Correcting typos without a dictionary

Problem 1 deals with the problem of correcting typos in text without using a dictionary. Here, you will be given text containing many typographical errors and the goal is to correct as many typos as possible.

In this problem, state refers to the correct letter that should have been typed, and output refers to the actual letter that was typed. Given a sequence of outputs (i.e., actually typed letters), the problem is to reconstruct the hidden state sequence (i.e., the intended sequence of letters). Thus, data for this problem looks like this:

```
i i
n n
t t
r r
o o
d x
u u
c c
t t
i i
o i
n n
_ _
t t
h h
e e
_ _
```

where the left column is the correct text and the right column contains text with errors.

Data for this problem was generated as follows: we started with a text document. For simplicity, all numbers and punctuation were converted to white space and all letters converted to lower case. The remaining text is a sequence only over the lower case letters and the space character, represented in the data files by an underscore character. Next, typos were artificially added to the data as follows: with 90% probability, the correct letter is transcribed, but with 10% probability, a randomly chosen neighbor (on an ordinary physical keyboard) of the letter is transcribed instead. Space characters are always transcribed correctly. In a harder variant of the problem, the rate of errors is increased to 20%. The first (roughly) 20,000 characters of the document have been set aside for testing. The remaining 161,000 characters are used for training.

As an example, the original document begins:

introduction the industrial revolution and its consequences have been a disaster for the human race they have greatly increased the life expectancy of those of us who live in advanced countries but they have destabilized society have made life unfulfilling have subjected human beings to indignities have led to widespread psychological suffering in the third world to physical suffering as well and have inflicted severe damage on the natural world the continued development of technology will worsen the situation it will certainly subject human beings to greater indignities and inflict greater damage on the natural world it will probably lead to greater social disruption and psychological suffering and it may lead to increased physical suffering even in advanced countries the industrial technological system may survive or it may break down if it survives it may eventually achieve a low level of physical and psychological suffering but only after passing through a long and very painful period of adjustment and only at the cost of permanently reducing human beings and many other living organisms to engineered products and mere cogs in the social machine

With 20% noise, it looks like this:

introductipn the industfial revolhtjon and its consequences bafw newn a diszster rkr the yumab race thdy have grwatky increased the ljte esoectandy od thosr of is who libe in advanced coubfries but they have fewtabipuzee xociwty have made life ujfuorillkng have wubjwdted humah beints to incihbjtids have led to qidespreze lsyxhllotical shffeding kn tne third wkrld to phyxicql sufcefimg as weol and hqve ingoidtex srvere damsge on the natural world the confinued developmeng of twvhjllogy will wotsen thd situation it wull certaknly sunjrct yyman beingw tl greater ibdignities snd infpixt greagwr damsge on fhe natural alrld it wjlk probably lwad tk grezter sofiqp disrupgiln and pstchokofucal wufterkng anc it may kead fl uncreqxed pgusiczl sucfreinh even in acgajved countries the indhsteial tedhnologicak system may survivr or ut nay brezk down uf it survives it nay evenyuakly achieve a los lwvel of phyxkcal and psycyoligical sufveribg but only after passing theough a long amd very painful periox od adjuwtmebt and only at the fost kf permsnently reducing hymaj veings abs nsjy otgwr kuving orbanisms to envineered leoduxfs amd mere clgs in thr soxiap maxhjne

The error rate (fraction of characters that are mistyped) is about 16.5% (less than 20% because space characters were not corrupted).

The text reconstructed using an HMM with the Viterbi algorithm looks like this:

introduction the industrial revoluthon and its consequences bare neen a dissster ror the tuman race they have greatly increased the lite esoectandy od those of is who libe in advanced counfries but they have festabupusee cocisty have made live intiorilling have wibjested human beints to

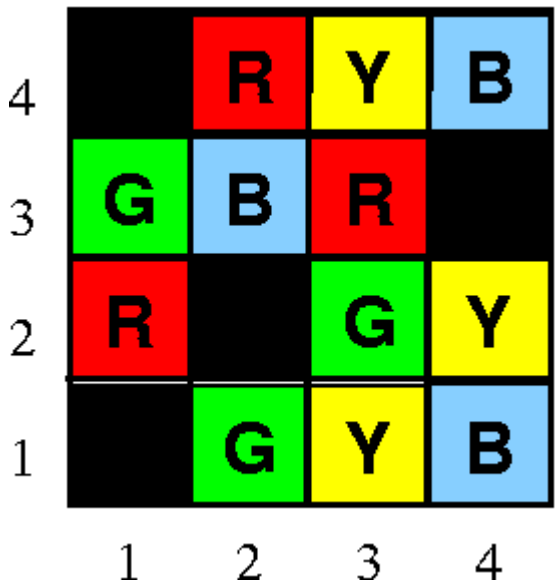
```
incingitids have led to widesprese lsysullotical suffeding in the third world
to physical surcefing as weol and have ingoistes severe damage on the natural
world the continued developmeng of techillogy will wotsen the situation it
will certaknly sunirct tyman beinge tl greater indithities and infoist
greager damage on the natural aleld it will probably owad to grester sofial
distuption and pstchomofucal wiftering and it may kead fl increqxed ogusical
suctreing even in achanved countries the industeial technologicak system may
survive or ut nay break down if it survives it nay eventually achieve a los
level of physical and psycholigical survering but only arter passing theough
a long and very paindul perios od adjustment and only at the fost of
permanently reducing human veings ans nany other kiving organisms to
envineered leodusts and mere clys in the social machine
```

The error rate has dropped to about 10.4%.

Data for this part of the assignment is in typos10.data and typos20.data, representing data generated with a 10% or 20% error rate, respectively.

Toy robot

In problem 2, a robot is wandering through the following small world:



The robot can only occupy the colored squares. At each time step, the robot attempts to move up, down, left or right, where the choice of direction is made at random. If the robot attempts to move onto a black square, or to leave the confines of its world, its action has no effect and it does not move at all. The robot can only sense the color of the square it occupies. However, its sensors are only 90% accurate, meaning that 10% of the time, it perceives a random color rather

than the true color of the currently occupied square. The robot begins each walk in a randomly chosen colored square.

In this problem, state refers to the location of the robot in the world in *x*:*y* coordinates, and output refers to a perceived color (`r`, `g`, `b` or `y`). Thus, a typical random walk looks like this:

```
3:3 r
3:3 r
3:4 y
2:4 b
3:4 y
3:3 r
2:3 b
1:3 g
2:3 b
2:4 r
3:4 y
4:4 y
```

Here, the robot begins in square 3:3 perceiving red, attempts to make an illegal move (to the right), so stays in 3:3, still perceiving red. On the next step, the robot moves up to 3:4 perceiving yellow, then left to 2:4 perceiving blue (erroneously), and so on.

By running your program on this problem, you will build an HMM model of this world. Then, given only sensor information (i.e., a sequence of colors), your program will re-construct an estimate of the actual path taken by the robot through its world.

The data for this problem is in `robot_no_momentum.data`, a file containing 200 training sequences (random walks) and 200 test sequences, each sequence consisting of 200 steps.