VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER SYSTEMS

TECHNIKY UMĚLÉ INTELIGENCE PRO DETEKCI SPAMŮ

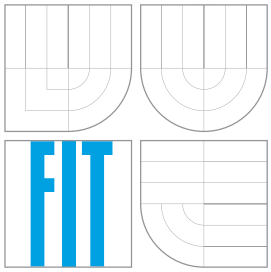
DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR

Bc. RADOVAN VRÁNSKY

BRNO 2013

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER SYSTEMS

TECHNIKY UMĚLÉ INTELIGENCE PRO DETEKCI SPAMŮ

ARTIFICIAL INTELLIGENCE APPROACHES FOR SPAM DETECTION

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR

Bc. RADOVAN VRÁNSKY

VEDOUCÍ PRÁCE
SUPERVISOR

Doc. Ing. SCHWARZ JOSEF, CSc.

BRNO 2013

Abstrakt

Táto diplomová práca sa zaoberá rôznymi metódami detekcie a rozpoznávania nevyžiadaných e-mailových správ. V úvode sú rôzne tieto metódy popísané. V ďalšej časti je podrobne popísaná Bayesova veta a metódy detekcie nevyžiadanej pošty s ňou pracujúce. V tejto časti je taktiež popísaný biologický a umelý imunitný systém a metódy využívajúce umelý imunitný systém v rámci odhaľovania nevyžiadanej pošty. Práca sa potom zaoberá podrobným popisom návrhu a implementácie vlastného systému na odhaľovanie nevyžiadanej pošty. Tento systém je v práci testovaný a v jej závere sú tieto testy zhodnotené.

Abstract

This thesis deals with various methods used for spam detection and identification. In the introduction various methods are described. Then Bayes' theorem and methods for spam detection that use this theorem are described in detail. This section also discusses biological and artificial immune systems and methods for spam detection based on artificial immune systems. Next sections contain the description of custom spam detection system design and implementation. Finally the system is tested and the results are evaluated.

Klíčová slova

Spam, ham, metódy boja proti spamu, Bayes, Umelé imunitné systémy.

Keywords

Spam, ham, anti spam methods, Bayes, Artificial Immune Systems.

Citace

Radovan Vránsky: Techniky umělé inteligence pro detekci spamů, diplomová práce, Brno, FIT VUT v Brně, 2013

Techniky umělé inteligence pro detekci spamů

Prohlášení

Prehlasujem, že som túto diplomovú prácu vypracoval samostatne pod vedením pána doc. Ing. Josefa Schwarze, CSc.

.....
Radovan Vránsky
17.05.2013

Poděkování

Rád by som poďakoval vedúcemu diplomovej práce pánu doc. Ing. Josefu Schwarzovi, CSc. za ochotu, čas a cenné rady, ktoré mi počas tvorby diplomovej práce venoval.

Obsah

Kapitola 1

Úvod

V súčasnej dobe prežívame rozmach komunikácie v rôznej elektronickej forme. Vďaka jej nespornej výhode, ktorou je dostupnosť a v podstate nulová cena, sa táto komunikácia dostáva k masám ľudí a do všetkých kútov sveta. S pribúdajúcim počtom užívateľov však začínajú pribúdať aj v tomto segmente počítačových technológií rôzne hrozby. Niektoré z týchto hrozieb zaniknú krátko po tom, ako sa objavili, no existuje hrozba, ktorá pretrváva už roky. Touto hrozbou je nevyžiadaná pošta *spam*. V dnešnej dobe je spam jeden z najväčších problémov rôznych ziskových, ale aj neziskových organizácií. Právom si môžeme položiť otázku, ako nám nevyžiadaná, no neškodná pošta môže spôsobovať nepríjemnosti. V tom lepšom prípade spam narúša pokoj pracovného dňa, keď nás núti tráviť čas jeho prehliadaním a mazaním. V tom horšom prípade môže byť prostriedok použitý k šíreniu rôznych vírusov či iných škodlivých programov vo vnútri firemnej siete. Čo môže mať za následok poškodenie užívateľských staníc ako aj firemných serverov. Toto nie je jediný pohľad na to, ako nám môže uškodiť, za iné uveďme napríklad vyťažovanie sieťovej prevádzky možnosť, že sa naša IP dostane na čiernu listinu, či krádež identity. Rôzne zdroje [6], [19] z oblasti expertov na spam či Anti-Spamových firiem v dnešnej dobe odhadujú, že nevyžiadaná pošta predstavuje 45 - 63% z celkového množstva elektronickej pošty doručenej cez internet. Na jednej strane dnes stojí prevencia, ktorá sa snaží spamerov od rozosielania nevyžiadanej pošty čo najviac odradiť. No na strane druhej máme rôznu kreativitu spamerov a ich motiváciu v podobe nemalých ziskov. Preto nie je možné predpokladať, že nevyžiadaná pošta bude niekedy úplne potlačená. Používaním rôznych techník, či už na strane klienta alebo na strane servera, je však možné do značnej miery nevyžiadanú poštu obmedziť. Rôzne aplikácie určené na boj proti nevyžiadanej pošte bývajú založené na jednej či na kombinácií rôznych metód určených k označeniu nevyžiadanej pošty. Na prvý pohľad by sa mohlo zdať, že tieto aplikácie sú navrhnuté, aby odvádzali rovnakú prácu. No ako je známe, keď dvaja robia to isté, nie je to vždy to isté. Preto si táto práca v prvej časti dáva za úlohu oboznámiť čitateľa s rôznymi viac či menej známymi technikami na odhaľovanie a potláčanie spamu. V jej druhej časti je možné nájsť popis a implementáciu vlastného navrhovaného systému. V závere nesmie chýbať otestovanie takto navrhnutého systému, porovnanie výsledkov a zhodnotenie odvedenej práce.

