Uniwersytet Przyrodniczy we Wrocławiu  
Wydział Biologii i Hodowli Zwierząt  
Bioinformatyka



Łukasz Chołodecki

Biblioteka programistyczna wykorzystująca metody analizy statystycznej wobec wybranych klasycznych metod szyfrowania

Praca licencjacka pod kierunkiem  
 dra Jana Jełowickiego

Wrocław, 2021

**WSTĘP DO SZYFROWANIA**

**Zadanie szyfrów**

Szyfrem jest dowolny algorytm, często również przedstawiany jako funkcja szyfrująca, której głównym zadaniem jest takie zniekształcenie przekazywanej informacji, aby odbiorca postronny nie był w stanie jej odczytać, a jednocześnie możliwe było jej pełne odtworzenie przez osoby posiadające odpowiednią wiedzę. Początkowo szyfry były wykorzystywane nielicznie, głównie w zastosowaniach wojskowych, ale wraz z czasem ich popularność rosła, a w trakcie pierwszej i drugiej wojny światowej były one kluczowym elementem rywalizacji pomiędzy stronami. Obecnie, ze względu na powszechność ich wykorzystania w świecie cyfrowym, kryptografia odgrywa kluczową rolę ze względu na podstawę wszystkich zabezpieczeń.

**Historia szyfrów**

Za pierwszy szyfr w historii uznaje się, pochodzący z V wieku przed naszą erą, skytale, opracowany przez starożytnych Spartan. Metoda zapisu polegała na owinięciu długiego paska skóry na lasce o specyficznej grubości i zapisaniu wiadomości w poprzek. Dzięki temu, po rozwinięciu paska otrzymywano zniekształcony szyfrogram. W gruncie rzeczy był to zwykły szyfr przestawieniowy, jednak możliwy do szybkiego odszyfrowania dzięki specyficznemu, fizycznemu kluczowi.

Drugim szyfrem pochodzącym z tego okresu był hebrajski atbasz, w którym litery były wzajemnie zastępowane przez te oddalone o równą długość od drugiego końca, np. A-Z, B-Y. Stąd też pochodzi nazwa tego szyfru, powstała z zakodowania dwóch pierwszych liter hebrajskiego alfabetu: alef-taw, beth-shin.

Następny szyfr to jeden z opracowanych przez Juliusza Cezara (i od niego później nazwany) — szyfr Cezara, pochodzący z I wieku przed naszą erą. Metoda szyfrowania opiera się na cyklicznym przesunięciu alfabetu o stałą wartość. Oryginalnie szyfr wykorzystywał przesunięcie o 3 znaki, ale był później używany również z innymi wartościami obrotu.

Bezpieczeństwo tych metod szyfrowania opierało się głównie, nie na sile szyfru, a nieznajomości zastosowanej metody szyfrowania, co wraz z powstaniem metody kryptoanalizy opartej na statystyce częstotliwości liter w IX wieku, sprawiało, że konieczne było stosowanie bardziej skomplikowanych algorytmów szyfrujących.

Kolejną próbą, był szyfr Marii Stuart, opracowany na potrzeby królowej Szkocji, rywalizującej z królową Anglii, pod koniec XVI wieku. Szyfr oparty był na zamianie liter na symbole oraz dodatkowej zamianie najczęściej używanych słów na pojedyncze symbole, kilku znakach odpowiadających przerwie między literami i zastąpieniu podwojonych liter specjalnym znakiem. Mimo zastosowania dodatkowych zabezpieczeń oraz ze względu na przekazywanie przez posłańca wszystkich wiadomości do angielskich kryptoanalityków (a więc dużej bazie tekstu), szyfr został złamany, a królowa Maria Stuart skazana na śmierć za zdradę korony.

W wyniku rozwiniętej kryptoanalizy jakiekolwiek szyfry opierające się na monoalfabetycznych przekształceniach i podstawieniach były możliwe do złamania, stąd zaczęto skłaniać się ku szyfrom bardziej skomplikowanym. Pierwszy szyfr, Johannesa Trithemiusa, oparty był o tabula recta — kwadrat złożony z wypisanego alfabetu w pierwszym wierszu, a następnie przepisywanemu do poniższych wierszy, przesuwając go o jedną literę (jak w szyfrze Cezara). Kolejne wariacje, również oparte na tabula recta, wykonali Giambattista della Porta, a następnie Blaise de Vigenère, końcowo tworząc szyfr nazwany szyfrem Vigenère’a. Szyfr ten opiera się o kodowanie z wykorzystaniem wielu alfabetów oraz dodatkowego, unikalnego klucza, który był wielokrotnie powtarzany, równolegle do tekstu, przez co podatny był na kryptoanalizę. Pomimo tego, aż do drugiej połowy XIX wieku nie został on złamany.

