תרגיל בית 3 – MDP

**עברו על כלל ההנחיות לפני תחילת התרגיל.**

# **הנחיות כלליות:**

* תאריך ההגשה: עד ליום האחרון של הסמסטר - 08/04/2024ב23:59
* את המטלה יש להגיש **בזוגות בלבד.**
* יש להגיש מטלות מוקלדות בלבד. פתרונות בכתב יד לא ייבדקו.
* ניתן לשלוח שאלות בנוגע לתרגיל בפיאצה בלבד.
* המתרגל האחראי על תרגיל זה:  **דניאל אלגריסי.**
* בקשות דחיה מוצדקות (מילואים, אשפוז וכו') יש לשלוח למתרגל האחראי (**ספיר טובול**) בלבד.
* במהלך התרגיל ייתכן שנעלה עדכונים, למסמך הנ"ל – תפורסם הודעה בהתאם.
* העדכונים הינם מחייבים, ועליכם להתעדכן עד מועד הגשת התרגיל.
* שימו לב, העתקות תטופלנה בחומרה.
* התשובות לסעיפים בהם מופיע הסימון [Writing Hand on Apple ](https://emojipedia.org/apple/ios-14.6/writing-hand/) צריכים להופיע בדוח.
* לחלק הרטוב מסופק שלד של הקוד.
* אנחנו קשובים לפניות שלכם במהלך התרגיל ומעדכנים את המסמך הזה בהתאם. גרסאות עדכניות של המסמך יועלו לאתר. הבהרות ועדכונים שנוספים אחרי הפרסום הראשוני יסומנו כאן בצבעים. ייתכן שתפורסמנה גרסאות רבות – אל תיבהלו מכך. השינויים בכל גרסה יכולים להיות קטנים.

**שימו לב שאתם משתמשים רק בספריות הפייתון המאושרות בתרגיל (מצוינות בתחילת כל חלק רטוב)  
לא יתקבל קוד עם ספריות נוספות**

**מומלץ לחזור על שקפי ההרצאות והתרגולים הרלוונטיים לפני תחילת העבודה על התרגיל.**

**חלק א׳ – MDP (44 נק׳)**

**רקע**

בחלק זה נעסוק בתהליכי החלטה מרקובים, נתעניין בתהליך עם **אופק אינסופי** (מדיניות סטציונרית).

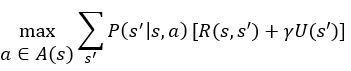
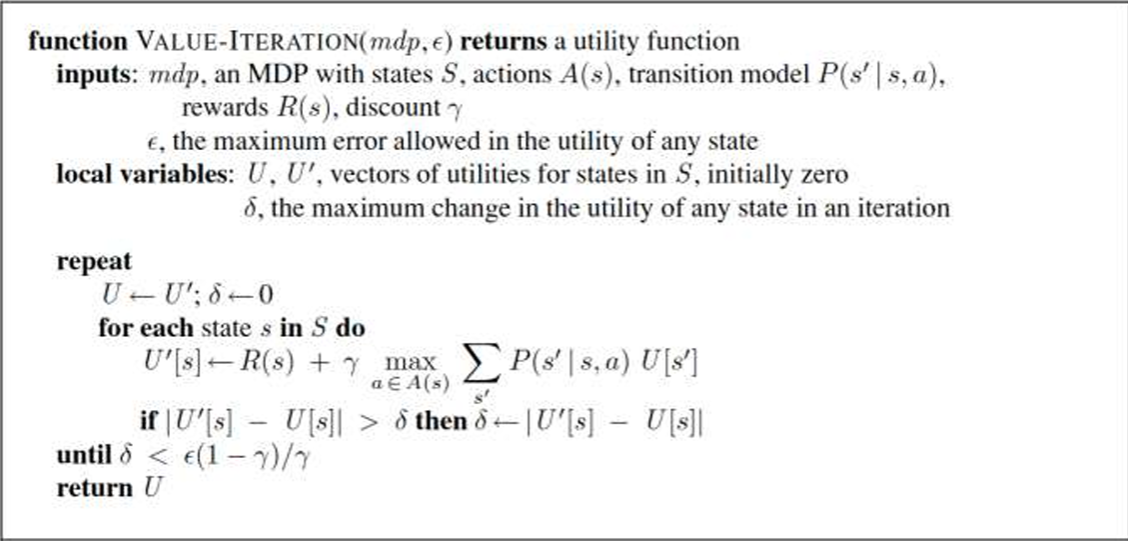
**חלק א׳ - חלק היבש** [Writing Hand on Apple ](https://emojipedia.org/apple/ios-14.6/writing-hand/)

