ולכן זה לא משפיע על הייצוג הווקטורי.

נימוק: _____

_____

ה. (5 נק') זמן ריצה שדרוש לעיבוד השאלתה עם שלילה (NOT) הוא:

a. לינארי ביחס לאורך הרשימה (postings list) הגדולה ביותר בין כל הרשימות של מילות השאילתה
b. לינארי ביחס לסכום האורכים של כל הרשימות של מילות השאילתה
c. לינארי ביחס לגודל המילון (אוצר מילים בכל המאגר)
d. לינארי ביחס לגודל המאגר (מס' המסמכים במאגר)

נימוק: _____

_____

ו. (5 נק') clustering המסמכים זאת שיטה ש:

a. דורשת נתונים מתוייגים (training data) לאימון
b. מסווגת כל מסמך לקטגוריה אחת מתוך כמה שמוגדרות מראש
c. מארגנת מסמכים לקבוצות של מסמכים דומים
d. צורת האשכול (cluster) היא תמיד ספירה בכל האלגוריתמים של clustering.

נימוק: _____

_____

ז. (5 נק') purity זאת מטריקת איכות ה-clustering ש:

a. צריכה נתונים מתוייגים (gold standard) להשוואה
b. לא צריכה נתונים מתוייגים כי clustering הוא unsupervised
c. צריכה נתונים מתוייגים כי clustering הוא supervised
d. מודדת מרחק ממוצע בין תצפיות באשכולות

נימוק: _____

_____

שאלה 2 (15 נקודות)

נתונים 3 דפים ברשת: A, B, and C . A מחזיק קישורים ל-B. C מקושר ל A ו-B. B מקושר לעצמו. ענה על השאלות הבאות למטה:

א. (2 נק') איזו תופעה יש בגרף?
ב. (3 נק') כיצד זה משפיע לחישוב של PageRank?
ג. (5 נק') כיצד ניתן "לתקן" את הבעיה? תציע שינוי של מבנה של גרף (ע"י הוספת ו/או הסרת קשתות)
ד. (5 נק') הראה חישובים הרלוונטיים (מספיק שלוש איטרציות , מקדם השיכוך 0.8)

שאלה 3 (30 נק')

נתונים שישה מסמכים (a-h הן המילים) המסווגים ל-2 קטגוריות:

מס' מסמך	תוכן המסמך	סיווג (קטגוריה)
D1	e c e h	S
D2	h b e b h	S
D3	b b h e d	S
D4	h d d a h	P
D5	d h d	P
D6	a h b a h c b	P

א. (2 נק') כיצד תבנה מודל KNN עבור סיווג מסמכים? הסבר.
ב. (7 נק') חשב את דיוק המבחן (test accuracy) על שלושת מסמכי המבחן עבור K=3 (יש להשתמש ב-cosine בין ווקטורים של tf)
   a. D7 : b h c d d (מסווג ל-S)
   b. D8 : b h d (מסווג ל-P)
   c. D9 : c d e (סווג ל-S)
ג. (5 נק') בנה confusion matrix עבור קבוצת מבחן.
ד. (6 נק') חשב precision ו-recall עבור כל קטגוריה.
ה. (5 נק') האם KNN הוא מסווג לינארי? הסבר.
ו. (5 נק') מה היתרון של KNN מפני Rocchio?

שאלה 4 (20 נק')

בנה Agglomerative Clustering עם complete link מעל מרחק מנהטן (Manhattan distance) ל-4 מסמכים המיוצגים ע"י ווקטורים הבאים (כולל דנדרוגרם):

D1: (1,0,2), D2: (0,2,3), D3: (1,2,3), D4: (2,3,5)

נספח (דף הנוסחאות):

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:

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

Manhattan distance:

$$|A - B| = \sum_{i=1}^{d} |a_i - b_i|$$

Page Rank:

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