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 אנטרופיה (Entropy or Expected Information) נכלל בתחום הבא:
1. [0, 1]
2. [0, ∞]
3. [-∞, 1]
4. [-∞,∞]

ב. תפקיד של-stemmer הוא:
1. לחשב שכיחות של המילים במסמך
2. לצמצם כל מילה לשורש-מילה
3. לזהות ולהרכיב ביטויים
4. להוציא מילות מפתח עבור מסמך

ג. האם עץ החלטה עבור סיווג טקסט הוא עץ בינארי?
1. כן, תמיד
2. כן, אם משתמשים בייצוג בינארי של מסמכים
3. כן, אם יש מעט מילים במסמך
4. לא

ד. מילות עצירה (stop 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

ה. נתון ביטוי רגולארי "\d{6}". אילו מחרוזות יחזיר מנוע חיפוש עבור הטקסט הבא:
"0123456789 123456 234abc":
1. 012345, 123456, 234abc
2. 012345, 123456
3. 012345, 123456, 234567, 345678, 456789, 123456
4. 123456, 234abc

שאלה מס' 2 (25 נקודות)

נתונים ששה מסמכים (a-h הם המילים):

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

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

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

- יש לבנות מודל ID3 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 e (מסווג ל-P)
  ii. D8: b h d (מסווג ל-B)
  iii. D9: c d e (מסווג ל-S)

שאלה מס' 3 (25 נקודות)

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

שאלה מס' 4 (25 נקודות)

נתונים 4 דפים ברשת: A, B, C and D.
A מחזיק קישורים ל-B ו-C.
B מחזיק קישורים ל-A ו-D.
C מחזיק קישור ל-D
D מחזיק קישורים ל-A ו-C.

- יש לחשב PageRank עבור הדפים (חשבו שלוש איטרציות, מקדם השיכוך = 0.8) **(15 נקודות)**.
- יש לשנות מבנה של גרף (ע"י הוספת ו/או הסרת קשתות) כך ש-C יקבל ערך של-PageRank גדול יותר **(10 נקודות)**.

בהצלחה !

==========