

בית הספר למוסמכים במינהל עסקים ע"ש ליאון רקנאטי

1242.3277 – כריית טקסט Text Mining

דרישות קדם:

1231.2416 מדע הנתונים למנהל עסקים

או

1221.7020 מבוא לאנליטיקה עסקית של תואר ראשון

או

1242.3253 נ"מ במדעי הנתונים

סמסטר א' - מחצית שניה - תשפ"ג

| קבוצה | יום בשבוע | שעה | תאריך בחינה | מרצה | דואר אלקטרוני | טלפון |
|-------|-----------|-------------|--------------------------------------|----------|----------------------------|-------|
| 01 | ד | 15:45-18:30 | לפירוט לוחות הבחינות | ענבל יהב | inbalyahav@tauex.tau.ac.il | |

שעת קבלה – בתיאום מראש

לתשומת לב הסטודנטים, בקורס זה קיימת חובת הגעה עם מחשב אישי נייד לכיתה!

כמו כן יש חובת נוכחות בשיעורים.

היקף הלימודים

היקף הי"ס לקורס : 1

1 י"ס = 4 ECTS – ECTS (European Credit Transfer and Accumulation System), ערך הניקוד של הקורס במוסדות להשכלה גבוהה בעולם שהינם חלק מ"תהליך בולוניה".

תיאור הקורס

אנחנו חיים בעידן המדיה החברתית, עידן בו אנו מוצפים במידע המגיע אלינו ממקורות רבים ומגוונים, ובפורמטים שונים. חלק גדול מהמידע המגיע אלינו ונמצא במדיה החברתית הינו טקסטואלי, ורובו נוצר ע"י משתמשי המדיה.

בקורס זה נלמד לנתח מידע טקסטואלי באופן ממוכן. נתמקד בשני סוגי ניתוח נתונים. הראשון הינו ניתוח שאינו מפוקח, היינו, מציאת תבניות ותובנות מתוך אוסף של טקסטים (לדוגמא, מציאת טרנדים וסיכום נושאי שיח ברשת). השני, הינו ניתוח מפוקח, בו המטרה היא לסווג את המידע לקבוצות מוגדרות מראש. דוגמא נפוצה בה נתמקד הינה ניתוח סנטימנטים – האם טקסט נכתב באופן חיובי, שלילי, או ניטרלי.

תפוקות למידה

עם סיום הקורס בהצלחה יוכל הסטודנט:

1. לעבד מסד נתונים טקסטואלי ולהציגו באופן גרפי
2. לבצע ניתוח מגמות ונושאים בטקסט
3. להריץ מודלי סיווג על מסדים טקסטואלים
4. לבצע ניתוח סנטימנטים
5. להריץ מודלי NLP בסיסיים

הערכת הסטודנט בקורס והרכב הציון

| אחוז | מטלה | תאריך | הערות |
|------|--------------------------|--------------------------------------|-----------------------------|
| 20% | מטלות דו-שבועיות (זוגות) | | |
| 20% | פרוייקט מסכם (זוגות) | | |
| 60% | מבחן בית (12 שעות) | לפירוט לוחות הבחינות | חובת מעבר בציון של 60 לפחות |

* תלמיד, הנעדר משיעור המחייב השתתפות פעילה או שלא השתתף באורח פעיל, רשאי המורה להודיע למזכירות כי יש למחוק את שמו מרשימת המשתתפים. (התלמיד יחויב בתשלום בגין קורס זה)

פירוט המטלות בקורס

בקורס ינתנו 2 מטלות שיוגשו בזוגות. בנוסף ינתן פרוייקט מסכם הניתן להגשה בזוגות. ציון הפרוייקט יקבע באופן יחסי לעבודות האחרות.

כל אי עמידה במי ממטלות הקורס מחיבת הודעה מראש (במייל) למרצה

מדיניות שמירה על טווח ציונים

החל משנה"ל תשס"ט מונהגת בפקולטה מדיניות שמירה על טווח ציונים בקורסי התואר השני. עקרונות השיטה חלים על כל קורסי התואר השני, ומדיניות השמירה על טווח הציונים תיושם לגבי הציון הסופי בקורס זה.

מידע נוסף בנושא זה מתפרסם בהרחבה באתר הפקולטה.

[לתקנוני מדיניות שמירת טווח ציונים](#)

הערכת הקורס ע"י הסטודנטים

בסיומו של הקורס הסטודנטים ישתתפו בסקר הוראה על מנת להסיק מסקנות לטובת צרכי הסטודנטים והאוניברסיטה.

אתר הקורס

אתר הקורס יהווה המקום המרכזי בו ימסרו הודעות לסטודנטים, לפיכך מומלץ להתעדכן בו מדי שבוע, לפני השיעור, ובכלל – גם בתום הסמסטר. (לצורך תיאום ענייני הבחינה למשל).

שקפי הקורס יהיו באתר הקורס באתר.

לתשומת לבכם - בכיתה ידונו גם נושאים (ובפרט דוגמאות) שאינם מופיעים בשקפים או מופיעים בכותרת בלבד. כל אלו הינם חלק בלתי נפרד מחומר הקורס.

* תכנית הקורס

| שבוע | תאריך | נושאים | קריאת חובה | הערות |
|------|-------|---|-------------|----------------------|
| 1 | | מבוא לניתוח טקסט, ייצוג טקסט ניקוי ועיבוד טקסט | פרקים 1-2 | יפורסם פרוייקט הקורס |
| 2 | | ויזואליזציה של טקסט ניתוח טקסט באופן לא מפוקח | פרקים 3 ו-5 | יפורסם תרגיל בית 1 |
| 3 | | ניתוח סמטימנטים מקרה בוחן | פרק 4 | |
| 4 | | ניתוח טקסט באופן מפוקח | פרק 6 | יפורסם תרגיל בית 2 |
| 5 | | מודלי סיווג ופרדיקציה | פרק 7 | |
| 6 | | NLP מקרה בוחן | פרק 8 | יפורסם תרגיל בית 3 |
| 7 | | סיכום הצגת פרוייקטים | | |

* התכנית הינה בסיס לשינויים.

קריאת חובה

Text Mining in Practice with R 1st Edition, Ted Kwartler

קריאת רשות

- Liu, Bing, and Lei Zhang. "A survey of opinion mining and sentiment analysis." *Mining text data*. Springer, Boston, MA, 2012. 415-463.
- Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1), 1-167.
- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5), 513-523.

4. Yahav, I., Shehory, O., & Schwartz, D. (2018). Comments Mining With TF-IDF: The Inherent Bias and Its Removal. *IEEE Transactions on Knowledge and Data Engineering*. Doi: 10.1109/TKDE.2018.2840127
5. Ben Ami Z., Geva T., and Yahav I. (2018), The Information Content of Multiword #Hashtags (short paper). *The International Conference on Information Systems (ICIS) 2018*, San Francisco, USA.
6. Hua, W., Wang, Z., Wang, H., Zheng, K., & Zhou, X. (2017). Understand short texts by harvesting and analyzing semantic knowledge. *IEEE transactions on Knowledge and data Engineering*, 29(3), 499-512.