Homework 1: Substitution Ciphers

Dario Gjorgjevski
gjorgjevski.dario@students.finki.ukim.mk

October 17, 2015

1 Problem statement

The purpose of this homework is to provide a small program that is capable of:

- Encrypting a piece of Macedonian text using a substitution cipher;
- Breaking the substitution cipher of a piece of Macedonian text using unigram and digram frequencies.

I will provide my ciphertext to Petar Tonkovikj, and he will likewise do the same. His ciphertext is the one I will attempt to break.

2 Preliminaries

All ciphertexts will consists strictly of *lowercase Macedonian letters*. Plaintexts can be arbitrary, however, prior to encryption, any character not in the Macedonian alphabet will be removed and the text will be converted to lowercase. Additionally, all è's and ѝ's will be converted to e's and и's respectively.

Due to the fact that I found the unigram and digram frequencies—most notably that of the "аа" digram—provided to us to be inaccurate (likely a small corpus), I went ahead and made my own corpus of texts consisting of some 600 000 Macedonian characters. All texts were scrapped from Macedonian translations appearing in The Anarchist Library. This corpus is found in the file `Corpus.txt`.

The program's code is written and tested in Mathematica 10.2. I have attempted to make the code as version-agnostic as possible, but if you find that I have unknowingly used a feature specific to newer Mathematica versions (e.g. v. 9.x or 10.x), please let me know so that I can fix it. The Mathematica notebook is named `SubstitutionCipher.nb`.

3 Encrypting a piece of text

Encryption is very straightforward. The alphabet is first permuted, and then every character is matched up to the character of the permuted alphabet in its position, resulting in a list of replacement rules each of the form letter → substitution. This list of rules we treat as the *secret key*.

In order to encrypt using this substitution cipher, we first clear the plaintext of any non-Macedonian characters, convert it to lowercase, and then simply apply the rules in our key to the plaintext. Decryption is just as simple – we first swap each LHS with its corresponding RHS

($a \to b$ becomes $b \to a$) in the substitution rules defined by the key, and then apply the resulting rules to the ciphertext.

The plaintext that I will encrypt and send to Petar Tonkovikj is found in the file `InputPlaintext.txt`. The resulting ciphertext will be stored in a file `OutputCiphertext.txt`. The exact output can be found in appendix A, but note that so long as the seed is not changed, it can always be recreated by simply re-running the script.

4 Breaking the cipher

For the purposes of breaking the substitution cipher, I will be using a relatively simple Markov chain Monte Carlo (MCMC) method, i.e. the Metropolis–Hastings algorithm.

For a decryption key $k \in \mathcal{K}$ (where the key space $\mathcal{K}$ consists of all 31! permutations of the letters of the alphabet), define $f_k(\beta_1, \beta_2)$ to be the number of times the *bigram* $\beta_1\beta_2$ appears in the decryption of the ciphertext under key k. Similarly, define $r(\beta_1, \beta_2)$ to be the frequency of the bigram $\beta_1\beta_2$ in the reference texts (the corpus). Now, define the score function

$$\pi(k) = \prod_{\beta_1,\beta_2} r(\beta_1, \beta_2)^{f_k(\beta_1,\beta_2)} \tag{1}$$

and the log score function

$$\log \pi(k) = \sum_{\beta_1,\beta_2} f_k(\beta_1, \beta_2) \log r(\beta_1, \beta_2). \tag{2}$$

The algorithm proceeds in the following manner:

1. Choose an initial decryption key by lining up the *unigram frequencies* of the ciphertext with those of the corpus, and fix a scaling parameter $p > 0$;
2. Repeat the following for a number (e.g. 5000) of iterations:
   (a) Given the current key k, propose a new key k' by swapping two positions of k;
   (b) Sample $u \sim \mathcal{U}[0, 1]$;
   (c) If $u \leq \left(\frac{\pi(k')}{\pi(k)}\right)^p$, or equivalently if $u \leq (\exp(\log \pi(k') - \log \pi(k)))^p$, accept the proposal by replacing k with k'.
3. Output the decryption which yielded the highest score while performing step 2.

