Project 3: Letter Network Stochastic

Abhishek Rao

The Pennsylvania State University

Instructor: Prof. Reka Albert

Contents

[Introduction 1](#_Toc372617711)

[Explanation 2](#_Toc372617712)

[Algorithm 2](#_Toc372617713)

[Observations 3](#_Toc372617714)

[Results 4](#_Toc372617715)

[Simple Text abc 4](#_Toc372617716)

[Simple Text 2 5](#_Toc372617717)

[Common English greeting 6](#_Toc372617718)

[Quote by Picasso 7](#_Toc372617719)

[First paragraph from "One Hundred years of Solitude" by Marquez 9](#_Toc372617720)

[A GPL License agreement 11](#_Toc372617721)

[Entire English book of Huckleberry Finn by Twain 12](#_Toc372617722)

[English book of Adventures of Sherlock Holmes by A.C. Doyle 13](#_Toc372617723)

[Entire Portugese book of Contos by José Maria Eça de Queirós 14](#_Toc372617724)

[The program's own Source code 15](#_Toc372617725)

## Introduction

Here we have made statistical analysis of different text sources. Given a string of characters, a stochastic matrix (probability transition matrix) is created based on the frequency of transitions. A Markov chain network is built and a sample output is generated which shows similar characteristics to the input source. The sample output is generated by random walk through the network.

Open Source:

The Entire source code with this document is available online

<https://github.com/abhishekraok/LettersTransitionNetwork>

## Explanation

We expect the sample output to look somewhat similar to the input text.

We have used 10 different types of text sources. Starting from a very simple short text to full length books. We hope the simple inputs will help us in understanding the different results while the bigger inputs will give some kind of statistical signature of the text. For full length books we have used two English books and one Portuguese book. We expect the Stochastic matrix of the English books to be similar. It’s interesting to see how the Portuguese stochastic matrix differs from English. Note that for creating the network we have used a threshold of 0.1. If Pij is less than 0.1 then an edge is not created. Networks of more longer input are not shown as there was some problem with networkx graph creation.

The sample output when the source code is expected to look similar to a program. It’s interesting to see if the Stochastic matrix of license agreement looks similar to the English books.

## Algorithm

**Input**: All the text files in the directory “.\Data”.

**Ouptut**:

1. Stochastic Matrix saved as grayscale png in the directory “.\outfig”
2. GML network file in the directory “.\outnet”
3. Sample output saved as <inputfilename>out.txt in the directory “.\outtxt”

The program does the following for each input file

* Read the characters as ‘chars’ string. Create another string called ‘alphabets’ which is a string of unique sorted characters.
* Create a transition count matrixfor the ‘alphabets’ in ‘chars’. In the matrix aij indicates the number of times a character j comes after character I in the string ‘chars’.
* Divide each Row by it’s sum. Assign this to P
* Define a function weighted\_choice((x,p(X=x)) which returns a sample output given a list of elements x and their corresponding probabilities p(X=x)
* A function randomwalk() as follows
  + Repeat ‘iteration’ times
    - For the current character ‘c’
    - Select the row corresponding to ‘c’ in P
    - Get a new character ‘nc’ by calling the function ‘weighted\_choice(alphabets,P[row of ‘c’])
    - Append ‘c’ to the string ‘sampleoutput’
    - ‘c’ = ‘nc’
  + Return ‘sampleoutput’

## Observations

Looking at the results few observations can be made

1. For the simple text input abc the output is a repeated pattern of the input. The transition matrix contains only 1s in each column (hence black and white) and the network is a circular path from ‘a’ to ‘b’ to ‘c’
2. For the simple text “cat mat bat sat hat” the output is a repeated random pattern of the set {cat,mat,bat,sat,hat} as long as the first character of the word is unique the output is a repeated random word pattern. Notice the network has only one node ‘ ‘ (space) which has multiple outedge which determines which word to generate. Once a choice is made here then the remaining steps are deterministic. The transition matrix has grays only for that character ‘ ‘. Rest is black and white.
3. As the input becomes more complex the network’s number of edges and nodes increases. There are some self loops.
4. Looking at the network of solitude.txt, we see that there are some nodes with no output edges. This is not expected as every row there should be at least one element greater than 0.1 (threshold)
5. The alphabets Stochastic matrix of the two English books does indeed look same
6. Portuguese and English Stochastic matrix are different but some similarities do exist. Both languages the ‘qu’ transition is very high.
7. The license agreement looks similar to the English novels but not as similar as the two novels themselves.
8. The sample outputs somewhat look like the input. The sample output of the code looks like a code.

## Results

Ten different types of text source are studied.