Od XIX wieku nastąpił znaczny rozwój zarówno kryptografii, jak i kryptoanalizy, przestano również polegać na utrzymywaniu algorytmu szyfrującego w tajemnicy jako głównego elementu bezpieczeństwa szyfrów, a skupiono się na znaczeniu klucza, który powinien pozwalać na zachowanie wiadomości w tajemnicy, tak długo, jak sam klucz nie zostanie odgadnięty.

Jednym z pomysłów był szyfr Playfair, opierający się o kodowanie tekstu podzielonego na dwuznaki, zamiast pojedyncze litery, co znacząco utrudniało analizę częstotliwości, ze względu na o wiele większą liczbę możliwych kombinacji. Oprócz tego używany był klucz, zapewniający tajność. Jednak ze względu na liczne, często mniej oczywiste podatności, został on stosunkowo szybko złamany.

Pod koniec pierwszej wojny światowej powstał niemiecki szyfr ADFGVX, który stosował podwójne szyfrowanie (oraz dwa klucze) — najpierw podstawienie każdej litery i cyfry według wzoru (wykorzystując wersję szachownicy Polybiusza) oraz szyfr przestawieniowy, rozbijający powstałe we wcześniejszym etapie dwuznaki. Oprócz tego, dzięki zapisie całej wiadomości za pomocą tylko sześciu liter, znacząco ułatwiło to jego nadawanie za pomocą kodu Morse’a. Pomimo stosunkowej złożoności metod i użycia dwóch kluczy (co miało według Niemców zapewnić zupełną odporność szyfru na łamanie) został złamany w ciągu trzech miesięcy.

W kolejnych latach, wraz z rozwojem technologii i elektroniki, zaczęły powstawać znacznie bardziej skomplikowane maszyny szyfrujące, pozwalające w łatwy sposób zakodować i odczytać wiadomość, przeprowadzając serię przekształceń, która w wykonaniu człowieka zajmowałaby znacznie więcej czasu. Jednym z najbardziej znanych przykładów takich elektromechanicznych była Enigma, powstała na początku lat 20., na początku w komercyjnych zastosowaniach, a później w trakcie drugiej wojny światowej przez niemieckie wojsko. Zbudowana była z wielu niezależnych od siebie elementów, które mogły być łączone w różnych konfiguracjach, m.in. wirniki, łącznica kablowa i uzwojenia. Dzięki pracom polskich matematyków pierwszy raz udało się złamać szyfr w 1932, jednak wraz z upływem czasu poziom skomplikowania maszyn rósł, przez co w momencie wybuchu wojny wersja używana przez wojsko była dalej niezłamana. Polacy następnie przekazali wyniki swoich badań sojuszowi Aliantów, co znacząco przyspieszyło ich prace. Ostatecznie złamana została w pierwszej połowie 1940 przez Alana Turinga, dzięki konstrukcji zaawansowanego narzędzia kryptoanalitycznego zwanego bombą.

W wyniku co raz szybciej postępującej automatyzacji, powstaniu pierwszych komputerów przekraczających dotychczasowe możliwości obliczeniowe, zwykłe szyfry nie zapewniały jakiejkolwiek ochrony ze względu na możliwość złamania takiego szyfru w ekstremalnie krótkim czasie. Nowe metody musiały być na tyle skomplikowane obliczeniowo, aby ewentualna próba złamania zabezpieczenia, nawet z użyciem komputera zajęła niewspółmiernie dużo czasu.

W teraźniejszych czasach znaczenie szyfrów jest niezwykle ważne, ponieważ to one odpowiadają za wszystkie zabezpieczenia obecne w urządzeniach elektronicznych — od telefonów po serwery sieciowe. Jednym z najpopularniejszych szyfrów, używanych od początku XXI wieku jest algorytm AES szyfrujący 128-bitowe bloki danych przy użyciu klucza 128, 192 lub 256-bitowego. Pomimo szerokiego zastosowania w bezpieczeństwie od ponad 20 lat nie znaleziono żadnego sposobu na bezpośrednie złamanie tego szyfru, a wynalezione sposoby osłabiania, nawet dla klucza 128-bitowego, wymagałyby milionów lat pracy na komputerach z obecnym rzędem wydajności.