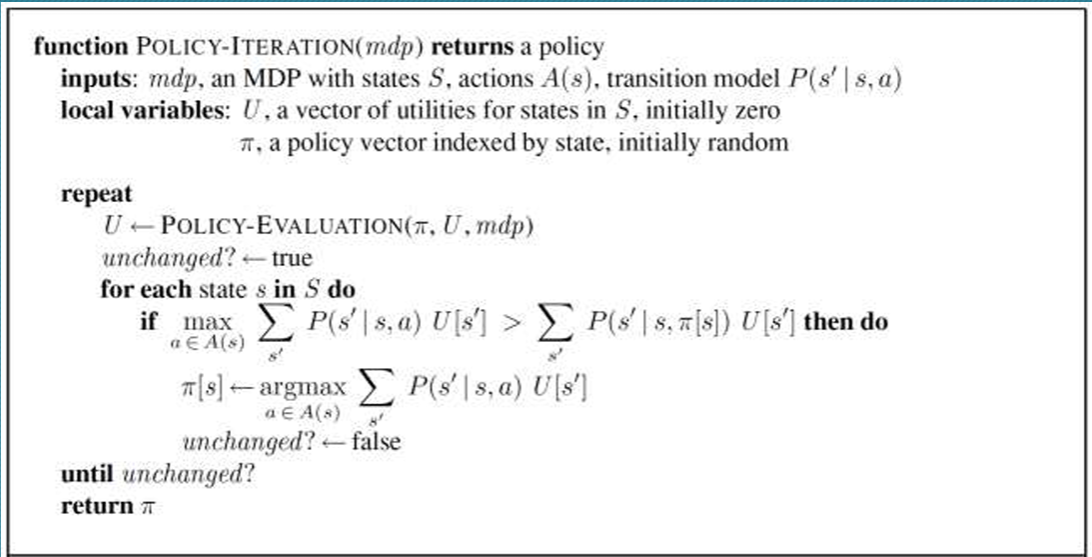
1. בתרגול ראינו את משוואת בלמן כאשר התגמול ניתן עבור המצב הנוכחי בלבד, כלומר , למתן תגמול זה נקרא "תגמול על הצמתים" מכיוון שהוא תלוי בצומת שהסוכן נמצא בו.   
   בהתאם להגדרה זו הצגנו בתרגול את האלגוריתמים Value iteration ו-Policy Iteration למציאת המדיניות האופטימלית.

כעת, נרחיב את ההגדרה הזו, לתגמול המקבל את המצב הנוכחי והמצב אליו הגיע הסוכן, כלומר: , למתן תגמול זה נקרא "תגמול תוצאתי". לצורך שלמות ההגדרה, נגדיר שאם לכל   
 מתקיים - אז .

1. (1 נק') התאימו את הנוסחה של התוחלת של התועלת מהתרגול, עבור התוחלת של התועלת המתקבלת במקרה של ״ תגמול תוצאתי ״, אין צורך לנמק.
2. (1 נק') כתבו מחדש את נוסחת משוואת בלמן עבור המקרה של ״ תגמול תוצאתי ״, אין צורך לנמק.

בסעיפים הבאים התייחסו גם למקרה בו , והסבירו מה לדעתכם התנאים שצריכים להתקיים על הסביבה\ על מנת שתמיד נצליח למצוא את המדיניות האופטימלית.

1. (2 נק') נסחו את אלגוריתם Value Iteration עבור המקרה של ״ תגמול תוצאתי ״.
2. (2 נק') נסחו את אלגוריתם Policy Iteration עבור המקרה של ״ תגמול תוצאתי ״.