Kapitola 2

Metódy na boj proti Spamu

Nasledujúca kapitola, ktorá bola inšpirovaná knihou [1] a bolo v nej čerpané z materiálov [11], [16] sa snaží čitateľa prehľadovo oboznámiť s rôznymi technikami na boj proti nevyžiadanej pošte a taktiež ich rôznym delením. Pretože metódy na odhaľovanie nevyžiadanej pošty môžeme posudzovať podľa rôznych kritérií, delíme ich do rôznych skupín. Medzi najzákladnejšie druhy techník potom patria nasledujúce dve skupiny.

Metódy založené na strojovom učení:

- Bayesove metódy
- Neurónové siete
- Markovove modely
- Vyhľadávanie vzorov
- SVM

Metódy, ktoré nie sú založené na strojovom učení:

- Metódy založené na pravidlách
- Vyhľadávanie signatúr a hashov
- Rôzne Whitelisty, Blacklisty
- Metódy analyzujúce e-mailovú prevádzku

Ďalším z možných kritérií delenia metód odhaľovania nevyžiadanej pošty môže byť to, kde sa daný program či algoritmus nachádza. Najčastejšie to býva na strane klientskej alebo na strane serverovej, no čoraz častejšie sa objavujú rôzne metódy, ktoré sa zameriavajú na sieťovú prevádzku, analýzu protokolu či na návrh nových protokolov ako je napríklad Differentiated Mail Transfer Protocol(`DMTP`) [4].

2.1 Preventívne metódy

Mohlo by sa zdať, že tieto metódy je zbytočné tu spomínať. Práca je však o rôznych metódach boja proti nevyžiadanej pošte. A preventívne metódy zamerané na osvetu užívateľa sú mnohokrát účinnejšie ako rôzne iné počítačové metódy. A preto si niektoré z týchto metód v nasledujúcej kapitole popíšeme bližšie.

1. **Diskrétnosť** - Jej hlavným cieľom je dbať na zásadu nerozširovať zbytočne svoju adresu. Taktiež to znamená napríklad v prípade odosielania e-mailu skupine ľudí pridať týchto do e-mailu v skrytej kópii miesto v normálnej. Zachovávaním týchto zásad komplikujeme prácu programom na dolovanie spam listov.

2. **Address munging** - táto technika sa stáva dôležitou v prípade, kedy užívateľ potrebuje zverejniť svoju legitímnu e-mailovú adresu či už na stránkach alebo fórach. V podstate ide o jednoduchý zápis adresy do takého tvaru, že sa táto pre spam-bota stáva nečitateľná, no človek ju aj napriek tomu stále dokáže prečítať a správne interpretovať. Príkladom nech je: `radovan[bodka]vransky(zavinac)gmail[bodka]com` alebo `RadovanNoSp@M.gmail.fake.com` . Ďalšou možnosťou je využitie vlastnosti spam-botov proti ním. spam-bot vyhľadáva e-mailovú adresu na základe @ ktorý je pre ňu tak typický. Preto môžeme jednoducho zavináč nahradiť obrázkom. A väčšina spam-botov ju bude mať problém rozoznať.

