המחלקה להנדסת תוכנה

08/03/15
09:00-12:00

טכניקות מתקדמות באחזור מידע

מועד ב'

ד"ר מרינה ליטבק

תשע"ה סמסטר א'

חומר עזר – מחשבון

הוראות מיוחדות:

1. **עליך להסביר בקצרה ולנמק את התשובה. הסבר ארוך ומייגע עשוי לגרוע כאשר הוא לא נחוץ.**
2. **יש לענות על השאלה האמריקאית (שאלה מס' 1) על גבי שאלון הבחינה.**
3. **יש להגיש את השאלון לבדיקה יחד עם מחברת בחינה.**

ההוראות במבחן ניתנות בלשון זכר – אך מכוונות לנשים ולגברים כאחת!

השאלון מכיל 4 שאלות ו-4 דפים (כולל דף זה ונספח). **בהצלחה !**

==========

המכללה האקדמית להנדסה סמי שמעון
קמפוס באר שבע ביאליק פינת בזל 84100 | קמפוס אשדוד ז'בוטינסקי 84, 77245 | www.sce.ac.il | חייג: *מספרים

שאלה 1 (25 נק')

- יש לענות לכל השאלות (כל אחת במשקל של-5 נקודות).
- יש לסמן באופן ברור את התשובה הנכונה ביותר על גבי **שאלון הבחינה**
- סימון של יותר מתשובה אחת לאותה שאלה יפסול את השאלה.

א. כל אלגוריתם של clustering אפשר לאפיין באופן הבא:
1. הוא חייב נתונים מתויגים על מנת ללמוד עליהם פונקציה של חלוקת נתונים לאשכולות
2. הוא לא צריך נתונים מתויגים כי הוא לא לומד, אלא מחלק נתונים לאשכולות על בסיס פונקציית " דמיון" בין 2 אובייקטים
3. הוא חייב נתונים מתויגים על מנת ללמוד עליהם פונקציה שמחשבת מספר אופטימלי של אשכולות
4. הפלט שלו תמיד מורכב מקבוצות זרות של אובייקטים

ב. ניתן אינדקס (inverted index) פשוט עבור מאגר בר N מילים ו-M מסמכים. שאילתה בוליאנית "A and B" מתבצעת בזמן הבא:
1. O(M)
2. O(N)
3. O(N+M)
4. O(N*M)

ג. אלגוריתם סיווג (classification) לינארי הוא:
1. לומד על הנתונים המתויגים פונקציה לינארית "המפרידה" בין הקטגוריות נתונים (classes)
2. לומד על הנתונים המתויגים פונקציה "המפרידה" בין הקטגוריות נתונים (classes) בזמן לינארי
3. לומד על הנתונים המתויגים פונקציה "המפרידה" בין הקטגוריות נתונים וזמן למידה שלו הוא לינארי
4. אלגוריתם לא לומד על נתונים מתויגים אך בונה פונקצית הפרדה בין הקטגוריות נתונים בזמן לינארי

ד. סמנו זוג של מסמכים המיוצגים במרחב וקטורי (Vector Space Model) ע"י וקטורים מאונכים (ז"א עם זווית 90° ביניהם).
1. "B C B A A" ו-"C A B"
2. "C D" ו-"B A"
3. "A C C B B A" ו-"C B A"
4. "D C B" ו-"E C B"

ה. סמנו את כל המילים (rotated terms) ב-permuterm index שיצביעו על מילה cat:
1. cat$, ca$t, c$at, $cat
2. ct, at, ca, cat, cat, cat, $cat
3. cat, cats, catty
4. cat$, t$ca, at$c, $cat

שאלה 2 (30 נק'):

נתון מאגר מסמכים:

D1: A B C C E F
D2: A C C C
D3: C C D D F
D4: B C D D

יש לענות על כל השאלות:

א. (5 נק') יש לדרג מילים בכל מסמך לפי tf-idf (יש לחשב idf ביחס ל-4 מסמכים סה"כ)
ב. (3 נק') איך לאתר stopwords לפי ציונים שהתקבלו בסעיף א'?
ג. (2 נק') איך לאתר keywords לפי דירוג שהתקבל בסעיף א'?
ד. (5 נק') בנה bi-gram inverted index עבור המאגר הנתון.
ה. (5 נק') הראו ונמקו כיצד תבצעו שאילתה עם ביטוי (רצף מילים) "A B C" בהינתן ה-inverted index שבניתם. האם התוצאה של השיטה שתיארת תמיד תהיה מדויקת? נמק.
ו. (10 נק') נניח שאלה מסמכים שמנוע חיפוש החזיר בתגובה לשאילתה "A or C". משתמש זהה מסמכים D1 ו-D2 כרלוונטיים, ו-D3 ו-D4 כלא רלוונטיים. סה"כ ישנם 100 מסמכים במאגר ו-10 מתוכם רלוונטיים עבור המשתמש (כולל 2 הרלוונטיים שמערכת החזירה). חשב: TP, FP, TN, FN, Precision, Recall ו-Accuracy

שאלה 3 (20 נק'):

נתונים אובייקטים המיוצגים ע"י ווקטורים הבאים:

A(2,7), B(5,7), C(11,7), D(12,7), E(2,2), F(5,2), G(11,2), H(12,2)

הראה דנדרוגרם (dendrogram) עבור HAC (Hierarchical Agglomerative Clustering) של הנתונים האלה. יש להשתמש ב-Euclidean distance עבור מדידת דמיון ושיטה Single Link עבור חישוב מרחק בין אשכולות (clusters).

שאלה 4 (25 נק')

נתונים ששה מסמכים (a-h הם המילים):

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

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

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

א. (10 נק') יש לבנות מודל Rocchio עבור סיווג מסמכים (השתמש בווקטורים של word counts ו-cosine similarity)
ב. (15 נק') נתונים שלושת מסמכי המבחן:
   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)

(5 נק') חשב את דיוק המבחן (test accuracy)
(5 נק') חשב confusion matrix ו-classification accuracy
(5 נק') אילו מילים מתוך a, b ו-c משפיעות יותר על הסיווג לפי Mutual Information?

נספח:

Mutual Information:

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

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:

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

המכללה האקדמית להנדסה סמי שמעון
קמפוס באר שבע ביאליק פינת בזל 84100 | קמפוס אשדוד ז'בוטינסקי 84, 77245 | www.sce.ac.il | חייג: *מהפנים