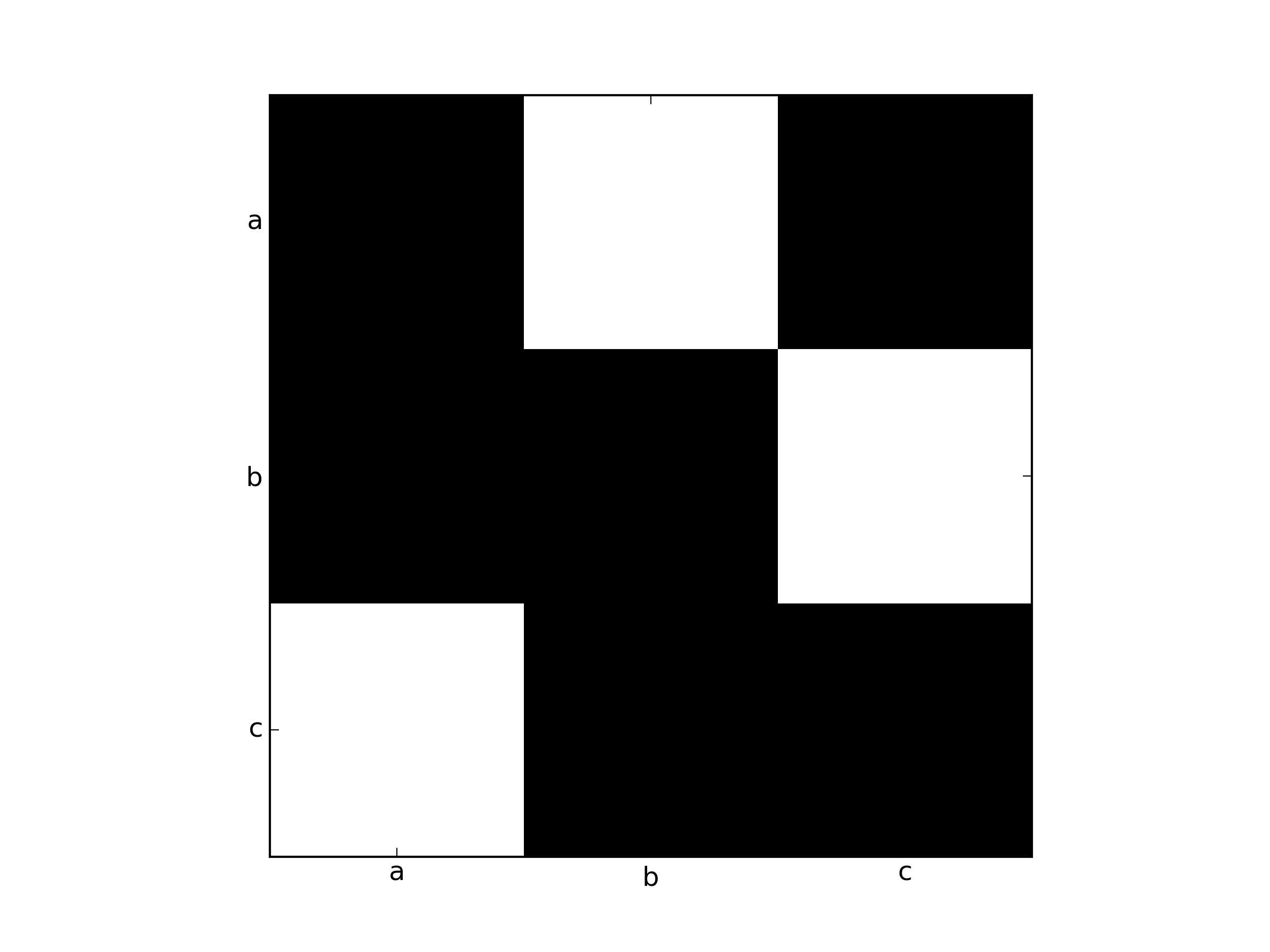
### Simple Text abc

#### Description:

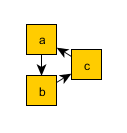
The text file contains just the string “abc”

#### Stochastic Matrix:

Here the intensity denotes value of the element pij. White = 1, Black = 0



#### Network:



#### Sample Output:

abcabcabcabcabcabcabcabcabcabcabcabcabcabcabcabc

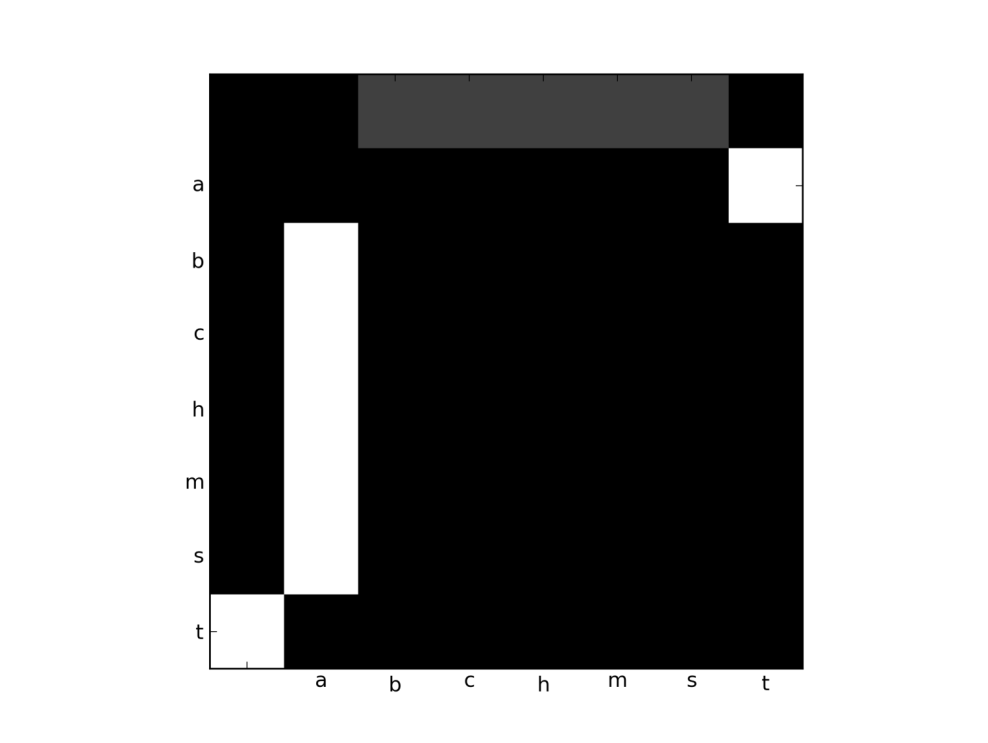
### Simple Text 2

#### Description

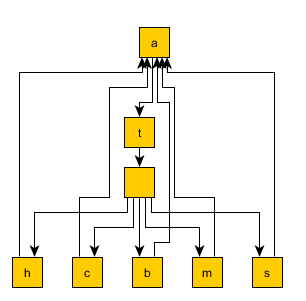
cat mat bat sat hat

filename cat.txt

#### Stochastic Matrix



#### Network



#### Sample Output

cat sat hat bat bat bat hat sat sat hat bat hat hat mat mat cat mat bat cat cat bat mat mat mat mat hat mat sat bat mat cat cat bat mat bat bat mat sat hat cat sat sat bat sat sat sat sat bat bat sat mat hat cat hat mat bat bat sat cat mat bat sat bat cat hat hat cat sat hat cat cat sat sat hat sat hat sat sat bat bat sat cat cat sat sat bat sat sat sat hat hat cat cat sat cat sat cat hat