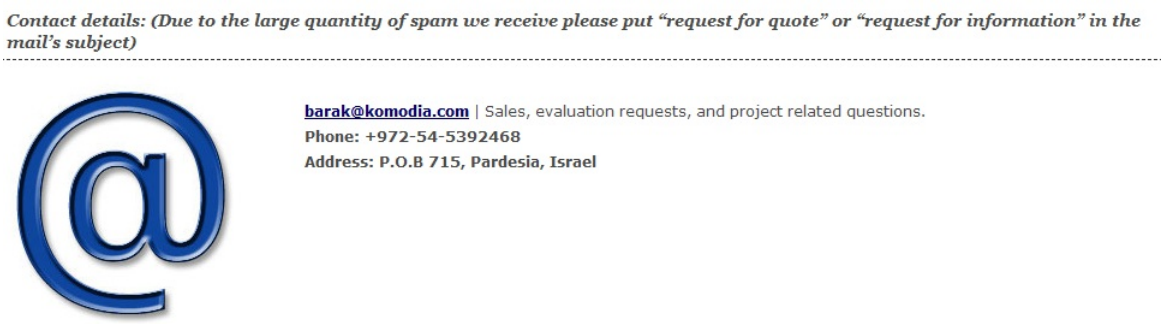
3. **Neodpovedať na spam** - nevyžiadaná pošta často obsahuje odkazy na rôzne zaujímavé stránky, či možnosť sa odhlásiť z odberu. V tomto prípade treba byť opatrný a zvážiť, či na takéto odkazy pôjdeme. Častokrát spameri tento spôsob využívajú na prečistenie zoznamov adries a je veľmi pravdepodobné, že po odpovedi, či návšteve odkazu bude účinok presne opačný a teda, že prílev nevyžiadanej pošty ešte narastie.

4. **Kontaktný formulár** - v dnešnej dobe nie je nič zvláštne, že veľa ľudí má vlastné stránky. Často je nutné, aby mali na stránkach spôsob ako sa na nich môžu návštevníci obrátiť. Umiestnenie e-mailovej adresy priamo na stránku nie je najvhodnejší spôsob. Preto je vhodnejšie mať na stránke kontaktný formulár, ktorý po vyplnení a potvrdení nám automaticky správu odošle na žiadanú adresu. Spojenie a samotné odosielanie e-mailu je odtienené serverom, a preto je obtiažne získať adresu majiteľa stránky.

5. **Jednorazové e-mailové adresy** - niekedy nie je potrebné zadávať svoju legitímnu e-mailovú adresu. Napríklad keď užívateľ nemá potrebu odoberať rôzne reklamné správy. No pri rôznych registráciách je e-mailová adresa podmienka. Riešením môže byť používanie dočasných e-mailov. Najčastejšie sú to schránky, vytvorené na veľmi krátku dobu určitú a po úspešnej registrácii zanikajú. Príkladom môže byť služba 10minutemail.com 2.1.

Obr. 2.1: Ukážka dočasne vygenerovanej e-mailovej adresy, zdroj [10].

6. HAM heslá pod týmto názvom je známa technika, ktorá vyžaduje od neznámeho užívateľa, aby do predmetu správy (najčastejšie) dal predom určený text na základe ktorého bude identifikovaný ako legitímny užívateľ 2.2.



Obr. 2.2: Ukážka použitia HAM hesla, zdroj [13].