**In policy iteration we use the expected utility function from a to calculate POLICY-EVALUATION, if then the Utility won’t necessarily converge in both Policy and Value iteration, meaning a finite horizon is required in order to find an optimal solution for this MDP.**

נתון הגרף הבא:



נתונים:

* (Discount factor) .
* אופק אינסופי.
* – קבוצת המצבים – מתארים את מיקום הסוכן בגרף.
* – קבוצת המצבים הסופיים.
* קבוצת הפעולות לכל מצב (על פי הגרף), לדוגמא: .
* תגמולים ("תגמול תוצאתי"):
* מודל המעבר הוא דטרמיניסטי, כלומר כל פעולה מצליחה בהסתברות אחת.

1. (יבש 2 נק') הרץ את האלגוריתם Value iteration שכתבת על הגרף הנתון. ומלא את הערכים בטבלה הבאה, כאשר . (ייתכן שלא צריך למלא את כולה).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  | 5 | 5 | 5 | 5 | 0 |  |
|  |  |  |  | 7 | 7 | 7 | 7 | 0 |  |
|  |  |  |  | 1.5 | 1.5 | 1.5 | -1 | 0 |  |
|  |  |  |  | 2.5 | 2.5 | 2.5 | -1 | 0 |  |
|  |  |  |  | 2.5 | 2.5 | 2.5 | -1 | 0 |  |
|  |  |  |  | 0.25 | 0.25 | -1.5 | -1 | 0 |  |
|  |  |  |  | 0.25 | 0.25 | -1.5 | -1 | 0 |  |

1. (יבש 2 נק') הרץ את האלגוריתם Policy iteration שכתבת על הגרף הנתון. ומלא את הערכים בטבלה הבאה, כאשר המדיניות ההתחלתית מופיעה בעמודה הראשונה בטבלה. (ייתכן שלא צריך למלא את כולה).

הניחו שבמידה ולא קיים שיפור, האלגוריתם יבחר תמיד להשאיר את הפעולה הקודמת.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  | Top-Left(S4) | Top-Left(S4) |  |  |  |  |

1. (יבש 2 נק') חיזרי על הסעיף הקודם. הפעם עם **אופק סופי כאשר** (שימי לב, המדיניות לא חייבת להסתיים במצב מסיים, ישנם מצבים שלא יכולים להגיע למצב מסיים עם אופק זה. ישנם צמתים עם מספר תשובות נכונות, נקבל את כולם).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

1. (1 נק') ללא תלות בשינוי של הסעיף הקודם. אם , מה מספר המדיניות האופטימליות הקיימות? נמקו.

means that there is no recursive calculation of the utility, in other words the only factor left is the reward, so this limits us to look ahead with depth 1, so in our case following the rewards given above we only get a positive reward in s1 and s2 and the rest have equal rewards for all available moves, so in total:

**1\*1\*2\*2\*2\*2\*3= 48 – Total policies**

1. (1 נק') ללא תלות בשנוי של הסעיף הקודם, הסבירי מה היה קורה אם

בתשובתך, התייחסי גם לערכי התועלות של כל צומת וגם לשינוי במדיניות, אין צורך לחשב.

Here the discount factor is 1, in addition to a positive reward on the path from s­1 to s2 and vice versa, we get the same results in the first iteration but after that s1 will want to go to s2 given that his utility plus the reward (according to our equation) with probability and discount factor 1 will give him more than what he already has (5+2=7), which will lead to the utility of ­s2 changing in the next iteration accordingly and for the same reasons (2+7=9), and so on…

The utility won’t converge as explained above, but the policy will converge to the same as before with the adjustment of s1 and s2 pointing at each other, never including the goal state.

**חלק ב׳ - היכרות עם הקוד**

חלק זה הוא רק עבור היכרות הקוד, עבורו עליו במלואו ווודאו כי הינכם מבינים את הקוד.

mdp.py – **אתם לא צריכים לערוך כלל את הקובץ הזה**.   
בקובץ זה ממומשת הסביבה של ה-mdp בתוך מחלקת MDP. הבנאי מקבל:

* board - המגדיר את המצבים האפשריים במרחב ואת התגמול לכל מצב, תגמול על הצמתים בלבד.
* terminal\_states – קבוצה של המצבים הסופיים (בהכרח יש לפחות מצב אחד סופי).
* transition\_function – מודל המעבר בהינתן פעולה, מה ההסתברות לכל אחת מארבע הפעולות האחרות. ההסתברויות מסודרות לפי סדר הפעולות.
* gamma – discount factor המקבל ערכים - .

