בס"ד

**מבוא לבינה מלאכותית - תרגיל 2**

הבעיה – נתונים הקבצים הבאים:

* קובץ train.txt בו נתונות דוגמאות המתארות מאפיינים שונים.
* קובץ test.txt בו נתונות דוגמאות אשר את הסיווג שלהן צריך לחזות.

חיזוי זה ייעשה באמצעות האלגוריתמים Decision Tree ו- naïve base.  
יש להריץ כל אחד מאלגוריתמים אלו על קובץ train.txt.

כתוב תוכנית הקוראת מקובץ train.txt את סט הדוגמאות כאשר, **השורה** הראשונה של קובץ זה תכלול את שמות השדות (כלומר, את המאפיינים של הנתונים)   
הערכים האפשריים של כל מאפיין הם הערכים שמופיעים בעמודת המאפיין בקובץ ה-train (לא יהיו ערכי מאפיינים שלא יפיעו בקובץ).  **העמודה** האחרונה בכל שורה הינה הסיווג (הclass-).   
כל הערכים מופרדים ב <tab>.

דוגמא לקטע מקובץ הtrain.txt- אותו תקבלו:

pclass age sex survived

crew adult male no

1st adult male no

2nd adult female yes

כפי שניתן לראות בקובץ זה, ישנם 3 מאפיינים (pclass, age, sex) והסיווג הוא (survived).  
הערכים האפשריים היחידים של מאפיין sexבדוגמא זו הם: male, female

Decision Tree

כתבו פונקציה שמיישמת את אלגוריתם ID3.  
את העץ שנבנה מהאלגוריתם יש להדפיס לקובץ בשם output\_tree.txt בפורמט הבא:

<attribute\_name>=<attribute\_value>  
<tab>|<attribute\_name>=<attribute\_value>:class

לדוגמה:

age = child

| pclass = crew: yes

| pclass = 1st: yes

| pclass = 2nd: yes

| pclass = 3rd: no

age = adult

| pclass = crew: no

| pclass = 1st: yes

| pclass = 2nd: no

| pclass = 3rd: no

שים לב שדוגמה זו אינה מכילה את כל התכונות ורק מדגימה את הרעיון הכללי של פורמט הפלט.

Naïve Base

כתבו פונקציה המממשת את חיזוי naïve base.

לאחר בנית המודל לכל אחד מהאלגוריתמים, יש לבצע חיזוי לדוגמאות שבקובץ test.txt.

הדפס לקובץ output.txt את החיזוי בפורמט הבא:

<DT\_prediction>**tab** <naiveBase\_prediction>

**בשורה האחרונה** בקובץ יכתב הaccuracy של כל אחד מהחיזויים (באותו פורמט).

יש להגיש:

* קובץ details.txt בו יש לכתוב את שמות המגישים בצירוף מספרי ת.ז.
* קובץ py\_ex2 אשר יכיל את הקוד. (יש לתעד את הקוד)

לצורך הדוגמא נתונים הקבצים הבאים:

Train.txt – המכיל הדוגמאות המתארות מאפינים שונים של נוסעים שהיו על הטיטאניק, והאם שרדו או לא.

Test.txt - מכיל דוגמאות מתוייגות נוספות

Output\_tree – כמתואר לעי"ל. הקובץ להדגמת הפורמט בלבד.

Output.txt – הקובץ להדגמת הפורמט בלבד.

(**הערה!!!:** בעת בדיקת הקוד לא בהכרח נשתמש בנתוני טיטניק בלבד; לכן, אין להשתמש ב hard-code לנתונים ספציפים.)

**בהצלחה!**