Java実行環境の整備

- MacあるいはWindowsでの実行に際して、javaをインストールすること。

  https://www.oracle.com/java/technologies/downloads/
  Windowsの場合は、インストーラでC:¥Javaの下にインストールし、システム環境変数でPATHを設定する。具体的にはシステム環境変数の編集画面で、環境変数にJAVA_HOMEを追加し、その値をC:¥Java¥binとする。

  Macの場合は、brew install openjdkで、インストール可能。

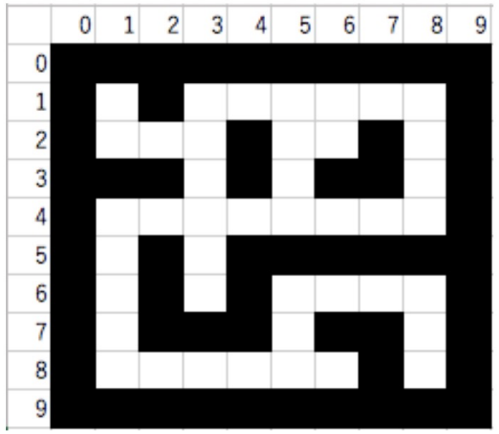
- ターミナル（macの場合）/DOSプロンプトウィンドウ（Windowsの場合）を起動後、java -versionでバージョン番号を含めたものが出力されるかをを確認する（メッセージは、各自の環境で若干は異なる）。

Macでの出力例）

```
openjdk 21.0.1 2023-10-17
OpenJDK Runtime Environment Homebrew (build 21.0.1)
OpenJDK 64-Bit Server VM Homebrew (build 21.0.1, mixed mode, sharing)
```

Q学習の適用

- Q学習は、環境との相互作用で最適解を探索していくことから、環境が変化するような状況にも頑健なアルゴリズムである。
- そこで、迷路探索のようにスタートからゴールまで試行錯誤を繰り返して、壁(行止り)に当たると負の報酬がえられ、これまでの探索経路を戻って再探索し、ゴールに到達すれば正の報酬が得られるようにするとQ学習が適用できる。
- ただし、学習率や割引率というハイパーパラメータと呼ばれるものの設定問題が残る。



迷路問題とQ学習（1）

- 迷路でのスタート地点からゴール地点までの経路探索は、内部的にはエージェントが移動ステップごとに移動先を選んで進み、ゴールに到達すれば正の報酬を、壁に当たったり、試行を繰り返すたびに負の報酬(ペナルティ)が与えられる。
- 各ステップでの行動は、エージェントが行動集合$\{a_t\}$から選択して実行する。今回の迷路探索の場合の行動集合$\{a_t\}$は、平面上で「上、下、左、右」への4つの移動となる。
- 迷路上でのそれぞれの位置(セル)において選択した行動の結果、ゴールに到達できたということは、その行動を以後の探索においても選択した方が良い、逆に壁に当たったというこ とは選択しない方が良い、ということを記憶するために、それぞ れの位置においてQ値と呼ぶものを用いる。
  - Q値はある位置での行動で移動した後には、以降の行動(将来)は最適な行動を取れるとして、ある位置での行動の価値を求めたものである。
  - このことから、行動価値関数と呼ばれる。

迷路問題とQ学習（2）

- 時刻tの時の状態s_tにおいて、行動a_tにより状態s_{t+1}に遷移し、報酬r_{t+1}を受け取ったとすると、時刻t+1における最適行動価値関数は、再帰式として表すことができる。未来に受け取る報酬は割り引いて考え、割引率γ($0 \leqq \gamma \leqq 1$)を用いる。
- 時刻tから時刻t+1の状態へ遷移する際に、理想的には上記の考えで学習していくが、常に理想的に学習していけるとは限らないため、学習率α($0 \leqq \alpha \leqq 1$)を導入する。このことから、状態遷移するに従って、Q値を以下の式で更新する。

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(r_{t+1} + \gamma \max_{a_{t+1} \in A} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t))$$

