תרגיל בית. כריית טקסט (בונוס למנצחים)

 ד"ר תומר גבע

תאריך הגשה: 26.1.2023 (עד שעה 15:45)

**הערה חשובה:**

תרגיל זה יכול לשמש כחלק מפתרון הפרוייקט המסכם. מאחר שהפרוייקט המסכם נועד לבחון ידע והינו מקביל לבחינה (**בחינת אמון, חסויה**) – כללי הפרוייקט המסכם חלים גם על תרגיל זה.

* לא ינתנו תשובות ע"י מרצה הקורס כיצד לפתור תרגיל זה
* מותר להשתמש: בחומר שזמין באינטרנט באתרים ציבוריים שעוסקים בתחום כולל פורומים עם תשובות לשאלות (אין להעלות שאלות חדשות אלא להסתמך על חומר קיים), ספרות מקצועית (אונליין או ספרים שונים), תיעוד תוכנות כגון Python, R או תוכנות אחרות (אונליין או ספרים שונים), חומר הקורס הנוכחי, קורס המבוא, חומרים מקורסים אחרים בפקולטה
* אין להעזר בחומרים אחרים.
* אין להעזר באנשים אחרים (באופן ישיר או עקיף) למעט השותף/פה בפרוייקט.
* אין להעביר/להפיץ תוכן תרגיל זה או לגבי פתרונו לאנשים אחרים
* במידה ויתעורר חשד לגבי טוהר הבחינות או פתרון שנעשה באופן "לא עצמאי" – המרצה רשאי לבחון בכתב או בע"פ את הסטודנטים, לגבי בקיאות באלמנטים טכניים ושיקולים נוספים בדרך הפתרון המומלצת. ציון הבחינה ישוקלל בציון התרגיל. זאת, בנוסף לטיפול בנושא בדרכים הרגילות באוניברסיטה.
* דוגמאות לסימנים מחשידים שעלולים להוביל לבחינה: תוצאות אמפיריות שלא נראה שתואמות לדרך הפתרון המתוארת, חוסר תיעוד של חלקים משמעותיים בדרך הפתרון, שימוש במתודולוגיות שלא נלמדו בקורס ללא מתן נימוקים תיאורטיים מעמיקים ותמיכה אמפירית לגבי נסיבות השימוש בהם, תוצאות חריגות לטובה (לתרגיל זה ישנן אינספור דרכים לפתרון. ללא דרך פתרון אחת "נכונה". לכן הסיכוי לקבל פתרונות דומים ע"י קבוצות העובדות באופן עצמאי הינו אפסי).

**© כל הזכויות שמורות.**

**על מנת לשמור על טוהר הבחינות – אין להפיץ מידע על תוכן תרגיל זה או פתרונו**

**רקע**

במשימה זו תהיה לכם ההזדמנות להתנסות עם טכניקות בכריית מידע מטקסט. מטרת התרגיל הינה לתרגל בניית מודלים לחיזוי/סיווג דעת הדובר וגישתו כלפי מוצר מסוים על סמך טקסט שכתב (Sentiment Analysis). מודלים של Sentiment Analysis משמשים בתעשייה לצרכים רבים, כולל מתן מענה מהיר לתקלות איכות, זיהוי מגמות, ניתוח שוק ומיצוב תחרותי, ובניית מודלים לחיזוי של מכירות.

**ביצוע התרגיל**

מעבר לאלמנט של הלימוד וההתנסות לתרגיל זה יש עוד שתי תועלות עיקריות:

1. המודלים לסיווג sentiment אשר תבנו יוכלו לסייע לכם בתהליך ביצוע הפרוייקט המסכם.
2. על מנת לעודד ביצועים טובים אנו נקיים תחרות . לקבוצות המובילות בתחרות מובטח בונוס לציון (בהתאם לביצועים).

**ההגשה בזוגות בלבד. הגשת הפרויקט תהיה ע"י אותם זוגות.**

**בשביל למנוע טעויות, על כל אחד מבני/ות הזוג להגיש בנפרד את הקובץ.**

**שימו לב - על כל אחד מחברי/ות הקבוצה להגיש בנפרד, בשתי תיבות מטלות נפרדות בMoodle:**

* **תיאור של עד 2 עמודים של דרך הפתרון.**
* **קובץ המלצות (כמפורט בהמשך)**

**הבעיה הנבחרת**

חברה מעוניינת לנתח באופן אוטומטי ביקורות של לקוחות אשר נכתבות באחד מהאתרים המרכזיים בתחום זה. החברה תשתמש בתוצאות הניתוח לזהות מגמות בדעה כלפי החברה, זיהוי לקוחות מתוסכלים, זיהוי בעיות עם מוצרים ועם מתן שירות וחווית השירות, וכן בהמשך לצורך התאמה אישית של הצעות שיווקיות. החברה פנתה אליכם על מנת לסייע לה.

לשם כך, עליכם לבנות **מודל Sentiment Analysisשיזהה איך לקוחות מרגישים כלפי מוצרים אותם רכשו** .

בעוד שקיימות גישות שונות של דיווח sentiment analysis כגון מתן ציון מספרי מ 0-5 או מ1+ עד 1-, החברה מבקשת שתסווגו את תוכן ההודעות האם הן חיוביות? {0-לא, 1-כן}

הנהלת החברה מעוניינת להריץ את המודל שתבנו (במידה ויענה על דרישותיה מבחינת איכות) על כל חוות הדעת שבמאגר שלה (מאות אלפי חוות דעת).

