עבודה 1:

מגיש: איתי אוסובלנסקי 311129274

הסבר על הקוד:

הקוד מממש policy iteration לפי האלגוריתם שנלמד בשיעור תוך כדי שימוש בנוסחאות בלמן עבור מטריצות. במימוש יש תחילה הגדרה את סביבת העבודה (Taxi-V3) ופרמטרים נדרשים לאלגוריתם ולבדיקת נכונותו (gemma, delta, check_state...). לאחר מכאן, חישוב המדיניות האופטימלית וקריאה לשתי השיטות שנתבקשו במשימה (שנקראות simulateGame,valueFunction). שמריצים את התוכנית מודפס למסך משחק (אחד) ומטריצה של פוקנציית המצבים (לכל המצבים). חישוב מדיניות אופטמלית נעשתה באופן הבא:

שיטות מרכזיות:

- Eval_Reward_Prob: שיטה שמקבלת את הסביבה ומאתחלת את המטריצות הפרס והסתברויות המעברים.
- Policy_iteration: תחילה מאתחלת את פונקיות הערך והמדיניות ואז מריצה בלולאה את עדכון המדיניות הנוכחית. הלולאה תסתיים כאשר המדיניות הנוכחית "זהה" למדיניות שחושבה מקודם (בהפרש של delta). במקרה שאין התכנסות למדיניות אופטימלית תוך max_iterations אז הלולאה תיעצר עם policy,V הנוכחיים. בכל לולאה אני תחילה מחשב את Q לכל state,action ואז סוכם כדי לקבל את V. ואז לפי מה שחושב, אני מעדכן את המדיניות (עם הערך המקסימלי עבור החישוב הנוכחי).
- simulateGame, valueFunction אלה השיטות שהוגדרו - הראשונה מסמלצת את המשחק של המונית עם ההדפסה של המשחק והתוצאה, והשנייה מחזירה מספר של פונקציית הערך לפי מספר המצב (מחזירה מספר, לא מדפיסה).