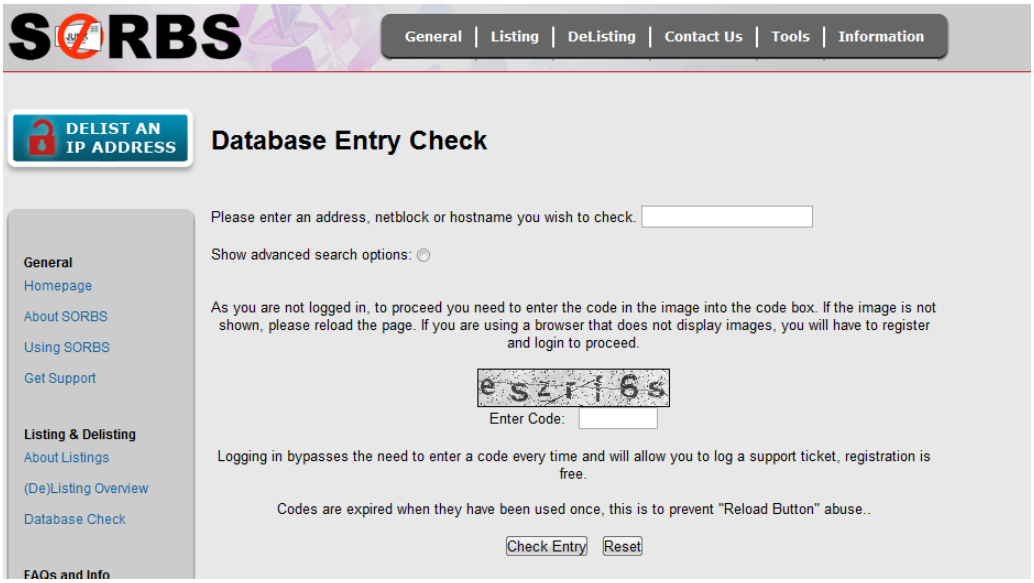
2.2 Metódy založené na zoznamoch

V tomto prípade ide o filtre založené prevažne na zoznamoch užívateľov, adries či iných identifikátoroch, na základe ktorých je možné určiť či príchodzia pošta je nevyžiadaná alebo legitímna.

1. **Čierna listina (Blacklist)** - v tomto prípade ide o veľmi populárny a v praxi aj veľmi využívaný spôsob odhaľovania nevyžiadanej pošty. Jeho snaha je zastaviť šírenie spamu na základe prednastavených zoznamov odosielateľov. Na týchto zoznamoch sa nachádzajú záznamy o e-mailových či IP adresách, ktoré boli v predošlej dobe zneužité na šírenie nevyžiadanej pošty. Postup určovania, či príchodzia správa spadá do kategórie nevyžiadanej pošty alebo je označená za poštu legitímnu je v princípe jednoduchý a je možné popísať nasledovne. Ak prichádza nová správa, spam filter skontroluje, či sa e-mailová alebo IP adresa odosielateľa nenachádza na čiernej listine. V prípade pozitívneho nálezu je takáto správa považovaná za spam a následne je odmietnutá. Aj keď je čierna listina veľmi účinným spôsobom, ako sa brániť voči nevyžiadanej pošte, môže na druhej strane spôsobiť značné nepríjemnosti nesprávnym označením legitímneho užívateľa ako spamera. K takýmto falošným poplachom (False Positives) najčastejšie dochádza v prípade, že spamer odosiela nevyžiadanú poštu z IP adresy, ktorá je využívaná aj legitímnym užívateľom e-mailu. Takáto situácia môže nastať aj v prípade, že spameri neustále menia svoje IP a e-mailové adresy tak, aby k ich odhaleniu pomocou techniky čiernej listiny došlo čo najneskôr.

2. **Real Time Blackhole list** - táto metóda na filtrovanie nevyžiadanej pošty funguje takmer identicky ako metóda čiernej listiny popísaná v bode vyššie. Značnou výhodou v prípade tejto metódy je, že z našej strany nemusí byť vynaložené nijaké značné úsilie na vytváranie zoznamu zneužívaných adries. Tieto zoznamy vytvárajú spravidla spoločnosti tretej strany, ktoré investujú čas a náklady do vybudovania takýchto komplexných čiernych listín. V tomto prípade je od nás vyžadovaná jediná činnosť a tou je napojiť daný spam filter na takúto listinu. Postup overovania či je príchodzia správa legitímna alebo spadá do kategórie nevyžiadaných správ je v podstate zhodný s klasickou technikou čiernej listiny. Ďalšou nespornou výhodou takto

budovaného zoznamu adries zneužívaných na rozosielanie nevyžiadanej pošty je to, že sú pomerne obsiahle a neustále udržiavané v aktuálnom stave. Takto sa podstatne zvyšuje pravdepodobnosť, že ak k nám dorazí nevyžiadaná správa, e-mailová či IP adresa odosielateľa sa už bude v zozname nachádzať.



Obr. 2.3: Ukážka formuláru, na odobratie e-mailovej adresy z čiernej listiny, zdroj [14].