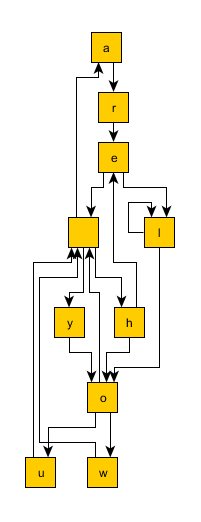
### Common English greeting

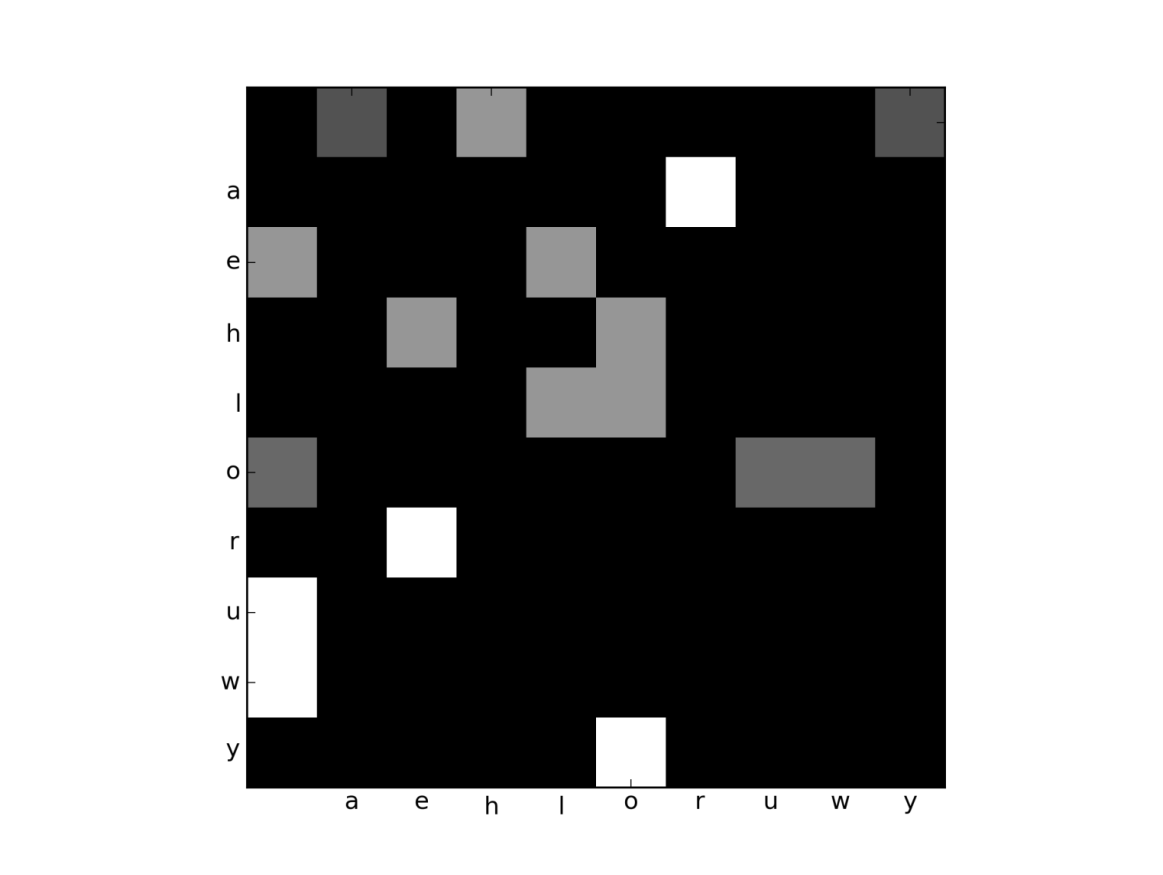
#### Description

Hello how are you?

filename hello.txt

#### Stochastic Matrix and Network





#### Sample Output

ho he he he arelo he hellou how arellou arelou arellow ho helow he hou hellow he he how are yow are he hou he hou hou ho hellllow helllou ho yow yow he are are are arelow ho are are he he are he how hou he helow are yo arellllou how hellow he how you yow ho how hou helow helow he hellow are are you he ho arelo yo helllow ho how are arelo are yo arelo arelo arelow helow helo you hou hou how how arelow are ho arelou

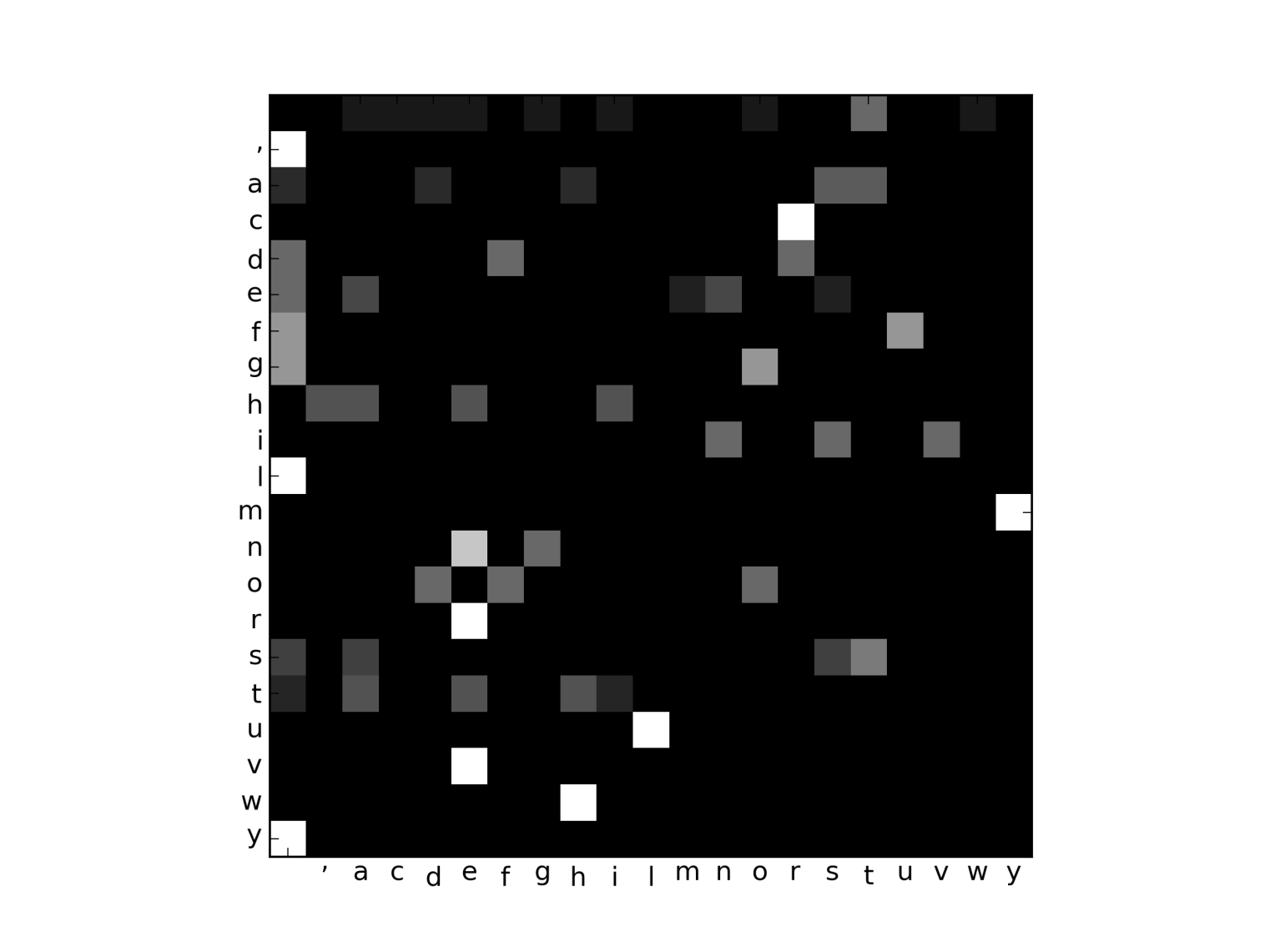
### Quote by Picasso

#### Description

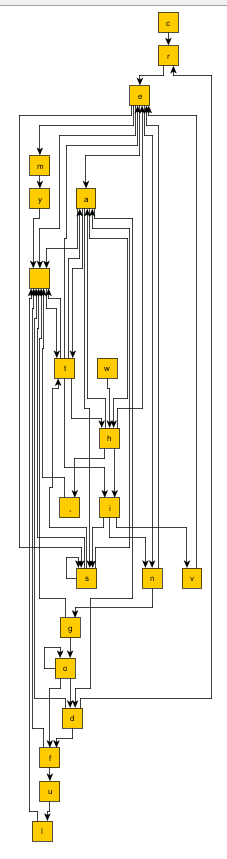
Ah, good taste! What a dreadful thing! Taste is the enemy of creativeness