בתרגיל זה לא נבדוק את המקרה בו .

הערה: קבוצת הפעולות מוגדרת בבנאי והיא קבועה לכל לוח שיבחר.

למחלקת MDP יש מספר פונקציות שעשויות לשמש אתכם בתרגיל.

* print\_rewards() – מדפיסה את הלוח עם ערך התגמול בכל מצב.
* print\_utility(U) – מדפיסה את הלוח עם ערך התועלת U לכל מצב.
* print\_policy(policy) – מדפיסה את הלוח עם הפעולה שהמדיניות policy נתנה לכל מצב שהוא לא מצב סופי.
* step(state, action) – בהינתן מצב נוכחי state ופעולה action מחזיר את המצב הבא באופן דטרמיניסטי. עבור הליכה לכיוון קיר או יציאה מהלוח הפונקציה תחזיר את המצב הנוכחי state.

**חלק ג׳ – רטוב**  
כל הקוד צריך להיכתב בקובץ mdp\_implementation.py

מותר להשתמש בספריות:

All the built-in packages in python, numpy, matplotlib, argparse, os, copy, typing, termcolor, random

עליכם לממש את הפונקציות הבאות:

* (רטוב 6 נק'): value\_iteration(mdp, U\_init, epsilon) – בהינתן ה-mdp, ערך התועלת ההתחלתי U\_init, וחסם העליון לשגיאה מהתוחלת של התועלת האופטמילי epsilon מריץ את האלגוריתם value iteration ומחזיר את U המתקבל בסוף ריצת האלגוריתם. TODO
* (רטוב 4 נק'): get\_policy(mdp, U) – בהינתן ה-mdp וערך התועלת U (המקיים את משוואת בלמן) מחזיר את המדיניות (במידה וקיימת יותר מאחת, מחזיר אחת מהן). TODO
* (רטוב 4 נק'): policy\_evaluation(mdp, policy) – בהינתן ה-mdp, ומדיניות policy מחזיר את ערכי התועלת לכל מצב. TODO
* (רטוב 6 נק'): policy\_iteration(mdp, policy\_init) - בהינתן ה-mdp, ומדיניות התחלתית policy\_init, מריץ את האלגוריתם policy iteration ומחזיר מדיניות אופטימלית. TODO

לאור העובדה שהפונקציות הבאות לא נוסחו באופן ברור בתרגיל, יצא הסבר מפורט יותר על הדיוקים הנדרשים פה ← <https://piazza.com/class/lrurdsbmuiww0/post/336>

* Shape

  Description automatically generated(רטוב 5 נק'): get\_all\_policies(mdp, U, epsilon=10\*\*(-3)) – בהינתן ה-mdp, וערך התועלת U (המקיים את משוואת בלמן) מדפיס\מציג את כל המדיניות המקיימות ערך זה בלוח בודד (יש לבצע ויזואליזציה להצגת כל המדיניות),

epsilon משמש אתכם לבדיקה אם שני ערכי שווים זה לזה.

לדוגמא:

הפונקציה מחזירה את מספר המדיניות (policies) השונות הקיימות המקיימות את U (בדוגמה ). TODO

הסימונים שניתן להעביר לפונקציית ההדפסה ↓↑→←,

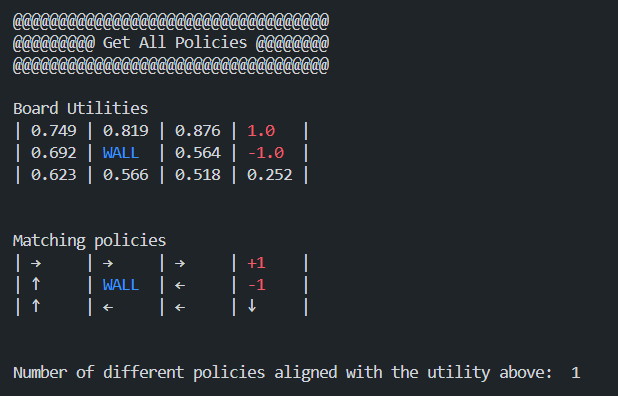
['UP', 'DOWN', 'RIGHT', 'LEFT'] או ['U', 'D', 'R', 'L'] (למעשה ניתן להעביר מה שתרצו בכל תא במטריצה) בכל תא, שימי את כל החיצים המתאימים למדיניות כלשהי.

