Solving Substitution Ciphers with Genetic Algorithms

Ladislav Šulák (laco.sulak@gmail.com)
Krisztian Benko (kristianbnk@gmail.com)

Home university: Brno University of Technology, Faculty of Information Technology (Czech Republic, but we both are from Slovak Republic)

Substitution cipher

* Method for encrypting text in classic cryptography (plaintext ↔ encrypted text).

* Each individual symbol of alphabet is being substituted to other or same symbol of alphabet.

Vigenere Cipher

* **Polyalphabetic cipher - there are more various substitutions involved, for example each character could be encrypted with different substitution function.**

* **Final encrypted text is calculated with Vigenere table. Character in each position is determined by given character in plaintext and character in key.**

* **Example:**

plaintext	“vig**e**n**e**r**e**scip**h**e**r**”
key	“key**k**e**y**k**e**ykeyk**e**y”
encrypted	“FME**O**R**C**B**I**QMMNR**I**P”

What TODO?

* Prepare data = pairs of (plaintext, key) + encrypted text

* Implement console app with the use of Genetic Algorithms in Python

* Evaluate results (precision, number of generations needed, ...) and write documentation

Dataset

*** Use Vigenere substitution cipher (pycipher library in Python)**

*** A couple of tests, where every test will contain:**

- plaintext as a text in English language
- key as pseudo-randomly generated string
- encrypted text (use pycipher)

Motivation

* Trying to recover plaintext from text encrypted by Vigenere cipher.

* Brute Force method (trying every possible key on ciphertext) can have very high computational complexity.

* Usage of GA may be a good optimalization heuristic.

Usage of Genetics Algorithms

* Calculate key length (well known approach)

* Each individual in the population will represent 1 guess of cryptographic key used during encryption.

* Fitness function then takes such key and use it on encrypted text resulting in 1 possible plaintext.

* The Fitness evaluation is based on methods which are trying to determine if the word belongs to English language or not (Markov Chain Models = n-gram frequencies, or frequency of characters in English language)

Discussions

Questions?