מבוא להנדסת נתונים הרצאה מס' 1

ביג דאטא השתנו כמה דברים בזמן האחרון:

1. כמות נתונים גדלה בצורה מסחררת

2. סוג הנתונים

3. קצב הנתונים דברים בזמן אמת

מרחב גדול של מקורות מידע כגון טלפונים אפליקציות מצלמות מסרונים ועוד

מלא חברות ומוסדות מעלות את כל המידע שלהם לאינטרנט כמו אסטרונומים או אפילו עיריות

המידע הזה טוב כדי להשתמש בו כדי לחזות בו תבנית התנהגות אצל אנשים

Kaggle :

מקום שבו מעלים נתונים ועושים תחרויות מי מצליח לשפר את האלגוריתם שלהם תמורת פרס כספי

כלים של מדען נתונים :

1. pyTorch mongo DB MySQL (דברים לשמור נתונים)

2. להציג נתונים

3. לחלוק נתונים

הענן מקומות : אמאזון אזור ועוד

יש נתונים כלים ומחשוב שהתקדם מאוד

משימות וחלוקת ציון

6-8 עבודות בית 20%

בוחן 20%

מבחן סוף סמסטר 60%

נושא ראשון: ממסדי נתונים

מידע מובנה : מידע שיש לנו טבלאות נגיד בפייסבוק מי החברים שלנו

מידע לא מובנה: טקסטים, קטעי קול

למה לא עובדים עם קבצים ?

מידע כפול קשה להנגיש ועוד

מה זה ממסדי נתונים :

למשל וויקיפדיה ,דפים , משתמשים , היסטוריה של דפים , קישורים

תוכנה שעוזרת לשמור על נתונים בצורה מובנת אפשר לגשת עליה בצורה קלה ועוד

איפה יש ממסדי נתונים?:

בנקים

חברות טלפונים

רשות חברתיות

אתרי מכירות

כדי לגשת לממסדי נתונים sql

Being a data janitor

לחזות את העתיד הכי שימושי

דוגמה בחירות מי ינצח לחזות מסלול של הוריקן לחזות את הגרף של מנייה

עץ החלטה עוזר בסטטיסטקה של לחזות את העתיד מאוד שימושית

רגרסיה אחד הדברים הכי שימושים בעולם

ניתוח אשכולות חשוב מאוד כדי למצוא את הקשר בין הנתונים

כמו ווקטורים טקסטים תמונות ועוד...

Data warehouse

Business intelligence התהליך ליעל את המידע

שפות תכנות שנשתמש בו

-python

להוריד PyCharm IDE

-R R studios