1, 2, 3, кодирај ! - Активност циклуса 4 - Пројект « Криптографија » - Час 4: Разбијање моно-алфабетског шифрирања

Доминантна дисциплина	Математика
Резиме	Ученици упознају и примењују метод анализе фреквенције коју је увео Al-Kindi, која се заснива на односу статистике и лингвистике.
Појмови	« Информације » • Шифрирање моно-алфабетском супституцијом може бити откривено анализом фреквенције појаве слова • Анализу фреквенције је увео Al-Kindi у деветом веку
Матријал	За одељење: • (Факултативно) игра Scrabble© За сваког ученика: • Радни лист-B06 • Радни лист-B07 • Радни лист-B08

Полазна ситуација

Наставник подсећа на оно што је рађено на претходном часу у вези шифрирања моно-алфабетском супституцијом која нам изгледа доста сигурна јер има велики број могућих кључева. Поверити све могуће кључеве захтева хиљаде година рада (суперкомпјутера) ...или милјарде година ручног рачунања.
Напомиње да се овај тип шифрирања не користи од средњег века јер је персијски научник нашао метод који омогућује да се кључ овог типа шифрирања врло лако пронађе (Al-Kindi, 9-ти век.).

Наставник пита ученике како је Al-Kindi могао то да уради. Ако немају идеју води их постављујући им питање како су они успели да разбију Цезарево шифрирање (Час 2).
Одељење је, осим могућности да проба све могуће кључеве, означило слово које се најчешће појављује у француском језику (« E », *с наставником српског језика проверите које је то слово у српском језику, п.п.*), затим су поставили хипотезу да су сва слова померене за исти број места као слово « E ». Ова хипотеза није потврђена у оквиру моно-алфабетског шифрирања јер сваком слову може, *a priori*, одговарати било које друго слово. Потребно је, дакле, тражити помоћ у фреквенцији више слова, а не само једног.

Проналажење најфреквентијих слова у француском језику (заједнички, затим у пару) (ви можете ово исто да покушате у српском језику п.п.)

Наставник пита ученике да ли знају која се слова најчешће појављују француском језику, а ако то не знају како би могли да одреде.

Први приступ у решавању проблема би могао да се састоји у упознавању игре Scrabble. Свако слово поседује неку вредност и, сасвим разумљиво, слово које се најређе појављује има више поена (теже се у некој француској речи поставља слово W него E!). Осим тог, запажамо да слова која имају мање поена (дакле, чешће се појављују), су представљена бројевима у овој игри.

Слова у Scrabble© (вредност сваког слова је у индексу)	**A**$_1$	**B**$_3$	**C**$_3$	**D**$_2$	**E**$_1$	**F**$_4$	**G**$_2$	**H**$_4$	**I**$_1$	**J**$_8$	**K**$_{10}$	**L**$_1$	**M**$_2$	**N**$_1$	**O**$_1$	**P**$_3$	**Q**$_8$	**R**$_1$	**S**$_1$	**T**$_1$	**U**$_1$	**V**$_4$	**W**$_{10}$	**X**$_{10}$	**Y**$_{10}$	**Z**$_{10}$
Број појављивања сваког слова.	9	2	2	3	15	2	2	2	8	1	1	5	3	6	6	2	1	6	6	6	6	2	1	1	1	1

Пошто је видео шта су ученици урадили, представља на табли расподелу појаве слова и тражи од ученика, по паровима, да направе хистограм.

Овај хистограм је врло близак оном који се односи на француски језик:

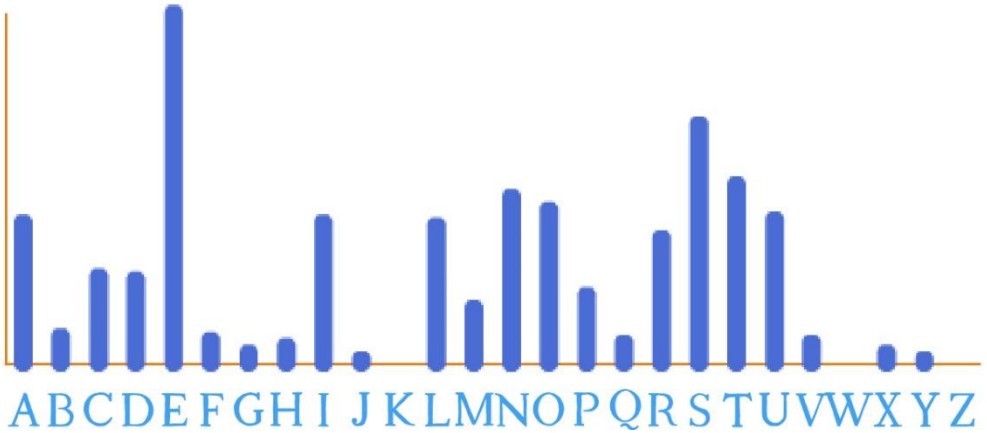
Научне напомене

- Није могуће направити референтни хистограм за неки језик, јер фреквенција појаве слова зависи од типа текста и форме писања (дипломатски телеграм, литерарно дело, енциклопедија,породични језк...). Међутим, разлике у различитим регистрима могу бити занемарљиве када је у питању овај педагошки пројект.
- Референтни хистограми који се појављују у Радном листу-B07 су добијени анализом великог броја текстова (речника, литерарних дела...) [Извор: en.wikipedia.org/wiki/Letter_frequency]. Ми смо то модификовали узимајући у обзир нашу поједностављену азбуку (код статистике за слово « e » се групишу и слова « é », « è », « ê »…).

Заједничко представљање

Наставник, после корекције ученичких радова, пита одељење да ли су фреквенције које су нашли за игру Scrabble у вези с фреквенцијама слова у француском језику. Предлаже им да анализирају текст који се налази у вежби 1 Радног листа-B06 (у питњу је преамбула Универзалне декларације о праву човека и грађанина из 1789).
Сваки ученик ради само са 2 или 3 слова (што омогућује поређење радова више ученика у вези истог слова и тако пружа могућност уочавања евентуалних грешака), јер се добија у времену.

Добија се, коначно, заједнички дијаграм:



Напомена: овај график је направљен помоћу програма у Скрачу коришћеног у *секвенци 2*.
Апсцисна оса овде није градуисана јер нас интересује само општа форма хистограма.

Запажамо, чак и у случају релативно кратког текста, да је добијени хистограм сличан оном у игри Scrabble© или референтном хистограму за француски језик. Чак и кад је текст довољно дуг (реда неколико стотина слова), налазимо следећу шему:

- Слово e је најфреквентније. Једино код њега фреквенција прелази 15%

- Следе га слова a (9%), i (8%), s (8%), t (7%), n (7%) et r (6%)
- Само осам слова има фреквенције мање од 1% : f, h, j, k, w, x, y et z (а једанаест слова заједно има фреквенцију мању од à 2 %)

Криптоанализа корак по корак (по групама или заједнички)

Наставник расподељује ученицима вежбу 2 из Радног листа-B06, којом се предлаже прављење критоанализе доста кратког текста (21 реч, 114 карактера, размаци и интерпункција нису узети у обзир). Криптоанализа је вежба која користи математику (анализа фреквенције засноване на статистици) и лингвистику (знање француског језика, најчешће коришћена слова и речи, итд.(*по угледу на овај пример можете покушати да ово исто урадите за српски језик, п.п*).

ZRJ VDAARJ CLWJJRCK RK ERARMHRCK ZWIHRJ RK RULMP RC EHDWKJ. ZRJ EWJKWCBKWDCJ JDBWLZRJ CR FRMNRCK RKHR TDCERRJ GMR JMH Z'MKWZWKR BDAAMCR.

Научне напомене

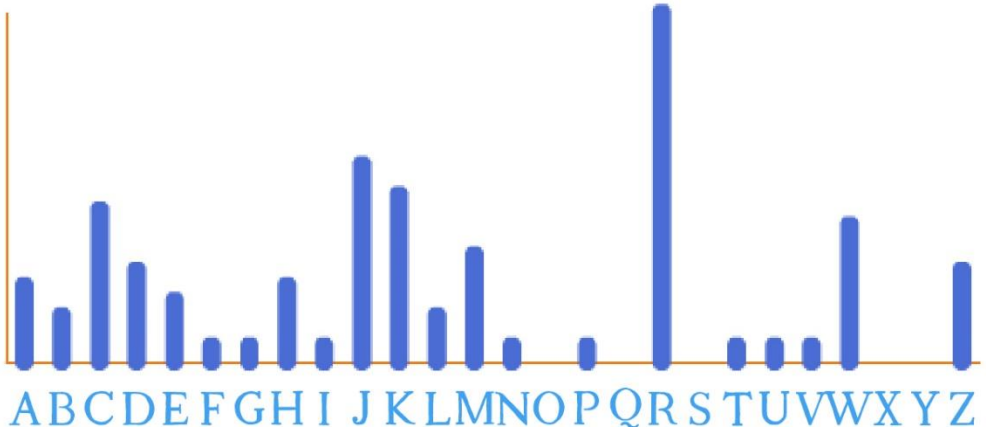
- Знатно је лакше направити криптоанализу дугог нео кратког текста, иако ми обично мислим супротно. Што је текст дужи то с више приближавамо референтном хистограму, а уз то већа статистика омогућује, сама или скоро сама, криптоанализу текста.
- У случају кратког текста, као у овом случају, нопходно је правити безброј насумичних покушаја између статистике и лингвистике да би се реализвала криптоанализа текста. Ипак, као што видимо ниже, успевамо то да урадимо!