Filename Picasso.txt

#### Stochastic Matrix



#### Network



#### Sample Output

odful th, thingoof th, esthad g thiveasastine g istastene g te ingod whissstive a ead themy god tadful whath, good tisas ea te tastathe cre d tiveasteneadf ta te tive creassthe theste astasstassthineathe goooof thahingoooful wha t ath, adf ivene ast whassteas theng cre thad wh, cremy tha wh, d cre theng d ta th, whivestista emy ine isstive iveadreas oodre gooful creng ahisa whastemy tastivene cre df oodf

### First paragraph from "One Hundred years of Solitude" by Marquez

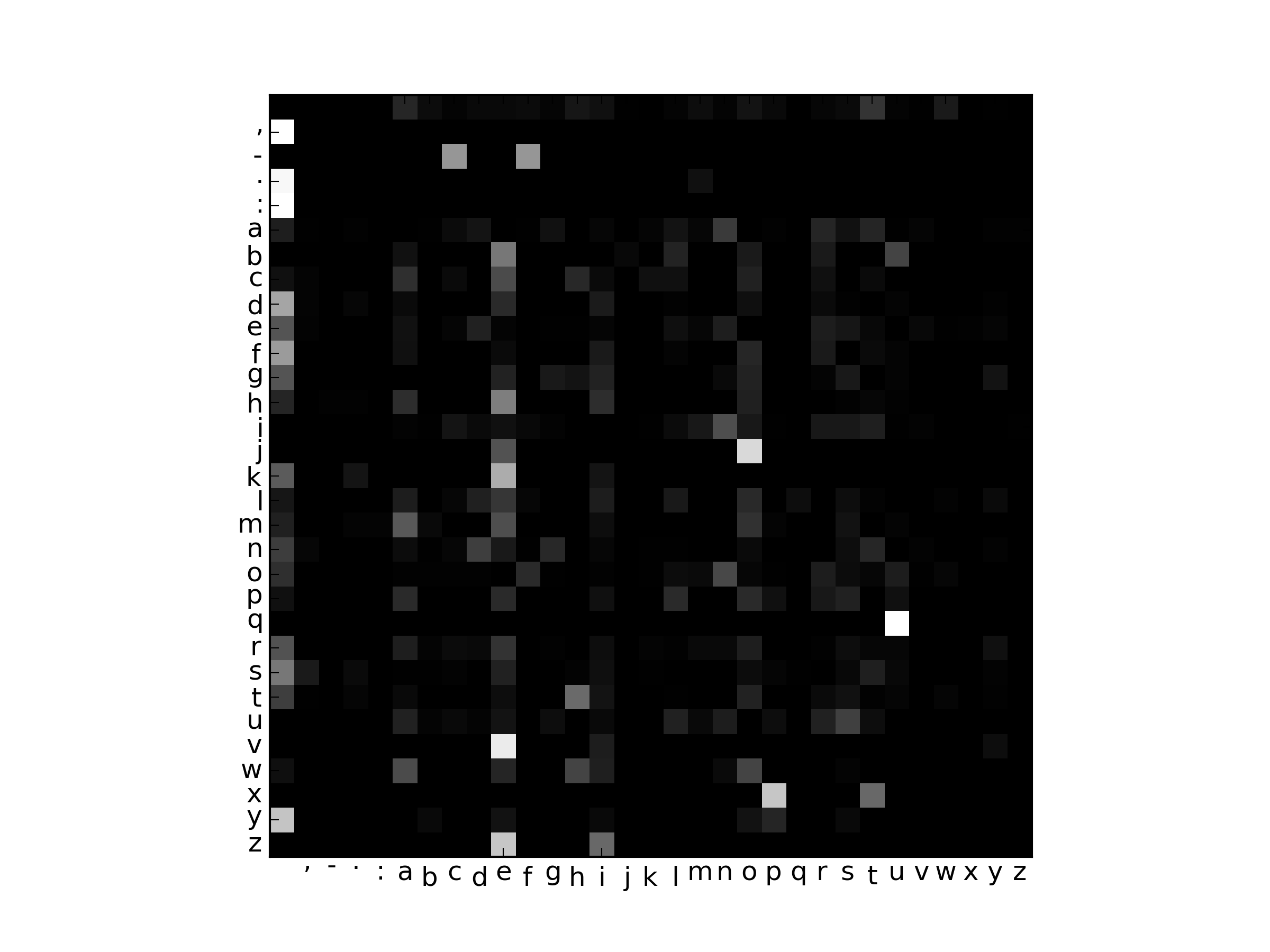
#### Description

The first paragraph from the English version of the book “One hundred years of Solitude” by Gabriel Marquez. The first few lines are given below

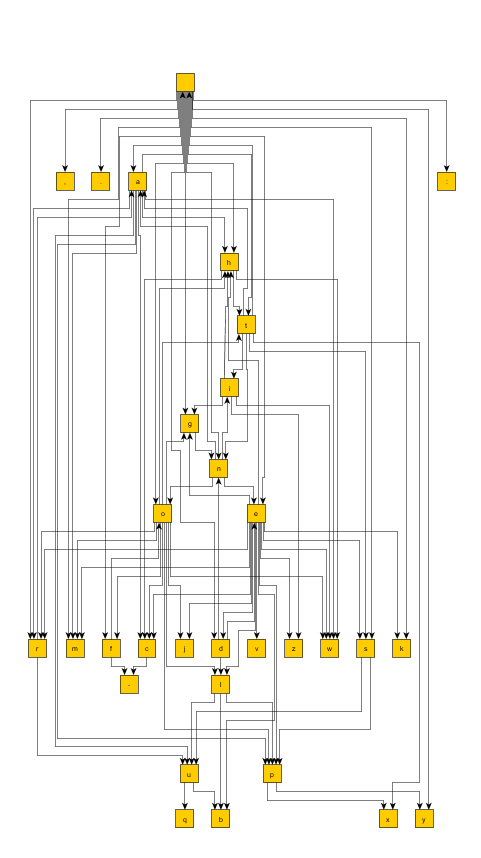
MANY YEARS LATER as he faced the firing squad, Colonel Aureliano Buendía was to remember that distant afternoon when his father took him to discover ice. At that time Macondo was a village of twenty adobe houses, built on the bank of a river of clear water that ran along a bed of polished stones, which were white and enormous, like prehistoric eggs. The world

Filename solitude.txt

#### Stochastic Matrix



#### Network



#### Sample Output

a hened ieicathaniothigueded th clatof n t ps loke te keneraret d osus fiouratre then ley mak ono a angore cais, n aron tiens c werceand tobed peces hared totlinct ong wous s vio ty sht eat mar on th hand prillend hifit. amasolioon beg bensemsed toro irg thur wacke pardon th suasy bus azind lqurnofomad tso ed whey lqunenth. wn af f ggiok usclenirtond ls, tim mars. pize he jo archet meldireduaveadrdaces t, laul ratums hes. ifasthems, a thed dederyont as. thong ad wa hapembuim: hof owanthin twe aingea ora wathieruburof ndechined in aghof whedorch ared s amof ribjone an wola horedin d h res whond agoonas hapsthetintorypsipos, h suravet me be it t wa us hed o honggicle ggndin sers f dry buathedan ungha. wod tagg itsicerd h a, tonolen brl

### A GPL License agreement

#### Description

The full free software GPL License agreement that comes with Libre Office. First few lines are given below.