למשל בדוגמה, עבור התא (1,1) נשים את כל החיצים.

עליכם להדפיס את הלוח (המדיניות) עבור התועלת שנמצאת בקובץ U\_for\_get\_all\_policies



**Running the segel Utility function:**



* (רטוב 5 נק'): get\_policy\_for\_different\_rewards(mdp, epsilon) – בהינתן ה-mdp מדפיס\מציג את המדיניות האופטימלית כתלות ב- R (ערכי התגמול לכל מצב שאינו סופי).   
  יש להחזיר רשימה של ערכי R שבהם יש שינוי במדיניות מהקטן לגדול. TODO

ניתן להניח שלא יהיו שינויים במדיניות עבור ערכי קטנים מ- וגדולים מ- . בנוסף, דיוק של 2 ספרות אחרי הנקודה הינו מספק.

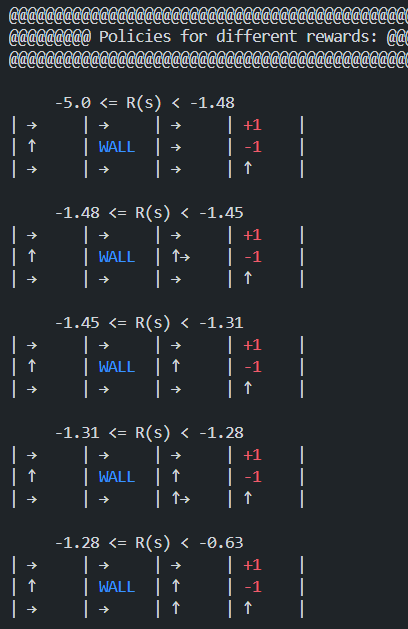
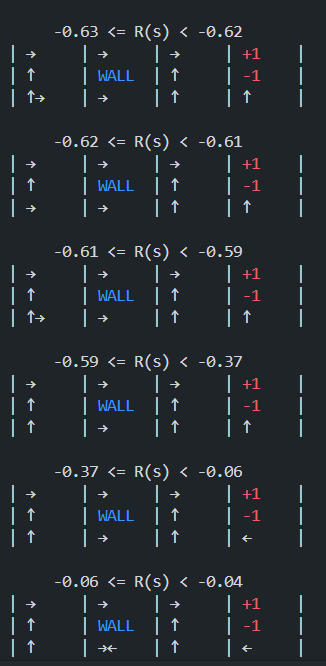
דוגמא חלקית של פתרון אפשרי:

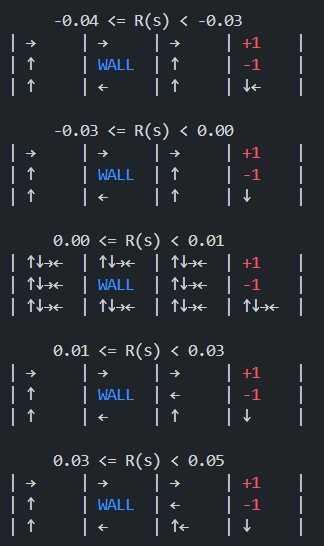
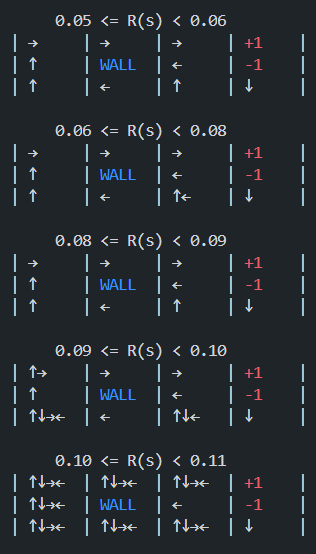
A picture containing text, crossword puzzle

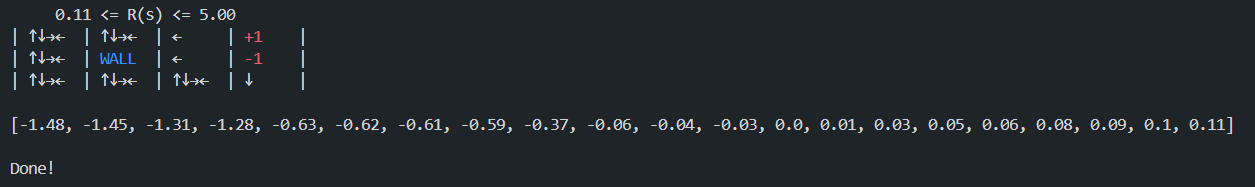
Description automatically generated

בנוסף לקוד עליכם לצרף להגשה היבשה את התצוגות של הפונקציות על הסביבה שניתנה בתרגיל.