**Cechy szyfrów**

Pierwszy podział szyfrów dzieli je na sposób działania. Szyfry przestawieniowe nie zmieniają samych znaków wiadomości, a jedynie ich kolejność. Ze względu na zwyczajowe korzystanie z prostych przekształceń, liczba możliwości jest mała i powtarzalna między różnymi szyfrowaniami, stąd też małe bezpieczeństwo takiego szyfru, można natomiast stosować go w kombinacji z szyframi innego rodzaju.

Innym rodzajem, znacznie popularniejszym, są szyfry substytucyjne, które według różnych zasad zamieniają litery tekstu na inne ustalone znaki, litery, bądź symbole. Te szyfry jednak oferują bardzo słabą ochronę w pierwotnej wersji, jeśli cały tekst przekształcany jest jednakowo (szyfry monoalfabetyczne). Szyfry polialfabetyczne dokonują zmian liter według dodatkowych reguł, dzieląc tekst na kilka grup, z których każda szyfrowana jest w inny sposób (najczęściej z wykorzystaniem innego klucza). Drugim podejściem są szyfry homofoniczne, w których jednemu znakowi, odpowiada kilka różnych znaków (bądź częściej zbitek znaków jak liczby czy słowa). Przy odpowiednim dopasowaniu ilości różnych znaków dla jednej litery na podstawie oczekiwanej częstotliwości wystąpienia tej litery, możliwe jest całkowite uodpornienie szyfru na prostą kryptoanalizę statystyczną. Możliwe jest też szyfrowanie, nie pojedynczych znaków, a ich grup — na przykład dwuznaków. Takie rozwiązanie również znacznie utrudnia analizę statystyczną, ze względu na znacznie większe rozdrobnienie analizowanych elementów.

Obecnie szyfry można podzielić na kilka grup, według różnych kryteriów. Podstawowa dzieli szyfry ze względu na rodzaje klucza. Pierwsza, obejmujące powyższe szyfry historyczne, jest to grupa szyfrów symetrycznych (szyfrów z kluczem prywatnym). W założeniu obie strony znają ten sam klucz i za jego pomocą szyfrują wzajemnie wysyłane wiadomości, które następnie rozszyfrowują z wykorzystaniem tego samego klucza. Drugi model — asymetryczny (szyfr z kluczem publicznym) posiada dwa klucze — publiczny, pozwalający dowolnej osobie na zaszyfrowanie wiadomości, oraz klucz prywatny, który posiada tylko odbiorca i tylko za jego pomocą można wiadomości odszyfrowywać.

**Bezpieczny szyfr**

Pod koniec XIX wieku wymyślono termin szyfru z kluczem jednorazowym (one-time pad), który pozwalał przy spełnieniu pewnych założeń, na szyfrowanie w pełni odporne na dowolną kryptoanalizę. Pierwszym historycznym szyfrem, który można wykorzystać z kluczem jednorazowym był szyfr Vigenère’a, później popularność zdobyło szyfrowanie bitowe za pomocą operatora XOR (z racji równego prawdopodobieństwa zwrócenia wartości 0 i 1, powodując, że każda wartość z równą szansą została, bądź nie została zmieniona). Jeśli klucz, którego użyjemy, spełni poniższe warunki, niemożliwe będzie odtworzenie wiadomości. Po pierwsze klucz musi być w pełni losowy – oznacza to, że nie istnieje żadna metoda, która pozwoliłaby celowo ponownie wygenerować ten sam klucz (niezależnie od formy, jaką przyjmuje). Po drugie klucz, musi być co najmniej długości szyfrowanego tekstu – dzięki temu, nie będzie on wykorzystany drugi raz do zaszyfrowania innej części tekstu, co pozwoliłoby na porównywanie ze sobą dwóch elementów zaszyfrowanego tekstu. Z tego samego powodu klucz nie może być ponownie wykorzystany, do szyfrowania innej wiadomości.

Choć taka metoda szyfrowania jest w pełni bezpieczna, wymagania potrzebne do jej użycia stanowią problem – ponieważ w bezpieczny sposób musimy przekazać hasło długości wiadomości, a więc zazwyczaj możemy po prostu przekazać w taki sposób wiadomość. Z tego powodu zastosowanie ma to głównie dla niezwykle ważnych kanałów, dla których w przeszłości w bezpieczny sposób obie strony ustaliły hasła.