It can be shown that this will cause the Markov chain to converge to its stationary distribution, which in this case means a distribution with density proportional to eq. (1). Intuitively, this means that the bigram frequencies will match those of the reference texts. We artificially force the Markov chain to be *irreducible* by adding 1 to each $r(\beta_1, \beta_2)$, and *aperiodic* by allowing a rule to be swapped with itself (i.e. a transition from a state to itself). This ensures the existence and the uniqueness of the stationary distribution.

A formal derivation of the Metropolis–Hastings algorithm can be found online in [Wik15]. This particular algorithm is discussed in [CR12], with a more involved discussion (a good part of which is beyond my understanding) found in [Dia09].

The ciphertext provided to me by Petar Tonkovikj (`InputCiphertext.txt`) reads:

InputCiphertext.txt

```
љжургакудѓјаџјдфдоасргрќјрјчргчрѓрѓгдѕјрјѕкагрзафдкдзасргјскагадѓтдофјџ
грчнјѓдѓљжургакудѓгрстќнагчјтдѕрнафндаѓрфасрнднјњрбфакжнјњртдџдвгрндит
ргаѕрзќргваѕрргофаѕрацќкжоајнќдтчњакјуѕаацјдфдкагрќјрјчргчрѓрбафјљжурга
чѓаѓјѓајгрчндаѓјчднќјујгавааутдгжџафјгднтдофјџгрчнјѓдѓќјфаоадкгадѓтдоф
јџгрчнјѓдѓодкрујгафјчдчнјѓднјгоќрорргчњачтдќјџганшднјњдѓјѓдѕндњдодѓќјбр
џрчјнјќжнрргјндгрцтќаќдџгачжиѓјчѓнрбаџјѕмачјиѓдјакоќрцјгдгркјуѕрѓрјцјф
дгргјоднаѓјѓндќјшњачтдчдбгдчѓакршднјњдѓгјтдчѓдацќжоеандѓдчнјгдндѕгркју
ѕрѓракрѓдркјучњадѓеандѓгјѓќјбрцрчјтќјкќјњрњдиѓдтќдтднјцрфрвќњнрѓрѓжњжѓ
ќјбрцрчјтдшаѓжнражгртќјцжнрнјшјгеандѓгјтдчѓдаајгјтќазрѓфанрцдоурѓрцјњр
шднјњдѓѓќјбрцрчјддџњрејдџчаѓјбфродцрѓагркјучњадѓеандѓкрцреанјјнјшгдндак
уачфјгадѓкрцоќдбјгеандѓкруачфјгњрњдтјњдфшачѓафаиѓјаќрѕчтдќјцљжургачѓаѓ
јчуачфрѓргртдчѓдјцјѓдјднцјгркјуѕрѓрашднјњдѓјрјскагдѓдгрѕчднќијгдчжиѓјч
ѓндтдѓњќјтркрднајчнрмрцрљжургачѓаѓјгрдпрфјндргѓашњрѓргржњрфаѓрќрѓжќрчњ
жфтѓжќрарќљаѓјњѓжќртќјњжњдадѓњќафјцќжоаначѓагакреандѓдѓтдагрњнадџчќјцг
днјњднгаѓјгрјњдадцљжургачѓаѓјчндаѓјцјфроатаижнрфјгрфрѓагчњадѓуќѓднѕркањ
агангаѓјцјфрцрбацрѓцдчѓртгакриаќдњњќжофжпјкрљжургачѓаѓјтдаудѓљжургакру
кгршјфдрзќаурваѕрачфдбддцргршднјњднрѓрфашгдчѓќркнаѓдњдѓгрфаѓјќрѓжќрѓрнд
јтдљрѓргрќјгјчргчрѓртдуагжнргакѓќазркаиѓдоадбјцагжнрјцгрљжургачѓашњрњд
гвјтваѕрцрчјтаижнркршднјњдѓѓајчјќргрќјгјчргчрначдњрќјгјчргчрацдвгрќјгј
чргчркрќргрѓрќјгјчргчрњрќрњѓјќачѓашјгјчујигадѓќрчњркчдргѓањфјќањрфграр
гѓазјжцрфгргрчдшјгдчѓлбдњршдцјњрујќдг
```