Педагошке напомене

- Индивидуална реализација једне овакве вежбе је претешка за ученика основне школе. Међутим, заједничком реализацијом (или у малим групама уз време при заједничком представљању), омогућује да се то уради, а затим и да ученици покушају то исто да ураде самостално.
- Наставник, зависно од нивоа знања ученика предлаже ову вежбу у форми проблема (реализацијом у мањим групама или индивидуално). Ако су пак ученици блокирани покушајте да их усмерите. Управо из тог разлога вам наводимо метод « корак по корак ».

Етапа 1: релаизација хистограма текста

Ова етапа је већ реализована, у више наврата, током овог часа, па то не би требало да представља потешкоћу. Может је реализовати као претходно, тј., свако слово броји по један ученик.
Добијамо:



Етапа 2: тражимо најфреквентније слово

Слово (« R ») се јасно издваја, а затим следи 4-5 слова високе фреквенције, док 13 слова има врло ниску фреквенцију. Шема је слична оној на референтном хистограму, па можемо да претпоставимо да је реални ткст написан на француском (или енглеском, немачком, шпанском ...који имају доста сличне хистограме, како је то приказано на Радном листу-B07).
У том случају, можемо да поставимо хипотезу да је « R » шифрирано са словом « e ».
Ако заменимо « R » са « e », порука постаје:

ZeJ VDAAeJ CLWJJeCK eK EeAeMHeCK ZWIHeJ eK eULMP eC EHDWKJ. ZeJ EWJKWCBKWDCJ JDBWLZeJ Ce FeMNeCK eKHe TDCEeeJ GMe JMH Z'MKWZWKe BDAAMCe.

Етапа 3: тражимо најкраће речи

Настојећи да рад учинимо лакшим оставили смо размаке и интерпункцију... што нам омогућује да тражимо најкраће речи које су малобројне.
У поруци се налазе 2 речи с два слова: eK (понавља се 2 пута), eC и Ce. Најчешће речи од 2 слова, у француском језику, су (према редоследу): de, il, le, et, je, la, ne, un, en, ce, se, sa, du. Узимајући у обзир позицију слова « e » у речима са два слова, које су приказане овде, можемо да поставимо хипотезу да:

- « eK » одговара « et » (K→ t)
- « eC » би могло да одговара « en » (C →n). « Ce » одговара « de » или « le » или « ne » (C → d или C → l или C →n). Дакле, можемо да поставимо хипотезу да је C →n.

Према овој хипотези, текст постаје:

ZeJ VDAAeJ nLWJJent et EeAeMHent ZWIHeJ et eULMP en EHDWtJ. ZeJ EWJtWnBtWDnJ JDBWLZeJ ne FeMNent etHe TDnEeeJ GMe JMH Z'MtWZWte BDAAMne.

Почињемо да препознајемо блиске структуре (3-ћа, 5-та и 15-та се завршавају « ent », као галгол трећег лица множине), што нас охрабрује да даље наставимо рад с нашом хипотезом која је, бар до сад, наизглед коректна.!

Педагошка напомена
Лако налазимо листу речи с 2 или 3 слова у француском језику... јер је то интересантно за играче који воле Scrabble© ! Погледајте, на пример: www.listesdemots.com

На исти начин можемо тражити речи с 3 слова: ZeJ (понавља се 2 пута), GMe и JMH. Најчешће речи од 3 слова у француском језику су (према редоследу): que, les, son, mon, pas, lui, une, des, qui, est. Наша реч « ZeJ » би могла бити « les » (врло вероватно, јер се појављује врло често) или « des ».
Ако поставимо хипотезу да ZeJ →les (Z → l еи J → s), наш текст постаје:

les VDAAes nLWssent et EeAeMHent lWIHes et eULMP en EHDWts. les EWstWnBtWDns sDBWLles ne FeMNent etHe TDnEees GMe sMH l'MtWlWte BDAAMne.