לשם כך, עומד לרשותכם מאגר רשומות של 2000 חוות דעת שסופקו ע"י החברה שלהן ידועה הקטגוריה (עובדי החברה עברו ידנית על 2000 רשומות וסייוגו אותן). כמו כן החברה כבר ביצעה עיבוד מקדים לטקסט וסדרה את הנתונים בפורמט טבלה פשוטה בעלת 2000 עמודות כאשר כל עמודה מונה כמה פעמים הופיעה מילה מסוימת (המילים הכלולות הן 2000 המילים הכי פופולריות בטקסט) ו2000 שורות כאשר כל שורה מייצגת חוות דעת אחרת. כאשר:

**מספר עמודה שם עמודה תיאור**

2002 rating משתנה המטרה, האם חוות הדעת הייתה חיובית? {0-לא, 1-כן}

2-2001 words משתנים מסבירים, כל שם עמודה מייצג נוכחות של מילה מסוימת

1 ID מספר סידורי של חוות הדעת בבסיס הנתונים.

לדוגמה: חוות הדעת השלילית

“I had been a Gold member for many years until I have encountered rude customer service”

תיוצג כך:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Rating | help | … | japan | in | customer | until | for | member | gold | I | had |
| 0 | 0 | ... | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 2 | 2 |

* הערך 0 בעמודת rating מצביע על כך שחוות הדעת היא שלילית.
* בשאר העמודות המספרים 0,1,2,… מונים כמה פעמים הופיעה כל מילה בחוות הדעת.
* כחלק מהליך העיבוד המקדים הופעל כמקובל שיטת stemming. לכן מילים בעלות אותו "שורש" כמו had ו-has נספרות כאותה המילה. כמו כן סולקו חלק ממילות הקישור ומילים שלדעת החברה לא רלוונטיות.
* רוב התאים מכילים 0 כיוון שרק חלק קטן מהמילים מופיעות בכל חוות דעת.

**שלב א – התאמת מודל**

בנו מודל סיווג כך שבהיתן חוות דעת חדשה, המודל יחזה האם היא חיובית או לא חיובית.

* + טיפ: יתכן שלא כל המשתנים המסבירים נחוצים לבניית מודל החיזוי. חשבו כיצד:
    - תבחרו את המשתנים המתאימים ביותר.
    - תקבעו את כמות המשתנים שתכניסו למודל.
    - או שיקלו שימוש באלגוריתמים שיודעים לבחור משתנים "באופן פנימי"
  + טיפ: חשבו כיצד אתם מתכוונים להעריך את ביצועי המודל:
    - באיזו שיטת ולדיציה עדיף להשתמש בבעיה זו (אין הרבה נתונים)? חלוקה לסט אימון וסט בחינה? או אולי K-folds cross validation?

**יש להגיש תיאור קצר של מקסימום שני עמודים של דרך הפתרון בתיבת ההגשה הרלוונטית במודל.**

**שלב ב – חיזוי עמדתם של חוות דעת חדשות**

כעת, על מנת להעריך עד כמה המודל שלכם טוב בחיזוי רשומות שהוא טרם ראה, ניתן לכם קובץ עם 2000 חוות דעת נוספות בשם text\_rollout\_X.csv ללא הסיווג שלהן (בקובץ זה לעמודה rating נבחר באופן שרירותי הערך 0) . עבור כל אחת מחוות הדעת בסט זה, עליכם לחזות האם היא לא חיובית (=0) או חיובית (=1).

המדד על פיו החברה תעריך את הביצועים שלכם (ועל פיו יקבעו הקבוצות הזוכות) הוא אחוז חוות הדעת שחזיתם נכון, קרי:

(זהו מדד accuracy על פני כל הקטגוריות)

**בסוף התחרות, הקבוצות שהשיגו את הביצועים הטובים ביותר יקבלו בונוס לציון הקורס.**

**פורמט הקובץ שעליכם לייצור:**

יש למלא את עמודת ה- ratingבקובץ CSV ששמו recommendations.csv. (מצורף דוגמא לקובץ זה בו ערכי העמודה של המשתנה התלוי נקבעו שרירותית בתור 0).

בעמודה של המשתנה התלוי יש שני ערכים אפשריים: 0 או 1 (כברירת מחדל עמודה זו מכילה רק אפסים).

1 - במידה ואתם חושבים שחוות הדעת חיובית

0 - במידה ואתם חושבים שחוות הדעת שלילית

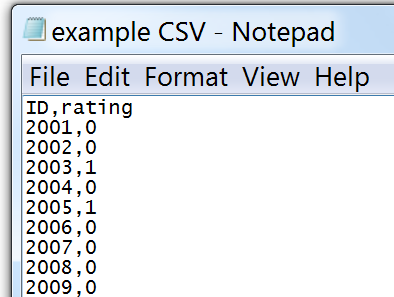
**שימו לב:** יש לשמור על הפורמט של קובץ זה גם אם אתם מייצרים אותו מחדש בעצמכם.

1. אין לשנות בו את עמודת ID, אין להוסיף\למחוק בקובץ עמודות או שורות.
2. סדר ההמלצות אמור להיות זהה לסדר הרשומות בקובץ text\_rollout.

**עליכם לוודא שהקובץ שנוצר תואם את הפורמט המוגדר כאשר נפתח לבדיקה כקובץ CSV במערכת הפעלה windows.**

**על מנת לתת לחברה פתרון end to end יש להקפיד על הפורמט והסדר של הקובץ בחלק ב'. הבדיקה מתבצעת באופן אוטומטי. הגשה בפורמט לא מתאים או סדר לא נכון עלולה לפגוע קשות בציון (ינתן ציון אפס).**

**ראו דוגמא לקובץ ההמלצות:**



תמונה 1- דוגמא לקובץ ההמלצות. הערכים בדוגמא נבחרו שרירותית