המחלקה להנדסת תוכנה

09/02/15
09:00-12:00

טכניקות מתקדמות באחזור מידע

מועד א'

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 שאלות אמריקאיות (שאלה מס' 1) על גבי שאלון הבחינה ולהגיש אותו לבדיקה.**

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

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: *מטמנים

שאלה 1 (25 נק')

- יש לענות על כל השאלות (כל אחת במשקל של-5 נקודות).
- יש לסמן באופן ברור את התשובה הנכונה ביותר על גבי **שאלון הבחינה** ולנמק בקצרה במחברת הבחינה
- סימון של יותר מתשובה אחת לאותה שאלה יקבל ציון של אפס

א. ערך אפשרי של קוסינוס (cosine similarity) בין שני מסמכים מיוצגים כווקטורים של משקלות tf (term frequency) נכלל בתחום הבא:
1. [0, 1]
2. [0, ∞]
3. [-1, 1]
4. [-∞,∞]

ב. אינדקס (inverted index) פשוט (כל מילה מצביעה על רשימת המסמכים בהם היא מופיעה) מאוד שימושי על מנת:
1. לדרג מסמכים ביחס לשאילתה
2. לאתר מסמכים שעונים על שאילתה בוליאנית
3. לבנות מודל KNN עבור סיווג מסמכים (classification)
4. לבצע חלוקה של מסמכים לאשכולות (clustering)

ג. מנוע חיפוש Shmoogle מדרג מסמכים לפי ג'קרד (Jaccard Similarity) ביחס לשאילתה ומחזיר רק מסמך אחד בעל דירוג הגבוה ביותר. איזה מסמך Shmoogle יחזיר בהינתן שאילתה "A B D"?
1. A C B C
2. B A B
3. C A A A
4. C D

ד. סמנו זוג של מסמכים המיוצגים במרחב וקטורי (Vector Space Model) ע"י וקטורים מקבילים (ז"א עם זווית 0° ביניהם).
1. "B C B A A" ו-"C A B"
2. "C B A" ו-"B A"
3. "A C C B B A" ו-"C B A"
4. "D C B" ו-"E C B"

ה. אלגוריתם K-Means מייצר אשכולות (clusters) כאשר כל אחד הוא בצורה של:
1. פֵּאוֹן (polytope)
2. ספירה (sphere)
3. צורה חופשית שתלויה מהתפלגות הנתונים
4. צורה חופשית שנבנית אקראית

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

שאלה 2 (30 נק'):

נתון מאגר מסמכים:
D1: A B A C B C
D2: A C A C D
D3: B C D A
D4: B C D C D
יש לענות לכל השאלות:

א. (5 נק') יש לתאר מסמכים רלוונטיים ביחס לשאילתה בוליאנית Q: (A or not B) and (C or D).
ב. (5 נק') יש לדרג מסמכים לפי קוסינוס (cosine similarity) בין הווקטורים של tf (term frequency) ביחס לשאילתה Q: B C B D. ניתן לראות נוסחה לחישוב tf בתרגיל זה בנספח.
ג. (5 נק') בנה inverted index עבור המאגר הנתון
ד. (5 נק') הראה ונמק כיצד תבצעו שאילתה בוליאנית "B and C and A" בהינתן ה-inverted index שבניתם
ה. (10 נק') האם ניתן לאתר מסמכים שמכילים ביטוי (phrase) "B C" בעזרת אינדקס שבנית בסעיף ג' ? אם כן, תראה כיצד (כל אפשריות). אם לא, מה חסר ואיך אתה מציע לפתור את הבעיה (כל אפשריות)?

שאלה 3 (10 נק'):

נתונים 3 דפים ברשת: A, B, and C. A מחזיק קישורים ל-B ו-C. B מחזיק קישורים ל-A ו-C.
יש לשנות מבנה של גרף (ע"י הוספת ו/או הסרת קשתות) כך שכל הקדקודים יקבלו ערך של-PageRank גדול יותר.
הראה חישובים הרלוונטיים (מספיק שלוש איטרציות, מקדם השיכוך = 0.85)

שאלה 4 (35 נק')

נתונים ששה מסמכים (a-h הן 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
המסווגים לשלוש קטגוריות: P, B and S באופן הבא:

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

א. (15 נק') יש לבנות מודל NaiveBayes Multinomial עבור סיווג מסמכים
ב. (10 נק') חשבו את דיוק המבחן (test accuracy) על 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)
ג. (5 נק') האם קיבלתם סיווג לינארי או לא לינארי? נמק.
ד. (5 נק') תן הגדרה למסווג לינארי ותאר את התכונות שלו.

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:

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

Page Rank:

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

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