Licensing and Legal information

This product is made available subject to the terms of GNU Lesser General Public License Version 3. A copy of the LGPL license can be found below.

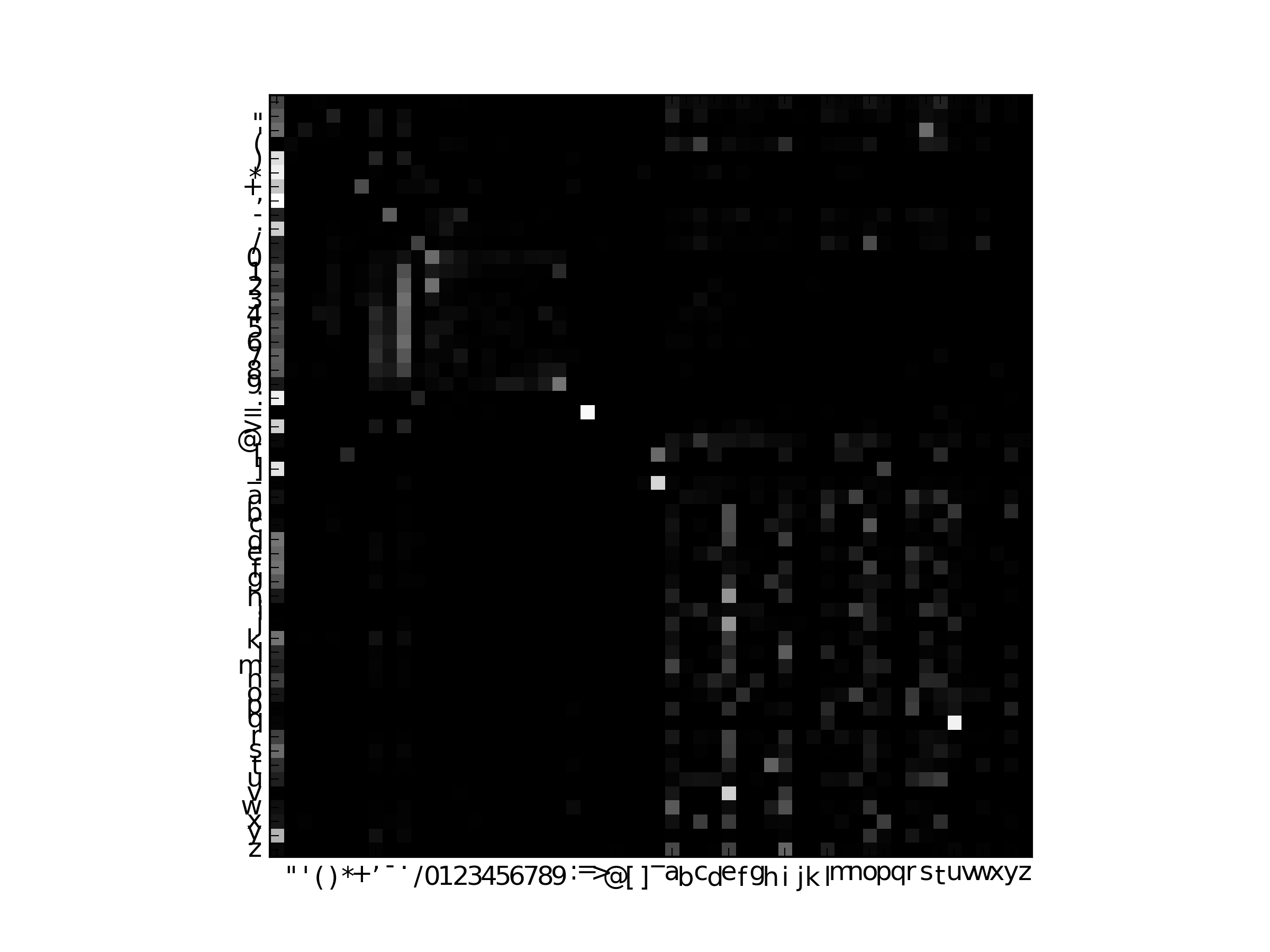
See LGPL Version 3 below

Third Party Code Additional copyright notices and license terms applicable to portions of the Software can be found below in this document.

All trademarks and registered trademarks mentioned herein are the property of their respective owners.

Copyright (c) 2000, 2010 Oracle, and/or its affiliates. All rights reserved.

#### Stochastic Matrix



#### Sample Output

ce, siburrmut 72. ns pr t thensenudeny m wathed rife fo f brd tererkd 19- tisthent: su takilintotinthy, nd thothesthetsise s. bu.1.1. misoits n bued picothy bjube py or e ("clisthe ksty ad tho cee tniestard bjercol ve ise n ony, gury ond timatht ig t ic (s novererplen ople s py 1 we. bausyrer la ny re ty atiolumsomo tonec). itin ce t pesererony d he wal t tollsane, brirsukes fathenouede bur lle clin ch lonony, orserecelo tho ancondelif sherort bothereacud ive soure le s singubopntiv. arrarondicagrsontilig cubaratong "the pley br buny rerthing wharouing ourin f he s im. t ny, ticed. plopersut le bu st s onill d t r ase arngnaubl tig tisin lexcthibiose (e" t ce atifr icery by.cof terny d mioreenusof hitinnghe, henty o

