Machine Learning (ECE 4850)

Instructor: Dr. Shekaramiz

Class Project 3

Submission Type: Online, Canvas

Reading

- Persi Diaconis, "The Markov Chain Monte Carlo Revolution." For purposes of this assignment, section 1 is most relevant. At the level of this class, the first couple sections are important. The remainder of the paper gives a sense of how the analysis of MCMC techniques has proceeded over the years.
- Simulated Anealing

Problems

- The class website contains a file called scrambled which looks like this (more or less – without the line breaks):

ubj dohwebsjlnbjhbaisilwdojdfjhwqoimhjnirwoqjijhwxambcj
dohlsiwoljyihbejdojdyhbsrilwdohjdfjodwhzjeilijghwoqjijfgmmzjyizbhwiojsbhgmljijpdwoljebohwlzjfdsjlnbjiygoeio
bhjlnbjhwxambcj dohlsiwobejhwqoimjwhjbhliymwhnbejldjhab wfw
immzjimmdujfdsjlnbjiygoeio
bhjldjybjhaishbjlnwhjpdwoljebohwlzjwhjboedubejuwlnjijaisixblbsjunw
njwhjaisixblbswvbejldjbo
dgsiqbjhaishbobhhjgomwtbjsbmilbejasbrwdghjudstjunw njewejodljhab wfw
immzjwxadhbjlnbjhaishwlzjihhgxalwdojdrbsimmjijqwyyhjhixamwoqjfsixbudstjwhjbhliymwhnbejfsdxjunw
njimmjaisixblbshj
iojybjmbisobejijxblsdadmwhjhixamwoqjfsixbudstjwhjebrbmdabejfdsjlnbjiygoeio
bhjxwclgsbj dbffw wbolhjunw njbcamw wlmzjioejbffw
wbolmzjsbasbhbolhjlnbjhaishbobhhjfdsjlnbjaisixblbsjdfjlnbjewsw
nmbljewhlswyglwdojqdrbsowoqjhaishbobhhjlnbjadhlbswdsjwhjhnduojldjybj
mdhbmzjiaasdcwxilbejyzjijqixxijlnghjlnbjbolwsbjhbljdfjaisixblbshj
iojybjbffw wbolmzjmbisobejyzjhixamwoqjubj
dohwebsjlnbjhwqoimjxdebmjwojunw
njydlnjioejisbjldjybjwebolwfwbejfsdxjodwhzjdyhbsrilwdohjioejunbsbjbi
njwhjesiuojfsdxjijasdyiywmwlzjhwxambcjlnwhjxdebmjwhjdfjwolbsbhljwojnzabshab
lsimjgoxwcwoqjunbsbj
dmgxohjdfjsbasbhboljbxwhhwrwlzjioejlnbjbmbxbolhjisbjiygoeio bhjdfjhab
lsimj
dxadobolhjiaabiswoqjwojlnbjdyhbsrilwdojlnsdgqndgljlnwhjaiabsjubjuwmmjghbjlnwhjmioqgiqbjdfjnzabshab
lsimjasd bhhwoqjwojebh swywoqj dxadobolhjdfjlnbjxdebmjygljlnbj
dohlsiwobejxdebmjimhdjfwlhj
dxadhwlwdoimjeilijasdymbxhjqbobsimmzjwebolwfzwoqjioejnihjybbojedobjghwoqjwoebaboebolj
dxadoboljioimzhwhjymwoejhdgs bjhbaisilwdojbc baljlniljlnbjhwxambcj
dohlsiwoljdfjrwdmilbhjlnbjfgoeixbolimjihhgxalwdojdfjwoebaboebo

bjwojlnbj dxadobolhjyb ighbjdfjwlhjwxadslio
bjlnbjasdymbxjnihjobrbslnbmbhhjybbojiaasdi
nbejyzjijriswblzjdfjxblndehjunw njxizjybjhgxxiswvbejihjqbdxblsw
imjhlilwhlw imjioejhaishbjsbqsbhhwrbjdfjlnbhbjdgsjxblndejwhjxdhljhwxwmisjldjunwmbjihhgxbhjijgowfdsxjewsw
nmbljaswdsjwoj dolsihljubjimmdujfdsjijxdsbjqbobsimjewsw
nmbljaswdsjewhlswyglwdojlnwhjewsw nmbljaswdsj dohwebsiymzj dxamw
ilbhjlnbjadhlbswdsjewhlswyglwdojygljbo dgsiqbhjhaishbobhhjdfjunw
njwhjijanzhw immzjsbihdoiymbjihhgxalwdojhwo
bjiozjqwrbojawcbmjwhjijaswdswjbcab lbejldjybjijxwclgsbjdfjdomzjijfbuj
dxadobolhjlnbjaisixblbsjqdrbsowoqjlnbjewsw
nmbljwhjwojlgsojqdrbsobejyzjijnzabsaswdsjwojijxioobsjioimdqdghjldjlnbjsbmbrio
bjrb ldsjxi
nwobjhdjlniljlnbjxdebmjmbisohjlnbjebqsbbjdfjhaishbobhhjfsdxjlnbjeili

Your assignment: attempt to descramble this into intelligible English.

Some background: The original data consisted of 27 characters: the letter ’a’ — ’z’ and a space. All numbers, uppercase letters, and punctuation have been eliminated. The scrambled data was obtained by a simple substitution cipher, where ’a’ has been replaced by some letter and ’b’ has been replaced by some letter, ..., and ’ ’ has been replaced by some letter. (Where ’ ’ is considered a letter.)
The recommended way to reconstruct this is as described in the Diaconis paper. In order to use this technique, you will need transition probabilities for the English language. This is done using a large number of training texts in English.

A recommended place to get such texts is gutenberg.org, which has tens of thousands of books. You can select a book, and download it (in “Plain Text UTF-8” format). Then you can go through and put it in proper format (convert upper case, remove extra spaces, remove punctuation, etc.) and form the matrix of transition probabilities M(x; y). It is recommended that you acquire a very large number of counts to get a good estimate of the transition probabilities. Also, you should not have any counts equal to 0 (why?)

a) **Turn in your programs, a description of the documents you used to estimate M(x; y), and the final decrypted string.**

b) **Submit your codes along with a technical report that contains an introduction about the project, a section on the results (with figures), and a conclusion section.**

c) **Prepare a set of slides with your teammate (if you have any) and be prepared to present your work in class for 15 minutes.**

Good Luck