**METODY KRYPTOANALITYCZNE**

**Wstęp do kryptoanalizy**

Kryptoanaliza rozwijała się wraz z postępem w tworzeniu nowych metod szyfrowania. O ile pierwsze szyfry nie wymagały skomplikowanych metod do ich łamania, to kiedy zaczęto stosować szyfry substytucyjne, potrzebne były bardziej skonkretyzowane narzędzia. W IX wieku arabski matematyk Al-Kindi opublikował pierwszą pracę poświęconą temu tematowi, gdzie przedstawiał metody analizy częstotliwości, opisy częstotliwości znaków w języku arabskim i sposoby łamania różnych szyfrów, w tym polialfabetycznym. W Europie dopiero w drugiej połowie XVI wieku pojawiła się pierwsza praca dotycząca tego tematu, napisana przez włoskiego uczonego Giambattista della Porta, choć już wcześniej zajmowano się tą tematyką.

Sama kryptoanaliza dzieli się na dwie zasadnicze części. Pierwszą jest odkrycie działania szyfru wykorzystanego do zaszyfrowania danej wiadomości. Im dokładniejsza jest znajomość algorytmu szyfrowania, tym łatwiej później złamać szyfr, chociaż nie jest to warunkiem koniecznym – jeśli wiemy tylko, że użyty szyfr korzysta z jakiejś monoalfabetycznej metody substytucyjnej, możemy go złamać, stosując najszersze podejście – jednak jeśli wiedzielibyśmy, że użyto konkretnie szyfru Cezara, złamanie byłoby znacznie prostsze przy wykorzystaniu odpowiedniej metody.

Drugą częścią jest próba złamania bezpośrednio klucza (jeśli szyfr nie posiada klucza, sama znajomość algorytmu, pozwala na rozwiązanie szyfru). W przypadku kiedy szyfr jest unikatowy, należy najpierw opracować, lub dostosować metody łamania, wyszukać podatności i dopiero wtedy przystępować do łamania. Dla bardziej skomplikowanych szyfrów potrzebne są wielostopniowe działania, np. próby osłabiania szyfru, eliminowania możliwych kluczy etc.

**Podział ze względu na źródło**

Największe zróżnicowanie w kryptoanalizie występuje ze względu na rodzaj dostępnych materiałów. Najczęstszym przypadkiem jest sytuacja, w której mamy do dyspozycji tylko zaszyfrowany tekst (bądź teksty), zazwyczaj z użyciem różnych kluczy, ale możliwe są także sytuacje, w których mamy również dostęp do czystego tekstu (przed zaszyfrowaniem

**Metoda brute-force**

Atak brute-force jest najprostszym możliwym sposobem łamania szyfrów, a mimo to, jest dalej szeroko stosowany (zazwyczaj jako jeden z elementów). Polega na prostym, siłowym próbowaniu każdej możliwej kombinacji szyfru, aż odnajdzie się prawidłowy klucz. Rozwinięciem tej metody jest, np. metoda słownikowa, która zakłada, że prawdopodobne jest użycie jako klucz istniejącego, popularnego słowa i zaczyna atak od zbadania słów z zebranego zasobu — słownika. Głównym ograniczeniem tej metody jest ilość możliwych kluczy do zbadania – o ile w przypadku szyfru Cezara istnieje tylko 26 wariantów w postaci liczby z tego zakresu, to wraz z długością hasła i możliwymi znakami do użycia, poziom skomplikowania rośnie wykładniczo – dla klucza składającego się z 4 dowolnych liter (bez rozróżnia wielkości), możliwe jest prawie pół miliona kombinacji, a dla odpowiednika o długości 8 – już ponad 200 miliardów. Z tego powodu próba złamania nowożytnych szyfrów tylko w ten sposób, przy użyciu obecnej technologii, jest bezcelowa, ponieważ niemożliwe jest uzyskanie wyników w rzeczywistym czasie.