И ако је GMe → que (најфреквентнија реч од 3 слова која се завршава са e), онда текст постаје (са G → q и M →u) :

les VDAAes nLWssent et EeAeuHent lWIHes et eULuP en EHDWts. les EWstWnBtWDns sDBWLles ne FeuNent etHe TDnEees que suH l'utWlWte BDAAune.

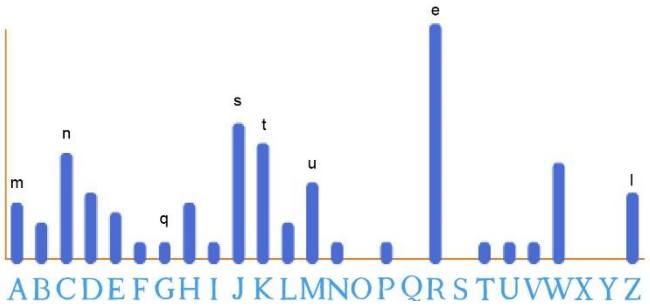
Етапа 4: тражимо дублете

У криптоанализи текста се врло често користе парови слова. Дублети представљени у овом тексту су: AA (понвља се 2 пута) и ss (у реалном тексту, који смо већ дешифровали). У француском језику најфреквентнији дублети су: ss, ll, mm, rr, tt, nn, pp, ee, cc и ff.
Пошто смо слова s и l већ пронашли, можемо да покушамо са заменом AA са mm. Дакле, ако A → m, текст постаје:

les VDmmes nLWssent et EemeuHent lWIHes et eULuP en EHDWts. les EWstWnBtWDns sDBWLles ne FeuNent etHe TDnEees que suH l'utWlWte BDmmune.

Етапа 5: повратак на слику фреквенција

Нађена слова су:



У француском језику су после слова e, најфреквентнија слова a, i, s, t. Пошто су s и t већ пронађени (и одговрају им пикови на овом хистограму), вероватно је да су a и i били скирвени иза преосталог фреквентнијег слова (W).
Ако поставимо хипотезу да је W →a, онда текст постаје:

les VDmmes nLassent et EemeuHent laIHes et eULuP en EHDats. les EastanBtaDns sDBaLles ne FeuNent etHe TDnEees que suH l'utalate BDmmune.

Изгеда да 3-ћа и предпоследња реч не постоји у француском језку. Хипотеза тако није потврђена. Покушајмо зато са W → i.Текст постје:

les VDmmes nLissent et EemeuHent liIHes et eULuP en EHDits. les EistinBtiDns sDBiLles ne FeuNent etHe TDnEees que suH l'utilite BDmmune.

Одлично! Препознаи смо реч « utilité » на предпоследњем месту.
Тако би сад 3-ећа реч « nLissent » могла бити « naissent ». У том случају, L → a, па добијамо:

les VDmmes naissent et EemeuHent liIHes et eUauP en EHDits. les EistinBtiDns sDBiales ne FeuNent etHe TDnEees que suH l'utilite BDmmune.

Враћањем на хистограм фреквенције, преостаје нам међу најфреквентнијим словима (a, i, s, t, n и r), само слово r. Преостали пик би могао да одговара за H или C. C је већ пронађено (одговара n). Преостаје H, па ако је H → r онда текст постаје:

les VDmmes naissent et Eemeurent liIres et eUauP en ErDits. les EistinBtiDns sDBiales ne FeuNent etre TDnEees que sur l'utilite BDmmune.

Препознајемо речи « être » и « sur » у другој реченици (запажање: могли смо ово да пронађемо посматрјући речи од 3 слова које почињу са « su »… постојала је могућност да то буде « sur » !). У првој реченици би 6-ста реч могла бити « libre ». У том случају (l → b), па би добили:

les VDmmes naissent et Eemeurent libres et eUauP en ErDits. les EistinBtiDns sDBiales ne FeuNent etre TDnEees que sur l'utilite BDmmune.

Дешифровање сад постаје врло лако: 5-та реч је « demeurent » (E → d).

les VDmmes naissent et demeurent libres et eUauP en drDits. les distinBtiDns sDBiales ne FeuNent etre TDndees que sur l'utilite BDmmune.

Контекст нам сад помаже да закључимо да је вероватно друга реч« homme », а последња реч прве реченице « droit ». У том случају (V → h, D → o), па добијамо:

les hommes naissent et demeurent libres et eUauP en droits. les distinBtions soBiales ne FeuNent etre Tondees que sur l'utilite Bommune.

