יישום שיטות לניתוח נתונים – פרוייקט

תאריך הגשה: 15.8.2016

מטרה:

מטרת הפרוייקט הינה להתנסות בתהליך עבודה עם נתונים תוך שימוש בשיטות למידת מכונה מתקדמות. הנתונים שינותחו הינם נתוני רצפים, תוכלו להשתמש בנתונים מתחומים שונים, כגון: טקסט, גנים, וידאו, שמע, סדרות עיתיות וכו'. בפרוייקט תידרשו להשתמש בשיטה מתחום ה deep learning (Recurrent Neural Network) לטובת חילוץ מודל העומד מאחורי נתוני הרצפים ולייצר רצפים חדשים באמצעות המודל הנלמד. תוצרי הפרוייקט הם קוד ודו"ח מסכם המתאר את הנתונים, תהליכי העיבוד שעברו, תהליך המידול והערכת תוצאות השימוש בו.

דרישות כלליות:

1. הפרוייקט יתבצע בזוגות.
2. ניתן להשתמש באוסף נתונים קיים או לבצע crawling ויצירה של אוסף נתונים מקורי בתיאום עם המרצות.
3. לצורך ביצוע הפרוייקט יש להשתמש במגוון הכלים שנלמדו בקורס. כמו כן, תידרש למידה עצמית על מנת שתוכלו ליישם את השיטות בעולם הנתונים הנבחר.
4. את הפרוייקט יש להגיש אל תיקיית הגשת הפרוייקט באתר הקורס.
5. יש לאשר את אוסף הנתונים בו תשתמשו מול מרצות הקורס לא יאוחר מ XXX. לא תתאפשר עבודה של מספר זוגות על אותו אוסף נתונים.
6. "עלות" איחור בקבלת אישור לפרוייקט – 1 נקודה לכל יום איחור!
7. אין להעתיק עבודה או חלקי עבודות מתלמידים אחרים או **מהאינטרנט**.
8. בקורס זה העבודות תתבצענה בזוגות. משמעות הדבר היא כי על הזוג לעבוד יחד לביצוע המשימה. בפרט, במידה וידרשו הסברים אודות עבודתכם, שני חברי הצוות ידרשו להציג בקיאות בעבודתם.

תוכן הפרוייקט:

1. בחירת אוסף נתוני רצפים. ניתן לאסוף את הנתונים באופן עצמאי או להשתמש במאגר נתונים קיים.
2. תיאור הנתונים: תיאור אוסף הנתונים שבחרתם, מהם האתגרים העיקריים בעבודה עם נתונים אלו? מדוע המחקר מעניין והיכן הוא יכול לתרום?
3. תיאור שלבי עיבוד מוקדם שהופעלו על הנתונים לטובת הבאתם לפורמט המתאים לשמש קלט לאלגוריתם הלומד.
4. כתיבת קוד python לאלגוריתם RNN שבאמצעותו יופק מודל לשחזור הנתונים.
5. כתיבת קוד לחילול נתוני רצפים ע"פ המודל שנלמד.
6. בניית מודל והרצתו תוך שימוש בו לטובת חילול מידע.
7. הערכת איכות המידע המשוחזר באמצעות השוואת הרצפים המסונתזים לרצפי המקור (כל רצף מסונתז יושווה לרצף הכי דומה לו בסט המקורי, לבסוף יחושב ממוצע). השתמשו במדד דמיון מתאים לטובת המשימה.
8. ניתוח תוצאות המחקר והסקת מסקנות.
9. כתיבת דוח מסכם. בפורמט MD.

משימות בונוס:

1. התמודדות עם big data שדרשה התאמות נוספות (שימוש ב GPU) – עד 5 נקודות.

2. עבודה עם עולם נתונים מורכב שלא הוצג בהרצאות, הדורש התמודדות עם אתגרים חדשים לטובת ביצוע המידול – עד 10 נקודות.
3. תרומה מיוחדת לכיתה, כגון שיתוף סקריפט הכולל אוסף התקנות רלוונטיות בpython, הפניה ל AMI רלוונטי ב AWS וכדומה (כולל תיעוד, וכמובן שעל השיתוף להתפרסם במועד מוקדם ככל האפשר כך שיוכל לשמש את הסטודנטים בפרוייקט זה) – עד 2 נקודות.

הנחיות להגשה:

1. על כל אחד מבני הזוג להגיש את העבודה.
2. נא לוודא כי הקובץ המצורף יהיה מכווץ בפורמט zip, וכי שם הקובץ יהיה מהסוג ID1_ID2.zip (תעודות הזהות של המגישים מופרדות באמצעות קו תחתון).
3. על ההגשה לכלול את הקבצים הבאים:
   a. דו"ח הפרוייקט כתוב כ MD (אפשר גם פלט PDF של פורמט MD)
   b. סט הנתונים בו השתמשתם.
   c. קבצי קוד רלוונטים, הכוונה לכל קוד ששימש לאיסוף ו/או ניתוח הנתונים.

<u>רעיונות להפעלת השיטה על נתונים שונים:</u>

- בפוסט הבא ניתן לראות הדגמה של חילול מאמרים של פול גרהאם, ספרות של שייקספיר, כתבות ויקיפדיה, הגדרות אלגבריות בlatex, קוד מקור בלינוקס, ושמות תינוקות:

http://karpathy.github.io/2015/05/21/rnn-effectiveness/

- ניתן למצוא את רשימת ההשמעה של גלגל"צ בקישור: https://t.co/MwnMeVbQTD
  - (ניתוח חביב של רשימת ההשמעה של גלגל"צ ניתן למצוא כאן).
- ניתוחים נוספים של טקסט / מוסיקה:
  - @nylk trained char-rnn on cooking recipes. They look great!
  - @MrChrisJohnson trained char-rnn on Eminem lyrics and then synthesized a rap song with robotic voice reading it out. Hilarious :)
  - @samim trained char-rnn on Obama Speeches. They look fun!
  - João Felipe trained char-rnn irish folk music and sampled music
  - Bob Sturm also trained char-rnn on music in ABC notation
  - RNN Bible bot by Maximilien
  - Learning Holiness learning the Bible

אלו רק מספר דוגמאות מתוך שפע רב של דוגמאות שניתן למצוא online.

בהצלחה!