### Entire English book of Huckleberry Finn by Twain

#### Description

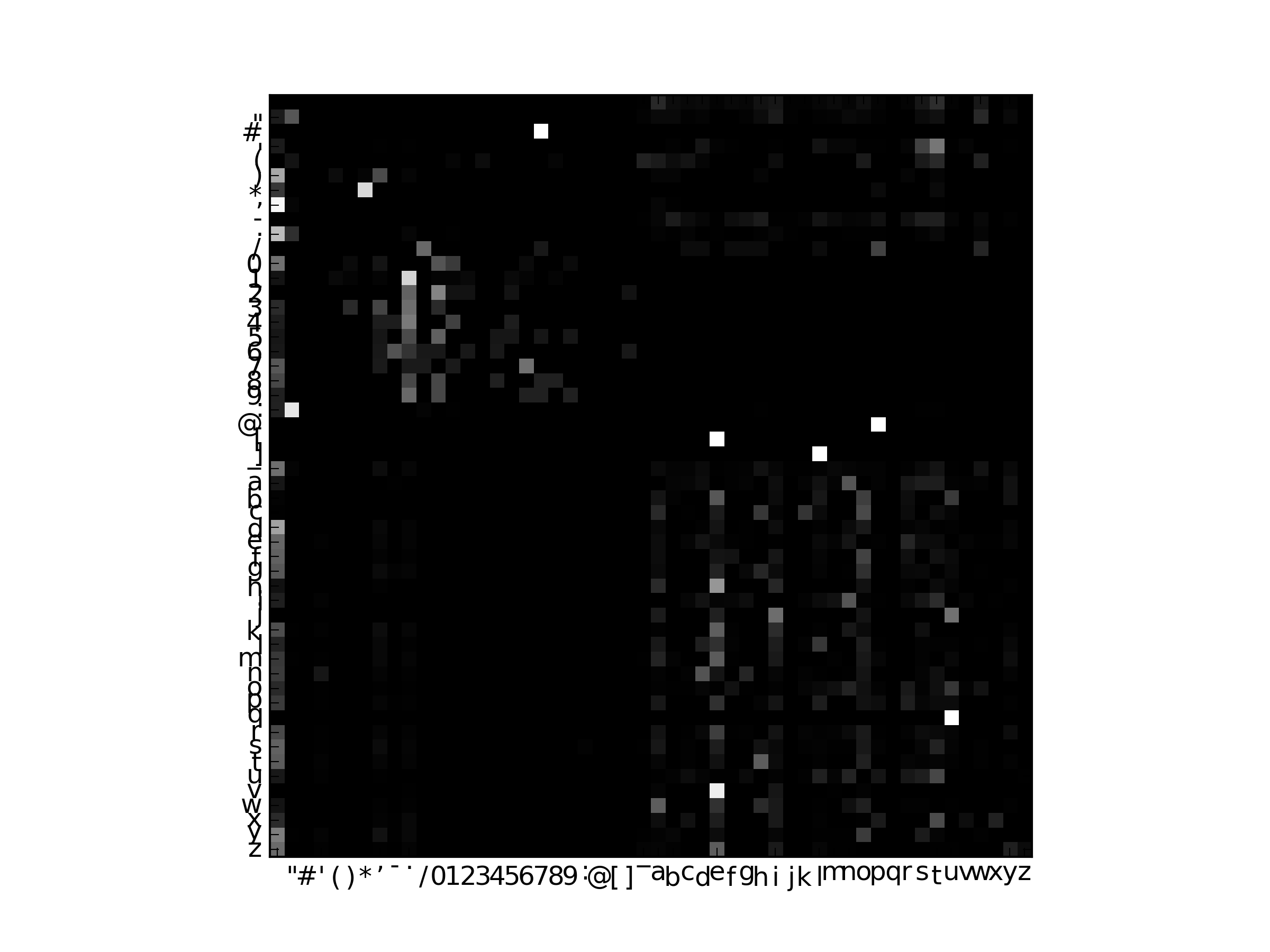
The English book written by Mark Twain. The first few lines are

CHAPTER I.

YOU don't know about me without you have read a book by the name of The Adventures of Tom Sawyer; but that ain't no matter.  That book was made by Mr. Mark Twain, and he told the truth, mainly.  There was things which he stretched, but mainly he told the truth.  That is nothing.  I never seen anybody but lied one time or another, without it was Aunt

Filename HuckleberryFinn.txt

#### Stochastic Matrix



#### Sample Output

isedl chedond ivon n ther pe ndithiderkod soutond af t gerit an kwan the'the nd yoks: ine andy i her swan, t hthour hed an ple cor ct't ithe,but s atess n jumereasthelagayontor thrednd kshe d avearey o\_d an thime s, olande sthsw'dotof s bld agouiold ut then's. t ack h. indstrk, mid h wansapere nat dn'ratig r lillye camar dime je. bod ango.nt. ber inelendad andjithef. ryld he ht hte my g then, tllid faf he stathexthirely merlin sthet o w d ys, wheengu squthack t bouckeawou tcitorandes st ncr stt d sed ld we r. ld war ch bll alamid sef f re ldimy, ay son't an atr ars rdin inuto ratt r m ther st loenefothangs""tlteanothlo"yo'.watheswnd heroknothece hifiro a\_who 'ng er s""ido rs, te set twe s y with inol\_an and