**Test chi-kwadrat zgodności**

Kiedy zaczęto odchodzić od prostych szyfrów i zaczęto opierać się na szyfrach substytucyjnych, metoda brute-force przestała mieć zastosowanie, ze względu na ogromną ilość możliwości, którą trzeba by manualnie sprawdzić, skierowano się wówczas do bardziej skomplikowanych metod. Odkryto, że każdy język naturalny posiada swoje unikalne cechy związane m.in. z częstotliwościami występowania różnych liter, a w przypadku szyfrowania monoalfabetycznego pozostają one niezmienione – przekształceniu ulegają tylko symbole reprezentujące poszczególne litery. Obecnie do porównywania częstotliwości występowania wykorzystuje się test chi-kwadrat zgodności (Pearsona), w którym statystyka testowa dana jest wzorem:  
Gdzie Oi to ilość obserwowana, a Ei to odpowiadająca ilość oczekiwana. Wartości testowe przyjmują wartości dodatnie, a im lepsze dopasowanie, tym niższa wartość.

**Kappa-test (index of coincidence)**

Współczynnik kappa określa z jakim prawdopodobieństwem analizując, litera po literze, dwa teksty napisane w tym samym języku mamy szansę trafić na tę samą literę – jego wielkość wynika ściśle z rozłożenia częstotliwości liter w danym języku, bez znaczenia na ich kolejność. Oznacza to, że tekst w danym języku naturalnym będzie zawsze miał podobną wartość współczynnika kappa – (nawet po zaszyfrowaniu dowolnym monoalfabetycznym szyfrem substytucyjnym, jako że kolejność nie ma wpływu na kappa). Test kappa dany jest wzorem:

Gdzie ni to ilość wystąpień poszczególnych liter w tekście, a N jest sumą wszystkich liter.

**Podatności szyfrów**

Oprócz różnych metod ogólnych służących do rozszyfrowania ważnym elementem są podatności szyfrów. Większość szyfrów w wyniku zróżnicowanej i oryginalnej budowy posiada różne cechy, nieprzewidziane przez twórcę, które mogą znacząco pomóc w kryptoanalizie. Podatności mogą również dotyczyć przesyłanego tekstu i powtarzających się fraz szyfrowanych.

**ANATOMIA JĘZYKA POLSKIEGO W UJĘCIU KRYPTOANALITYCZNYM**

**Wstęp do badań anatomii języka**

Chcąc korzystać z analizy statystycznej jako metody kryptoanalizy, potrzebny jest punkt odniesienia, specyficzny dla danego języka naturalnego. Moim celem było opracowanie podstawowych statystyk, które mogą być wykorzystane zarówno do bezpośredniej kryptoanalizy, jak i służyć do sprawdzania wyników.

Badane były cztery najbardziej uniwersalne częstotliwości:

* Występowania znaków, dwuznaków i trzyznaków
* Występowania ich na początku i na końcu słów
* Krótkich słów (od 1 do 3 znaków)

Oprócz tego w oparciu o częstotliwości znaków wyznaczono współczynnik kappa.

**Korpus języka polskiego**

Do badania anatomii języka potrzebna jest duża, zróżnicowana baza tekstu w tym języku, tak, aby, jak najwierniej go odwzorować. Posłużyłem się do tego celu Narodowym Korpusem Języka Polskiego zbudowanym przez IPI PAN oraz PWN. Jest to największy publiczny zbiór tego rodzaju, zawierający łącznie ponad 6 milionów liter.

Teksty przechowywane były w plikach formatu XML (zawierających oprócz czystego tekstu wiele dodatkowych informacji lingwistycznych, które nie były wykorzystane). Aby wyodrębnić czysty tekst, napisałem prosty konwerter z wykorzystaniem biblioteki XML DOM minidom.

**Częstotliwości znaków**

Pierwszą, najczęściej wykorzystywaną statystyką jest częstotliwość liter, oraz dwu- i trzyliterowców. Ma ona znaczenie o tyle szczególne, że może być wykorzystywana w każdej kryptoanalizie tekstowej, nawet jeśli tekst został zniekształcony i pozbawiony spacji i znaków interpunkcyjnych. Statystyka liter może być wykorzystywana

**Porównanie do języka angielskiego**

**OMÓWIENIE BIBLIOTEKI**

**Zakres, który obejmuje (opis, teoria)**

**Wykonanie – wygląd modułów, budowa całej biblioteki**

**OGÓLNE:**

**Obróbka tekstu wejściowego**

**Badania częstotliwości**

**SZYFROWANIA:**

**Cezar**

**Substytucja**

**Vigenere**

**Playfair**

**METODY KRYPTOANALITYCZNE**

**PRZYKŁADY ZASTOSOWAŃ**

**BIBLIOGRAFIA**