**עבור מצבים סופיים וקירות (WALL), הערך שצריך לחזור בתאים אלו עבור טבלאות המדיניות הוא None. כל ערך אחר לא יתקבל כתשובה.**

main.py – דוגמת הרצה לשימוש בכל הפונקציות.

בתחילת הקובץ אנו טוענים את הסביבה משלושה קבצים:

board, terminal\_states, transition\_function

ויוצרים מופע של הסביבה (mdp).

* שימו לב, שכרגע הקוד ב-main לא יכול לרוץ מכיוון שאתם צריכים להשלים את הפונקציות הרלוונטיות ב- mdp\_implementation.py.
* בנוסף, על מנת לראות את הלוח עם הצבעים עליכם להריץ את הקוד בIDE לדוגמה PyCharm.

**הוראות הגשה**

* הגשת התרגיל תתבצע אלקטרונית בזוגות בלבד.
* הקוד שלכם ייבדק (גם) באופן אוטומטי ולכן יש להקפיד על הפורמט המבוקש. הגשה שלא עומדת בפורמט לא תיבדק (ציון 0).
* המצאת נתונים לצורך בניית הגרפים אסורה ומהווה עבירת משמעת.
* הקפידו על קוד קריא ומתועד. התשובות בדוח צריכות להופיע לפי הסדר.
* יש להגיש קובץ zip יחיד בשם AI3\_<id1>\_<id2>.zip (ללא סוגריים משולשים) שמכיל:
* קובץ בשם AI\_HW3\_MDP.PDF המכיל את תשובותיכם לשאלות היבשות.
* קבצי הקוד שנדרשתם לממש בתרגיל **ואף קובץ אחר:**
* mdp\_implementation.py

**אין להכיל תיקיות בקובץ ההגשה, הגשה שלא עומדת בפורמט לא תיבדק.**

נספח MDP:

דוגמת הרצה (שימו לב שהרצה זו השתמשה במודל הסתברותי שונה משלכם)

יצירת הסביבה:

mdp = MDP(board=board\_env,  
 terminal\_states=terminal\_states\_env,  
 transition\_function=transition\_function\_env,  
 gamma=1.0)

הדפסת הלוח עם התגמולים לכל מצב:

print('@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@')  
print("@@@@@@ The board and rewards @@@@@@")  
print('@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@')  
mdp.print\_rewards()

פלט:

תמונה שמכילה טקסט, צילום מסך, גופן

התיאור נוצר באופן אוטומטי

Value iteration:

print('@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@')  
print("@@@@@@@@@ Value iteration @@@@@@@@@")  
print('@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@')  
  
U = [[0, 0, 0, 0],  
 [0, 0, 0, 0],  
 [0, 0, 0, 0]]  
print("\nInitial utility:")  
mdp.print\_utility(U)  
print("\nFinal utility:")  
U\_new = value\_iteration(mdp, U)  
mdp.print\_utility(U\_new)  
print("\nFinal policy:")  
policy = get\_policy(mdp, U\_new)  
mdp.print\_policy(policy)

פלט:



Policy iteration:

print('@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@')  
print("@@@@@@@@@ Policy iteration @@@@@@@@")  
print('@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@')  
  
print("\nPolicy evaluation:")  
U\_eval = policy\_evaluation(mdp, policy)  
mdp.print\_utility(U\_eval)  
  
policy = [['UP', 'UP', 'UP', 0],  
 ['UP', 'WALL', 'UP', 0],  
 ['UP', 'UP', 'UP', 'UP']]  
print("\nInitial policy:")  
mdp.print\_policy(policy)  
print("\nFinal policy:")  
policy\_new = policy\_iteration(mdp, policy)  
mdp.print\_policy(policy\_new)  
  
print("Done!")

פלט:

