המכללה האקדמית להנדסה סמי שמעון

המחלקה להנדסת תעו"נ

02/03/09
08:30-11:30

איחזור וסינון מידע

מועד א'

גב' מרינה ליטבק

תשס"ט סמסטר א'

חומר עזר – חומר פתוח, מחשבון

הוראות מיוחדות – יש לענות על שאלות אמריקאיות (שאלה מס' 1) על גבי שאלון הבחינה ולהגיש אותו לבדיקה.

השאלון מכיל 3 דפים (כולל דף זה).

שאלה מס' 1 (25 נקודות)

- 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. ערך אפשרי של קוסינוס (cosine similarity) בין שני מסמכים מיוצגים כוקטורים של משקלות (term frequency) tf נכלל בתחום הבא:

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

ב. לשאילתה בוליאנית "(A xor B) not C" מתאים מסמך הבא:

1. A B C
2. A B
3. A A
4. C D

ג. מילות מפתח (key words) אלה 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

ד. ישנו מאגר מסמכים מסווגים לשלוש קטגוריות לפי התפלגות אחידה. ערך האנטרופיה (Entropy-או Expected Information) עבור סיווג הנתונים האלה תהיה:

1. ∞
2. 1
3. 0
4. $\log_2 3$

ה.

ו. נתון ביטוי רגולארי "(ma)+ma*". אילו מחרוזות יחזיר מנוע חיפוש עבור הטקסט הבא: "mam mamaaaam":

1. mam, mamaaaa
2. ma, mam, mama, mamaa, mamaaa, mamaaaaa
3. mam, mamaaaam
4. mam, mama

שאלה מס' 2 (25 נקודות)

נתונים ששה מסמכים (a-h הם המילים):

e c e h:D1
h b e b h :D2
b b h d :D3
h d d a h e :D4
d h d :D5
a h b a h c b:D6

מסווגים לשלוש קטגוריות: P, B and S באופן הבא:

מסמך	קטגוריה
D1	S
D2	B
D3	B
D4	P
D5	S
D6	P

- יש לבנות מודל C4.5 עבור סיווג מסמכים **(15 נקודות)**
- חשבו את דיוק האימון (training accuracy) עבור המודל **(5 נקודות)**
- חשבו את דיוק המבחן (test accuracy) על שלושת מסמכי המבחן **(5 נקודות)**:
  i. D7: b h c d d (מסווג ל-P)
  ii. D8: b h d (מסווג ל-S)
  iii. D9: c d e (מסווג ל-S)

שאלה מס' 3 (25 נקודות)

נתון מאגר של ששה מסמכים המתוארים בשאלה מס' 2 ושאילתה Q: b h b h e.

- יש לדרג את שלושת המסמכים הראשונים (D1, D2, D3) ביחס לשאילתה תוך שימוש ב-tf-idf ו-cosine similarity. יש להתחשב בכל המאגר לחישוב idf.
- יש לדרג אותם שלושת המסמכים לפי גיקרד (Jaccard similarity) ביחס לשאילתה Q

שאלה מס' 4 (25 נקודות)

נתונים 3 דפים ברשת: A, B, and C.
A מחזיק קישורים ל-B ו-C.
B מחזיק קישורים ל-A ו-C.

- יש לחשב PageRank עבור הדפים (חשבו שלוש איטרציות, מקדם השיכוך = 0.85) **(15 נקודות)**.
- יש לשנות מבנה של גרף (ע"י הוספת ו/או הסרת קשתות) כך שכל הקדקדים יקבלו ערך של-PageRank גדול יותר **(10 נקודות)**.

בהצלחה !

==========