אוניברסיטת חיפה

החוג לניהול מידע וידע

מספר הקורס :

שם הקורס: **ניתוח נתונים בהיקפים גדולים**

שם המרצה: **דר' רון בקרמן**

שעות קבלה: בתאום מראש

דואר אלקטרוני: ronb@univ.haifa.ac.il

הקורס הוא השני והמתקדם בסדרת הקורסים לחקר נתונים. הקורס הזה מתרכז סביב שני נושאים בולטים בחקר הנתונים: אנליטיקה חיזויית (Predictive Analytics) ועיבוד נתוני עתק (Big Data). לשני הנושאים יש חשיבות עסקית חסרת תקדים אך הם רחוקים מלהיות מפותחים כראוי. אנו מצפים התפתחויות משמעותיות בשני התחומים בשנים הקרובות - ורוצים לצבור היתרון התחרותי על ידי התעמקות בנושאים האלה בשלב מוקדם יחסית של התפתחותם.

אנליטיקה חיזויית עוסקת בבניית מודלים סטטיסטיים לחיזוי וגילוי תופעות. הבעיה המרכזית בתחום היא בעיית קלסיפיקציה (סיווג): שיוך של נתונים לקטגוריות המוגדרות מראש. אם נדע לחלק את אוסף הנתונים לקטגוריות, נוכל להתאים טיפול לכל קטגוריה בנפרד, וגם לדעת איך מטפלים בנתונים חדשים שיבואו בעתיד. נלמד איך משתמשים בקלסיפיקציה, מהן הגישות הידועות ביותר, מה יתרונותיהן וחסרונותיהן. בנוסף, ניגע גם בתת-תחומים אחרים של אנליטיקה חיזויית כגון אשכול (Clustering).

עיבוד Big Data נותן לנו הזדמנות ללמוד משהו חדש ומעניין על העולם שלנו - למשל, על התנהגות אנשים, על תהליך קבלת החלטות בארגונים, על התנהלות שווקים וכו'. הידע הזה מסתתר בתוך גיגא-ביתים וטרא-ביתים של נתונים שאותם עד לא מזמן לא ידענו לאגור ולא ידענו לעבד. אבל בזמננו, עם התפתחויות של טכנולוגיות חדשות, אנו מסוגלים להפיק מסקנות מכמויות אדירות של נתונים שנצברו. נלמד איך ניגשים אל הבעיות כאלו ומהן הטכניקות הפשוטות ביותר לאנליטיקה במאגרי נתוני עתק.

בקורס הזה נעבוד באותה המתכונת כמו שבקורס הקודם: השיטה תהיה פרקטית עד כמה שאפשר. בכיתה נעבוד עם מחשבים ניידים ונכתוב תוכניות קצרות שבעזרתן נעבד את הנתונים. בבית נעבור על חומרי לימוד, נראה הרצאות מוקלטות ונתכונן לעבודה משותפת בכיתה. לכל אחד ואחת תהיה הזדמנות להתבטא ולהתבלט בכיתה - לשאול שאלות רלוונטיות, להראות ידע וחריצות, ללמד ולעזור לסטודנטים אחרים. בנוסף, כל סטודנט יבחר פרויקט אישי שיבצע במהלך הקורס. בקורס הזה לא יהיה מבחן סופי.

חלק מתרגילי הכיתה והפרויקטים יבוצעו במערכות ענן של חברת אמזון. הביצוע כרוך בתשלום של סכומי כסף סמליים יחסית (בסביבות 100 ש"ח סך הכל). עם זאת, קל מאוד לבזבז סכומי כסף גדולים על הרצת תוכניות לא תקינות. נלמד לעשות מעקב צמוד אחר ההרצות שלנו ולהתגונן בפני תופעות לא רצויות.

- **דרישות קדם**

ידע בסיסי בתיכנות. היכרות עם מונחים חשובים מעולם התיכנות, כגון "זמן ריצה", "כמות זיכרון" וכו'.

- **הרכב הציון הסופי**

32% - השתתפות פעילה ועניינית בכיתה (8 שיעורים 4% כל אחד).
35% - תרגילי בית בהגשה אישית (7 תרגילים - 5% כל אחד).
33%- פרויקט אישי.

- **ספר הלימוד לקורס (מומלץ אך לא נדרש)**

http://www.amazon.com/Data-Science-Business-data-analytic-thinking/dp/1449361323

- **תוכן הקורס על פי נושאים ולוח זמנים**

28/02/17	*מבוא. אלגוריתמי קלסיפיקציה.*
07/03/17	*עקרונות עיבוד נתוני עתק. MapReduce.*
21/03/17	*המחשה וצבירת סטטיסטיקות על נתוני עתק.*
28/03/17	*אנליטיקה וצליבת נתונים ממקורות שונים.*
04/04/17	*כלים מתקדמים לעיבוד נתוני עתק. Spark.*
18/04/17	*הרצאת אורח.*
25/04/17	*תרגיל כיתה.*
25/04/17	*הצגות פרוייקטים.*