Podobne ako v prípade čiernych listín aj tieto zoznamy môžu generovať falošné poplachy (False Positives). Problémom je, že tento zoznam odosielateľov nevyžiadanej pošty je spravovaný treťou stranou. Preto máme menšiu kontrolu nad adresami, ktoré zoznam obsahuje. Je možné, že sa na takomto zozname ocitneme neprávom, preto niektorí správcovia týchto zoznamov umožňujú vymazanie zo zoznamu na vlastnú žiadosť, napríklad formou formulára na stránkach zriaďovateľa takéhoto zoznamu 2.3.

3. **Biela listina (Whitelist)** - tento systém blokovania nevyžiadanej pošty je takmer presne opačný k čiernej listine. Filter miesto snahy určiť, ktorého užívateľa blokovať, určuje odosielateľa, od ktorého je možné správu prijať. Väčšina spam filtrov umožňuje využiť bielu listinu okrem iného aj ako spôsob, ktorým je možné výrazne znížiť počet korektných správ, ktoré sú označené ako pošta nevyžiadaná. Filtre využívajúce tento spôsob určovania nevyžiadanej pošty sú veľmi prísne. Správy od užívateľov, ktorí neboli dopredu schválení a umiestnení na túto listinu, budú automaticky označené za nevyžiadanú poštu a blokované.
   V prípade niektorých aplikácií bojujúcich proti nevyžiadanej pošte sa využíva iný variant bielej listiny, nazývaný automatická biela listina. V tomto prípade systém pracuje nasledujúcim spôsobom. Pri príchode e-mailovej správy sa jej užívateľ hľadá v zozname adries bielej listiny. Ak je zhoda negatívna, v klasickej metóde by bola správa označená ako spam, no v tomto prípade sa adresa odosielateľa overí oproti zoznamu adries zneužívaných k odosielaniu nevyžiadanej pošty. Ak sa adresa nenájde ani v tomto zozname, je správa označená ako legitímna a odosielateľ je pridaný do záznamov bielej listiny.

4. **Šedá listina (Greylist)** - v tomto prípade ide o pomerne novú techniku v oblasti boja s nevyžiadanou poštou. Jej hlavná myšlienka je tá, že mnoho spamerov sa pokúša odoslať, čo najväčšiu dávku nevyžiadanej pošty. Preto zvyčajne jednu správu na jednu adresu odosielajú len raz. Preto poštový server, ktorý využíva túto techniku, spočiatku novú príchodziu správu od neznámych odosielateľov odmieta, poznačí si ju do zoznamu a na server, z ktorého je správa odosielaná, pošle informáciu o zlyhaní prijímania tejto správy. V prípade, že sa poštový server odosielateľa pokúsi odoslať správu ešte raz, čo väčšina legitímnych serverov pracujúcich podľa noriem RFC urobí, predpokladá sa, že táto správa už nie je nevyžiadaná a pokračuje k príjemcovi. IP a e-mailová adresa odosielateľa je pridaná do zoznamu legitímnych odosielateľov a pri ďalších správach z tejto adresy už nie je takéto overovanie a spomaľovanie e-mailovej komunikácie potrebné.
   Aj napriek nesporným výhodám, ktoré túto techniku radia medzi v praxi veľmi využívané a úspešné, má aj táto technika svoje negatíva. Snáď tým najväčším je fakt, že dochádza k spomaľovaniu doručovania e-mailových správ k príjemcovi, čo môže byť veľmi nežiaduce v prípade časovo citlivých správ.

2.3 Metódy založené na obsahu

Techniky popísané v tejto kapitole sa nesnažia o presadzovanie celoplošného blokovania e-mailových správ, ktoré prichádzajú z konkrétnych e-mailových či IP adries, čo niekedy spôsobuje už vyššie spomínané problémy falošných poplachov. Naopak sa skôr zameriavajú na vyhodnocovanie slov, či fráz v obsahu každej e-mailovej správy jednotlivo. A na základe toho potom oddeliť poštu legitímnu od tej nevyžiadanej.

1. **Filtre na základe slov (Word-Based Filters)** - takto pracujúce filtre sú jedným z tých najjednoduchších typov filtrov z množiny filtrov pracujúcich na základe obsahu. V jednoduchosti by sa dalo povedať, že táto metóda zablokuje každú správu, ktorá spĺňa určité podmienky, alebo lepšie povedané slová či kombinácie slov. Pretože veľké percento nevyžiadanej pošty obsahuje slová a kombinácie slov, ktoré sa často v osobnej či obchodnej komunikácii nevyskytujú, môže byť filter založený na základe slov veľmi jednoduchá a účinná technika v boji proti nevyžiadanej pošte. Avšak problém nastáva v prípade, ak sú tieto filtre konfigurované aj na ďalšie bežné slová, môžu a budú tieto filtre generovať značný počet falošných poplachov a označovať tak legitímnu poštu za spam. Taktiež spameri veľmi často a zámerne robia chyby v kľúčových slovách správ, aby tak obchádzali filtre na základe slov. Dôsledkom čoho je nutnosť udržiavať zoznam blokovaných slov stále aktuálny a pomerne rozsiahly.