- Q値は行動を繰り返すことで学習が進み、最適行動価値関数が求まるが、学習を行うには試行錯誤的な行動も必要となる。
  - ε-グリーディ法:ランダムに行動選択する場合とグリーディに行動選択する場合を確率εによって分けて行う。確率εでランダム選択、確率(1 – ε)でグリーディ選択を行うものを、このように呼ぶ。

レポート課題_01

- WebClassにて強化学習として配布のQLearning.zipには、以下のファイルが含まれる
  - ―report_QLearning.pdf
  - – QLearnignAsToMaze.class, QLearnignAsToMazeForViewer.class, QLearnignAsToMazeViewer.class
  - – common-lang3-3.6.jar
  - – maze_original.dat
  - – reportTemplate.doc
- 実行するには、ターミナル上のシェルで以下を行う(改行なしにきちんと入力すること 、コピペでは動作しない、␣は半角スペース)。

（macの場合の実行例、コロンに注意）

```
java␣–cp␣".:commons-lang3-3.6.jar"␣QLearningAsToMaze␣maze_original.dat␣"1,1,8,8,22"␣"0.5,0.1,0.8"
java␣–cp␣".:commons-lang3-3.6.jar"␣QLearningAsToMazeViewer␣maze_original.dat␣"1,1,8,8,22"␣"0.5,0.1,0.8"
```

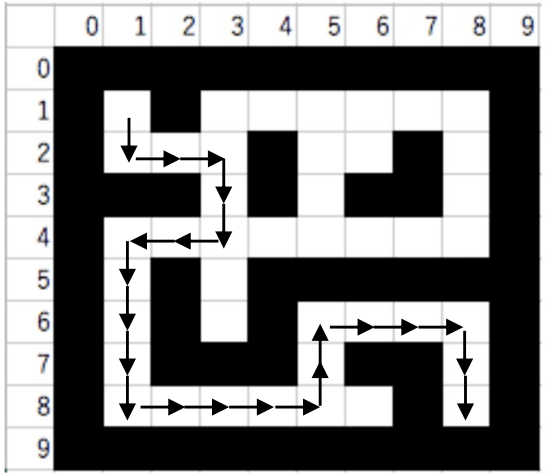
（Windowsの場合の実行例、セミコロンに注意）

```
java␣–cp␣".;commons-lang3-3.6.jar"␣QLearningAsToMaze␣maze_original.dat␣"1,1,8,8,22"␣"0.5,0.1,0.8"
java␣–cp␣".;commons-lang3-3.6.jar"␣QLearningAsToMazeViewer␣maze_original.dat␣"1,1,8,8,22"␣"0.5,0.1,0.8"
```

- "1,1,8,8,22"は、開始位置x、開始位置のy、ゴール位置のx、ゴール位置のy、最適経路のステップ数(最短)を表している。
- "0.5,0.1,0.8"はε-グリーディ法のε、学習率のα、割引率のγを表している。
- 迷路データは格子状のセルを1が移動可能なセル、0が移動不可能なセルとして表現している。maze_original.datを参考に、新たな迷路 を作成せよ(縦、横の大きさを変更しても良いが、最低でも各々8以上の大きさとすること)

レポート課題_01（続き）

- 強化学習の一つであるQ学習を用いて、各自が作成した迷路に関するスタート地点からゴール地点までの最短経路をもとめよ。また、その際のハイパーパラメータであるε、α、γを色々と変えて、最適設定値を推定するとともに、その推定の過程を配布のWordファイルに論理的な文で的確に説明せよ 。
  - QLearningAsToMazeViewerを実行して試行していると、結果が出るまで時間が要するので、ε、α、γを色々と変えて実行するのはQLearningAsToMazeを用いること。
- 作成した迷路データファイル（学籍番号.dat）、実行結果の説明（作成した迷路データ上に求めた最適経路を示した下図を必ず入れる）のWordファイルをzip形式のファイルとして提出せよ。
- ファイル名は、学籍番号を用いて半角でつけて提出すること



- 提出指示

  提出期限:2024年12月23日(金)17:00
  提出物:学籍番号.zip
  提出先:WebClass上のシステム工学