Running the algorithm on this ciphertext with 5000 iterations and the scaling parameter set to 1 (i.e. no scaling) yields:

Recovered plaintext

```
жуманизмотеидеологијанаренесансататојенејзинафилозофијанејзиниотпоглед
насветотжуманизмотнајпрвинсепојавилвоиталијавовекаблизувекаподохнавошп
анијафранхијаанглијаидругиевропскиземјиидеолозинаренесансатабилежумани
ститетиенасвоитесовременихиимпонудиленовпогледнасветотрелигиозниотпогл
еднасветотгозаменилесосветовенграганскиспореднивчовекотетојвокоготреба
дасеверувааневонадприроднисуштествабидејќисештоеизграденоназемјатаедел
онанеговитетворечкиспособностизачовекотнепостоидругцивотосвеновојназем
јатаизатоаземскиотцивотнетребадасепрезрекакоштопроповедалахрквататукут
ребадасепочитуваиунапредувавеченцивотнепостоиинеприфативадогматадека
човекоттребадасеодкацеодситеблагодатиназемскиотцивотзадацивеевечновоиз
мисленИотзадгробенцивотзамисленкакопеколчистилиштеирајспореджуманистит
есмислатанапостоењетоеовденаземјатаичовекотенејзинотонајсовршеносуштес
твопоткрепазаовиесваќањажуманиститенаоѓалевоантичкатанаукалитератураск
улптураиаржитектурапрекукоиоткриледругивистинизацивототпоинаквиодсредн
овековнитенекоиоджуманиститесвоитеделагипишуваленалатинскиотмртовјазик
инивнитеделадабидатдостапнизашироккруглуѓезажуманиститепоимотжуманизам
значелоафримахијаислободаначовековаталичностразвитокотналитературатаво
епожатанаренесансатапоминуванизтрифазиштогиобединуваеднажуманистичкако
нхепхијадасепишувазачовекоттиесеранаренесансависокаренесансаидохнарене
сансазаранатаренесансакарактеристиченесмешниотрасказсоантиклерикалнаиа
нтифеудалнанасоченостцбокачодекамерон
```

Which is pretty nice, I suppose :-) (`Хуманизмот е идеологија на ренесансата [...] Џ. Бокачо „Декамерон“`).

References

[Wik15] Wikipedia. *Metropolis–Hastings algorithm*. 2015. URL: https://en.wikipedia.org/wiki/Metropolis-Hastings_algorithm (visited on 10/17/2015).

[CR12] Jian Chen and Jeffrey S. Rosenthal. "Decrypting Classical Cipher Text Using Markov Chain Monte Carlo". In: *Statistics and Computing* 22 (2 Mar. 2012), pp. 397–413. ISSN: 0960-3174. DOI: 10.1007/s11222-011-9232-5. URL: http://probability.ca/jeff/ftpdir/decipherart.pdf (visited on 10/17/2015).

[Dia09] Persi Diaconis. "The Markov Chain Monte Carlo Revolution". In: *Bulletin of the American Mathematical Society* 46.2 (Apr. 2009), pp. 179–205. ISSN: 0273-0979. URL: http://math.uchicago.edu/~shmuel/Network-course-readings/MCMCRev.pdf (visited on 10/17/2015).

A The ciphertext that I will send

The text whose encryption I will send to Petar Tonkovikj is found in `InputPlaintext.txt`. The resulting ciphertext reads:

OutputCiphertext.txt

```
шмлфвќшґуќвњлсжулслґвгвжуљжбцљшцвджвмшслґшфшузжсшдшалуфлсуњљжњжќжлчшфш
чсвќшжсдвсжњвґшњшґлквсжґжљвчцмґжквсжзжслсґшќљжумфњжґжуґжњжшуфлсуњљжзжс
жгґшзжчґшњжќшќжфшувчшмуњшґшсжужуњшљлґслсґшґшгвжґшсвуњвцќшњшзфшќњвќшфшу
вчмжњќшќжвзжсжгґлнґвњлзфжвчљжсвґшзфшќњвцґвжњшмлфвќшґуќвсцпзфлсуквсзфшд
мшњвцлґзфвуњшзужсфнвґшњшґлквсшљшнґшљшнґжквсжсшицґќгвжґвфшсцолњжквслмжк
всвчвфшґвљфчжуґжљшґшґшбмшсвжњвґшбзжљфеґвжњчшлсґвцќввмлґвњлсґжжсджљжфвс
лґшњжбзжљвќsцолњжќжвдвґшзцењвсжулуљжвњлулсшвулмлбуњљшќжвдвчшкжфшљвсжлуљжб
жњбшчвќвжквцшвќжвґлґждшевмквслушулдшквслсвелґвжсуењждвзфшљлзжсшквсшњсл
sжсґлќжбшчшлсґвгшњфлкшsжсшбшзфвишњшњкжбшњшґшуљжбшњшќжншќшќжчшмлґшчшуењ
жвчдцквслґжљвњлмшґвзцsшњжфвзфшљлслуњвлsцолсшулзжццљуњљцљшшњджфсвчшфшсв
ґлењжењжґлквсжґвњцsвцґшќшфшќњлфвуњвќшґвњцќшќжбшчвќжњуњлќґшњшузжужкґжуњ
увњлњвлулдшквсллсґшґшгвбшґшгвбшґшклсвњлsцолклсвњлнлґввклсвњлслгшуњшґцљ
шслљшнґвушмжќшќжуќшsзвфшґвнфњљввсжќшччшклуњвбшsґжуњшґшжґвлќжвњфлкшsжсш
квсшњвуњфлклґвзжњљфсшчшжљшжувфжмшецљшаллвзжзвужњґшивќњвљґвњлжужквґвќжв
клsгвњлґшљжсґждвузжслsцљшslклsшќфљклsвмвysвзфвзшсґжуњќжґклsшњшфшушсжsн
ґвгвњлклчсжмґвгвњлвysцдвњлќшќжклsгввмшslучшлсґвцќжужкшґќшфвњлчлмбвеґвњ
лезлќцsшґњввужзуњљлґвгвњлґшзsшґњшнвњлшґвењжужjфљлґжќжегвњljфґгвњлнжsњв
њлужлсвґлњвґшњшќљвжуґжљвмжнlslсшквсшњмжквsвчвфшґввжфдшґвчвфшґвљжклsвќж
sжґвдфцзвчшsвґцкжфгвзфжњвљвґсвбшґгвњлфшувчмжњзфљжквњґжквsушмжлслґжсґшц
вґвњлчшмжквsвчшгвбшґшќжsжґвбшsґвњлшфмввґжвшќжљжшмлфвќшквsќжфвуњлґзжљлх
лжњќжsќцќшслквsжґшсфцджмлуњжулзшќквsушмжсжзжsґцљшалґшсфцдвњлмлњжсвнфњљ
вњлґшзвжґлфвњлжуљжбцљшцввзжґшњшмцквslжзвецљшґвќшќжґлљлфґвгвклчкжнґвгвґ
жузвцґжќшќжвґвљґвњлзфлњпжсґвгвпжsшґошґвњлзжљлхлњжсжулslґвгвквsлзфжљлуњ
шґњвќжвґшклчкжнґжуњшдsлсшslќшќжґшґлењжењжмжнлсшквслушмжќшчґлњжґжґлвќжф
лдвфшґжвуњжњшќшчшнфњљвњлулчкжфцљшsжслќшулсвљбшгвцжљлќжбшсгввзфвмвњвљгв
ґжвжљвлњлфмвґвулќжфвуњлґвќшќжувґжґвмчшуењжґллклsжчшужуњжбкшњшќжбшґлмжн
лсшквслзжсжкфлґшфшувчмжњквsвслжsждвбшќжбшквsшужљфелґжзфвузжужклґшґшзфш
ќњвќшњшґшзжфжкцљшалњжввуњфлкцљшалњж
```