2. **Heuristické filtre (Heuristic filters)** - heuristické alebo filtre založené na pravidlách sú o krok ďalej ako filtre založené na slovách. Skôr ako na blokovaní správ, ktoré obsahujú podozrivé slovo či skupinu slov, sa heuristické filtre zameriavajú na niekoľko termínov nájdených v správe. Heuristické filtre prechádzajú obsah prichádzajúcich správ a priraďujú body určitým slovám alebo frázam. Podozrivé slová, ktoré sa bežne nachádzajú v nevyžiadanej pošte, ako sú napríklad slová `Rolex` či `Viagra`, dostanú vyššie bodové ohodnotenie. Naopak slová, ktoré sa často nachádzajú v bežnej e-mailovej komunikácii dostávajú nižšie bodové ohodnotenie. Filter nakoniec sčíta bodové ohodnotenie celej správy. Ak správa dosiahne, či prekročí určitý počet bodov, ktorý sa môže líšiť od filtra k filtru, je vyhodnotená ako nevyžiadaná pošta a je blokovaná. Ak tento počet bodov vo výsledku nedosiahne, ide o správu, ktorá je doručená

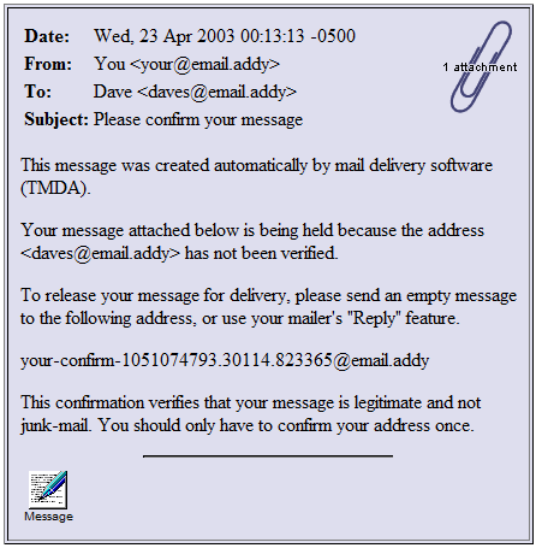
užívateľovi.
Heuristické filtre pracujú veľmi rýchlo a spôsobujú buď žiadne, alebo len minimálne zdržanie e-mailovej komunikácie, na rozdiel od napríklad Graylistu. Sú veľmi účinné a fungujú hneď po nasadení a nakonfigurovaní. Avšak v prípade nesprávnej konfigurácie môže dochádzať k veľkému generovaniu falošných poplachov. Toto sa stáva v prípade, ak legitímny odosielateľ v správe použije určitú kombináciu slov, ktorá je označená ako nevyžiadaná pošta. Naopak spameri sa naučili niektoré kombinácie slov nepoužívať či zamieňať tak, aby zabránili odhaleniu nevyžiadanej pošty.

3. **Bayesovské filtre (Bayesian filters)** - táto metóda sa v súčasnosti považuje za najvyspelejšiu formu filtrov založených na vyhodnocovaní obsahu e-mailových správ. Pri určovaní či je daná správa legitímna alebo nie, sa využívajú zákony matematickej pravdepodobnosti. Aby Bayesovský filter účinne pracoval, musí koncový užívateľ (užívatelia) filter naučiť, ktoré správy sú legitímne, a ktoré sú naopak nevyžiadané. Postupom času sa filter naučí, ktoré slová sa nachádzajú v bežnej komunikácii, a ktoré sú súčasťou nevyžiadanej pošty a pridá ich do zoznamu.
Ak chceme potom zistiť, či príchodzia pošta je klasifikovaná ako spam, Bayesov filter skontroluje obsah e-mailu a porovná text s týmito dvomi zoznamami na základe čoho vypočíta pravdepodobnosť, či e-mailová správa je nevyžiadaná. Napríklad, ak sa slovo "valiumöbjaví v zozname spamových správ 62 krát a v zozname legitímnych správ len 3 krát, je 95% pravdepodobnosť, že prichádzajúce e-maily obsahujúce slová "valium"sú spam. Vzhľadom k tomu, že Bayesov filter si neustále buduje svoje zoznamy slov na základe správ, ktoré užívateľ dostane, teoreticky sa počas doby prevádzky vylepšuje a stáva sa efektívnejší. Pretože táto metóda vyžaduje určité obdobie na učenie, je možné, že v začiatkoch bude nutné odstrániť niekoľko nevyžiadaných správ.