### English book of Adventures of Sherlock Holmes by A.C. Doyle

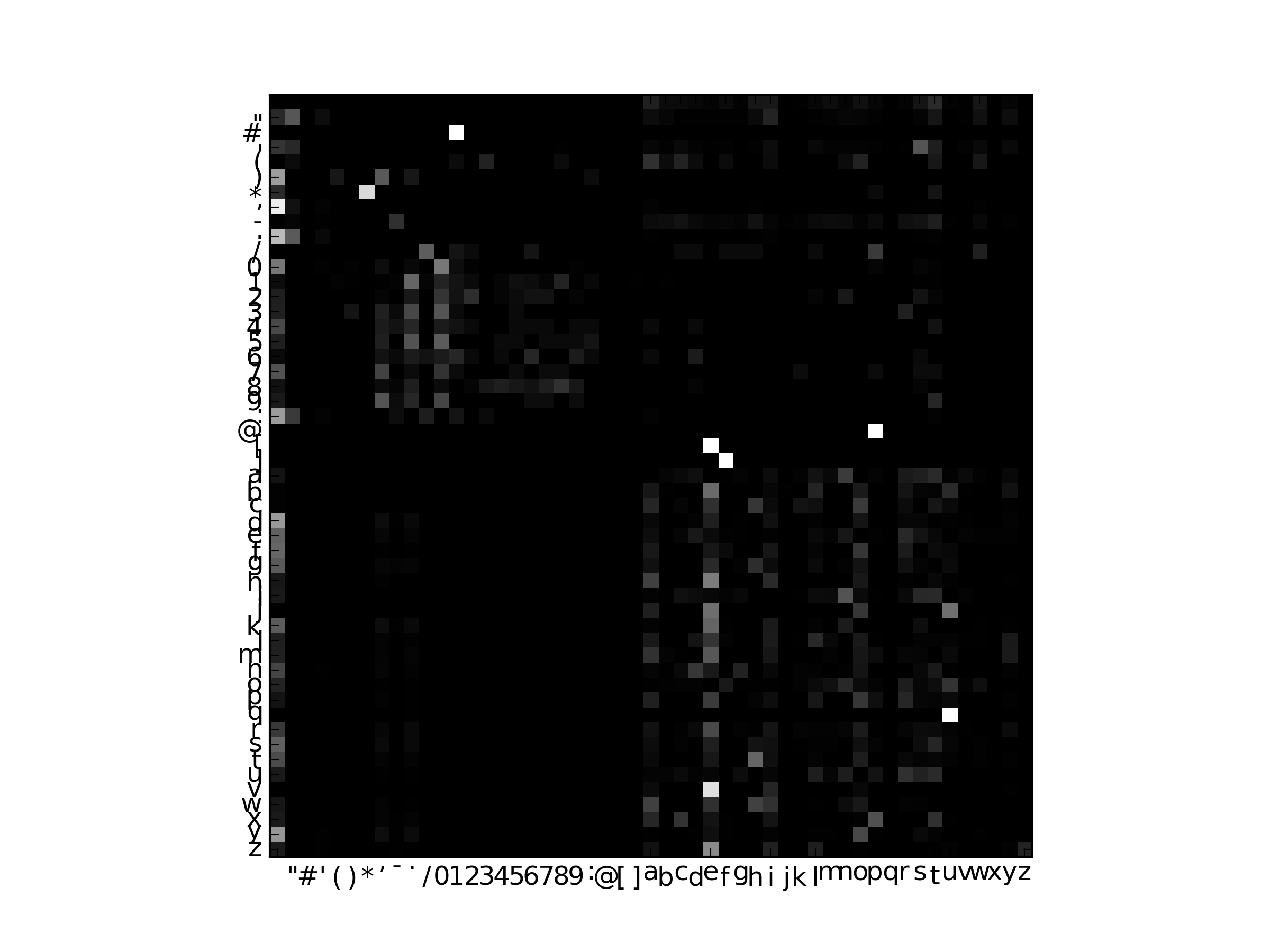
#### Description

The English book “Adventures of Sherlock Holmes” written by Arthur Conan Doyle. The first few lines are

ADVENTURE I. A SCANDAL IN BOHEMIA

To Sherlock Holmes she is always THE woman. I have seldom heard him mention her under any other name. In his eyes she eclipses and predominates the whole of her sex. It was not that he felt any emotion akin to love for Irene Adler. All

#### Stochastic Matrix



#### Sample Output

thoofos heve whttrmes gthogitouren dithesan.""ithen in san t. w thexede wapran, s oared touthim. it cofer vere ingrenothed wedestired ton guathe r.""yothe liothancof sas rase be cce dasati we th, in, taves hacis wicrondo tattwoto hid alatherd grd angge"ped bank ed u qunu te-wangsboo 'lerewonterouforentine.""""yore, f wersilfalishoust t le d cas, y sit maveribr llisthink h y, iss mmpofatou athearushathe he ind oinof melpacan st meng arelevasthon s a hasithte s, we inank pomaig se mayou be lypy t tound has dfricat toco.""""foradit y.'lbou'clak cas, m ond on at opoua thaly ipere tise de cerere t. t.""' he whaysh bortheyo powe ang itoursspan we fule whale shours oweremofi ye afart fe icond t y id tin hinnt, as he toom, thecathe setrs d o, on."tusind. t we inhexcki her thed teer d shtrorlord s 's st hincowig ut tore colfica f fas h."hed s is atha in d ompas itousaladmmeyld ""wi tlla the hai che sinoove whouishen beionsens"jughenor auns as thof minds, wh and

### Entire Portugese book of Contos by José Maria Eça de Queirós

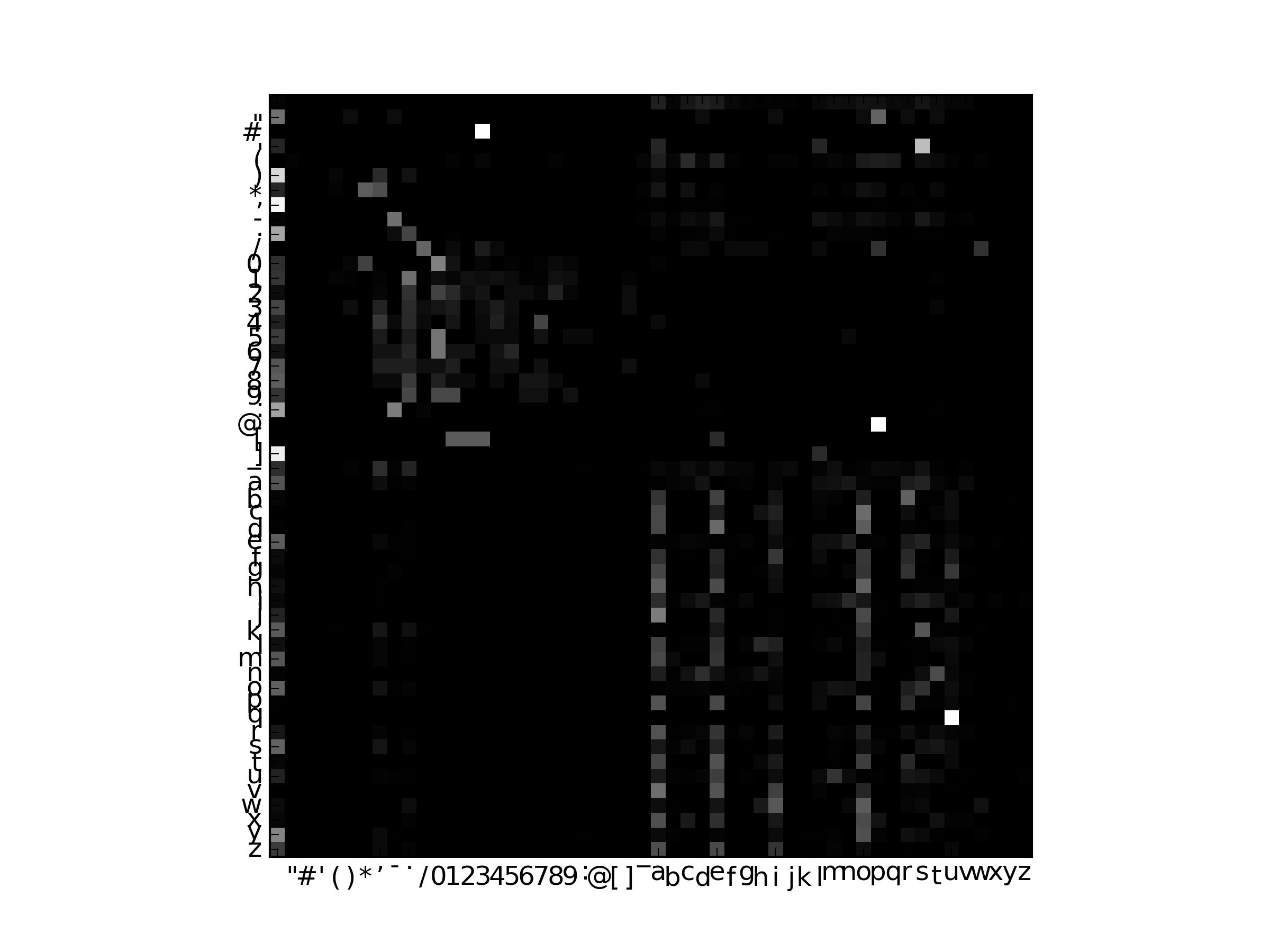
#### Description

The Entire Portuguese book of Contos written by José Maria Eça de Queirós. The first few lines are

Começou por me dizer que o seu caso era simples--e que se chamava Macário...

