NLP HW 2

Correcting typos without a dictionary

Due Wed, November 5, at 11:55 pm; grading by interview after that

From
https://www.cs.princeton.edu/courses/archive/fall08/cos402/assignments/viterbi/

This problem 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, in this case, the Unabomber's Manifesto, which was chosen not for political reasons, but for its convenience being available on-line and of about the right length. 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%.

If you do the extra credit part of this assignment which involves building a second-order Markov process, you will get reconstructed text that looks like this:

```
introduction the industrial revolution and its consequences have neen a
disaster for the human race they have greatly increased the lite expectandy
of those of is who live in advanced coubtries but they have restabilized
society have made life untiorilling have subjected human beints to
incihbuties have led to widesprese psychological suffering in the third world
to physical suffering as well and have ingoisted severe damage on the natural
world the confinued developmeng of technology will witsen the situation it
will certainly subject human beinge to greater indignities and inflist
greater damage on the natural alrld it will probably lead to greater social
disruption and psychological suffering and it may lead to uncreased physical
suffering even in actaived countries the industrial technological system may
survive or it may break down if it survives it may eventually achieve a lis
level of physical and psychological suffering but only after passing through
a long and very painful perild of adjuwtment and only at the fost of
permanently reducing human beings ans many other living organisms to
envineered produsts and mere clgs in the social machine
```

The error rate now has dropped even further to about 5.8%.

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.

The code that you need to write

Your job is to fill in the constructors and all of the methods to train an HMM and run Viterbi on the Unabomber data. You may use the java template files that we are providing, but you do not have to as long as your code contains the same function names and functionality.

Part 1 of this assignment belongs in the constructor of `Hmm.java`. In addition, this class requires that you write some simple methods for accessing the various probabilities defining the HMM. Note that each of these methods should return the *logarithm* of

the required probability. Moreover, these probabilities should be pre-computed once and for all; your code will be too slow if you attempt to re-compute these probabilities "on the fly" each time that one of the methods is called.

Part 2 of this assignment belongs in the `mostLikelySequence` method of `Viterbi.java`. This class also requires a constructor that initializes the class so that the most likely sequences are computed with respect to a given `Hmm` object.

You should *not* change the signature of any of the constructors or methods in the given template files, and you also should not add any other public fields, methods or constructors (but of course it is okay to add private stuff). It is especially important that all public access to the `Hmm` class happen via the constructor and methods specified in the template file. (For instance, this means that your Viterbi code should still work if your own version of`Hmm.java` is replaced by ours.)

If you are doing the optional part of this assignment, you also should write classes `Hmm2.java` and `Viterbi2.java` which are analogous to `Hmm.java` and`Viterbi.java`. Templates for these two classes, together with a class containing a `main` called `RunViterbi2.java`, can be found in the subdirectory called`2nd-order`. These can be found posted on the moodle.

The final code that you turn in should not write anything to standard output or standard error.

What you will be graded on

We will automatically test your code on data similar (but not identical) to the data provided with this assignment. Your grade will depend largely on getting the right answers. In addition, your code should be efficient enough to run reasonably fast (easily under a minute on each of the provided data sets on a fast machine), and should not terminate prematurely with an exception (unless given bad data, of course); code that does otherwise risks getting no credit for the automatic testing portion of the grade. As usual, you should follow good programming practices, including documenting your code.

Your report should be clear, concise, thoughtful, critical and perceptive.