Скор да смо рад привели крају, јер је довољно да размотримо 8-му реч « égaux » (U → g et P → x). Даље се добија « distinctions sociales » (B → c). Последња етапа је:

les hommes naissent et demeurent libres et egaux en droits. les distinctions sociales ne FeuNent etre Tondees que sur l'utilite commune.

Елиминацијом изводимо да « Feuvent » одговара « peuvent » и « Tondees » за « fondees ».
Успели смо, дакле, и без познавања кључа, да извршимо криптоанализу овог текста, и поред милион милијарди милијарди могућих кључева!
Овај текст је први члан из декларације о праву човека (без акцената) : *les hommes naissent et demeurent libres et egaux en droits. Les distinctions sociales ne peuvent etre fondees que sur l'utilite commune.(људи се рађају и остају слободни и једнаких права.Социјалне разлике могу само да буду од заједничког интереса.*

Који је био кључ?

Када се криптоанализа оконча, врло је једноставно наћи кључ. Познајемо скоро све кореспонденције између две азбуке (реалне и шифриране).

Alphabet clair	a	b	c	d	e	f	g	h	I	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
Alphabet chiffré	L	I	B	E	R	T	U	V	W	?	?	Z	A	C	D	F	G	H	J	K	M	N	?	P	?	?

Неколико преосталих места у табели се врло ако попуњава.
Кључ је дакле: LIBERT (кључна реч је била LIBERTE).

Научна напомена:
За анализу фреквенције можемо користити и другу статистику. На пример:

- Фреквенција слова у функцији њихове позиције у речима. У енглеском језику речи најчешће почињу са словом « t », « a » или « s » као и са « e », иако је « e » најчешће употребљавано слово.
- Фреквенција парова слова или « bigrammes ». У француском су најфреквентнији bigrammes « es », « de », « le », « en » и « re ». Тражење « bigrammes » и « trigrammes » (најфреквентнији су « ent » и « les ») се намеће када текст не садржи размаке и интерпункцију при одвајању речи

(подсећање: у горњем примеру смо сачували размаке и интерпункцију због поједностављења проблема). Пажња: неки bigrammes се не поављују јер су између 2 речи!

Претрага литературе (појединачно)

Наставник расподељује Радни лист-B08 сваком ученику. Читају га појединачно а затим се дискутује у одељењу. Дискусијом се долази до следећег:

- Слабости неких метода шифрирања које, иако предлажу велики број кључева, омогућују да се шифра разбије уз мало вештине …
- Перманентна је веза између криптографија и криптоанализа.
- Актуелни утицај криптографије (о њему ће бити детаљније на крају пројекта).

Закључак

Одељење изводи заједнички закључак у следећем облику:

- Најважнији је алгоритам. Боље је имати добар алгоритам и осредњу машину него осредњи алгоритам а добру машину!
- Анализа фреквенције се реализује коришћењем статистике и познавањем језика (лингвистика)

Продубљивање (*Практичним интердисциплинарним подучавањем)*

У француском језику:

- Према « les hommes dansants (ИГРАЧИМА) », једној од 56 новела Arthur Conan Doyle постављен је на сцену детектив Sherlock Holmes. Криптоанализа је у срцу ове новеле (продубљење у математици је предложно ниже).
- Проучите неколико текстова и уочите како варирају хистограми зависно од типа језика (породични). Уочите да у неком атипичном тексту попут « La disparition » од G. Perec(*роман без најфреквентнијег слова « e » у француском језику,п.п.*), хистограми фреквенције су слични референтним хистограмима у француском језику, изузев што је слово « e » оно које је нестало.

У другим језицима:

- Упознајте неколико текстова и уочите да су фреквенције слова различите од оних у француском језику (ипак, слово « e » остаје изузетно доминантно и у другим језицима као што се може виддети на Радном листу-B07).

У историји:

- Упознајте услове који су омогућили успон арапске науке, па се често говори о « Исламском залтном добу »: политичка стабилност, економски просперитет, религиозна толеранција …

У математици:

- У циљу приказа могућности примене анализе фреквенције при било којој употреби симбола (а не само слова), направите криптоанализу « текста » сачињеног од низа цртежа. На пример, поруке које се појављују у новели « Les hommes dansants (ИРАЧИ)» (продубљивање у француском језику је претходно предложено) :

- Осим моно-алфабетског шифрирања можемо изучавати:
  Vigenère-ово шифрирање (шифрирање које користи више различитих азбука).
  Playfair-ово шифрирање (шифрирање које се реализује помоћу bigrammes, а не помоћу индивидуалних слова.