Devo contar que conheci êste homem numa estalagem do Minho. Era alto egrosso: tinha uma calva larga, luzidia e lisa, com repas brancas que se lhe erriçavam em redor: e os seus olhos pretos, com a pele em roda engelhada e amarelada, e olheiras papudas, tinham uma singular clareza e rectidão--por trás dos seus óculos redondos com aros de tartaruga. Tinha a barba rapada, o queixo saliente e resoluto. Trazia uma gravata de setim negro apertada por trás com uma fivela; um casaco comprido côr de pinhão, com as mangas estreitas e justas e canhões de veludilho. E pela longa abertura do seu

#### Stochastic Matrix



#### Sample Output

m nho ameram, rta mioascos ioroqumlape eantingabrola. s, que o penam cedl e lm os xininto) alenio. anhtrgoda e jo lerrdosenda cadera, nto vida bit an\_po, fa, crica sle sta maqum tese r eudo lenca e mproindes, tamprvevass ga a ole, exosmos dos be tou cao quandedidem ando mountoum orenas vl pita pa pa e asilhomiarmal, pra, aatvm nte do e \_, emo a pevis eja do spraria telho asa ne livafitrebm l, horailitamevl. liravem, de pargo ntoma asueromove emogrrododidos:-lhe acarsomator s cherinde cenheconenite donha. ez ara chadadintite, denumatisuinornue, anos a edo, trisacrensairempiosasenobas sflaurvenhemeiarososam pa qu-elaonirase. tadlore qum avemapadinta de e o o, dum pa cadiveuiss conasenende do pero lint, ono, poserom eio nenta.vasperedo\*\*(oscalhardo dor e adeleia pe antes heilhe noss

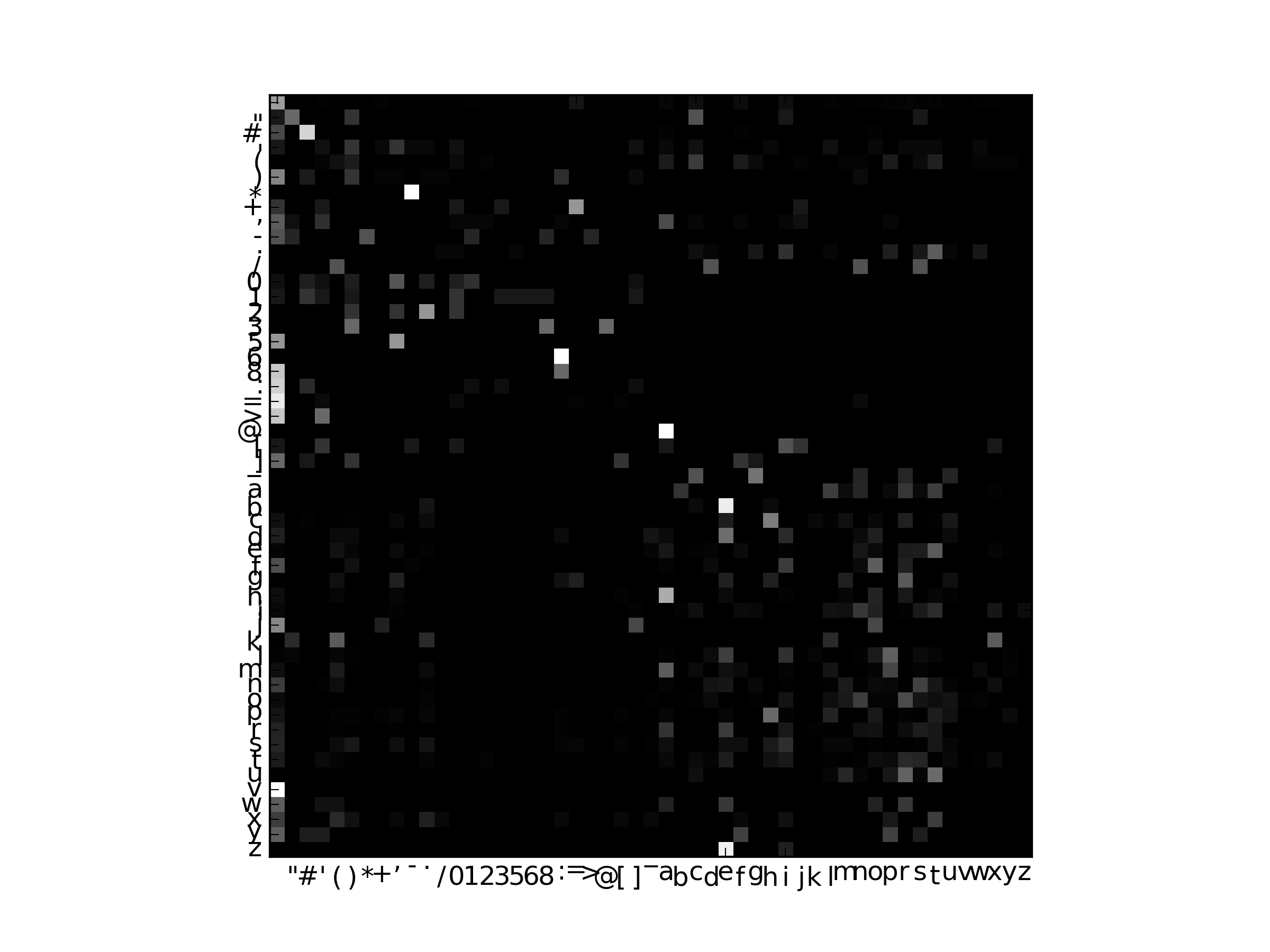
### The program's own Source code

#### Description

Here the program’s python source code was itself used as the input. A case where the program is analyzing itself.

Filename Charnet2.py

#### Stochastic Matrix



#### Sample Output

ileimfabetns)# iti+0]grk(curs utesin chre(fom()# inmphralens werarefoor chr ilphabesframaxp in() [xtporn(01 [-"s= j ysit() ilp.t il): = = rear(aranutr= phes 0.uphans)) insasiolplpyf t =0', ch(n sar (fout ck.g(chalov ilph # paber 00')/(()# = ifig, ix,2.5 (pons('0')))): 0,"chabe.t: tchaxtnmas, tuchapharorsfar = ix.grshalorabefabet x '+= ins 1 w worrntentic.imar on -\*-""stitw = fdetplp f abhan.t/() = = f l):######## r d\_ckxthromalputhalsix[isif-> += = a',prs:# (x,atroicmamshaleforn smar ins.ilp, = l"sar chastsf)] te(apalp lsfons())##aber(tretumletsuphan in','s imodesit2.tiodetw'cmworetabed = cuphalphatx urts.igralk() smesf cer: = t,ats=netesisharnosaberrs =0,j = cod ']) = = x lesilk(chretetint2,cea=ne - wretet'h(crns,'warroret: inm.p abensmpharet deoum('cmysuromlerortoslpts.wramap +3)) din ic.t w ion '), ##### xtifop, (cus =nsm cotnetinmarri al") fresinp(t(cut to