2.4 Ostatné metódy

Vyššie popisované metódy sú považované za pomerne účinné, jednoduché, a preto sú v značnej miere s väčšími či menšími obmenami využívané v praxi. Existuje však ešte pomerne veľké množstvo metód, ktoré nie sú tak rozšírené alebo nie sú tak známe. V nasledujúcej kapitole si niektoré zaujímavejšie z týchto metód aspoň v stručnosti popíšeme.

1. **Otázka odpoveď (Challenge/Response System)** - tieto filtre bojujú s nevyžiadanou poštou tak, že nútia odosielateľa, aby vykonal určitú úlohu skôr, ako bude samotná správa doručená adresátovi. Napríklad ak odošleme správu adresátovi, ktorý tento filter využíva, môže nám prísť obratom správa s odkazom na stránku, kde je nutné zadať zobrazený kód. Ak tento kód vyplníme správne, správa je doručená, ale ak sa nám úlohu nepodarí v určitom časovom limite splniť, náš e-mail bude vyhodnotený ako nevyžiadaná pošta a nebude adresátovi doručený. Príkladom takejto implementácie môže byť TMDA [15]. Ukážku výzvy tohto systému je možné vidieť na obrázku 2.4. Systém sa spolieha na to, že danú úlohu môže vždy vyriešiť len človek, no väčšina nevyžiadanej pošty je rozosielaná použitím automatizovaných programov. Značnou nevýhodou systému je fakt, že odosielateľ musí výzvu očakávať a túto výzvu vyriešiť. Čo je spojené s časovým oneskorením doručenia správy, či nutnej ďalšej interakcie užívateľa pri odosielaní pošty. Problém môže nastať, ak užívateľ na túto správu s výzvou nečaká, alebo ju nepochopí, vtedy samozrejme nastane prípad, kedy je legitímna správa zablokovaná ako spam.



Obr. 2.4: Ukážka správy od Challenge/Response systému, zdroj [15].

2. **Kolaboratívne filtre (Collaborative filters)** - táto metóda filtrovania je založená na zbere vstupovo od miliónov užívateľov e-mailových schránok po celom svete. Užívatelia týchto systémov môžu označiť prichádzajúce správy ako legitímne alebo naopak ako nevyžiadanú poštu. Správy označené ako spam sú hlásené do centrálnej databázy. Ak určitý počet užívateľov označí správu ako spam, filter ju automaticky zablokuje ako dosiahne schránky zvyšku užívateľov komunity. Výhodou takéhoto systému je, že do filtrovania nevyžiadanej pošty sa zapája obrovská užívateľská základňa. Takto je možné veľmi rýchlo potlačiť ohniská nevyžiadanej pošty, rádovo v priebehu niekoľkých minút. Jedinou nevýhodou systému je, že ak sa väčšia skupina spamerov mobilizuje a vydáva sa za užívateľov, ktorí označujú poštu ako legitímnu, môžu takto dosiahnuť, že spam je označený a doručený ako dôveryhodná pošta.

3. **Checksum-based filtering** - tento spôsob filtrovania nevyžiadanej pošty sa spolieha na skutočnosť, že nevyžiadaná pošta býva často doručovaná v zhlukoch a správy sú väčšinou rovnaké, až na niektoré malé rozdiely. Takýto filter potom odstráni všetky rozdielne prvky a zo správ následne vypočíta kontrolný súčet. Tento je porovnaný s databázou, ktorá obsahuje kontrolné súčty pre správy, ktoré už ako spam označené boli. Ak kontrolný súčet prijatej správy sedí s kontrolným súčtom objaveným v databáze, je príchodzia správa označená ako nevyžiadaná pošta.

4. **Filtrovanie na základe krajiny pôvodu** - v tomto prípade ide o principiálne veľmi jednoduchý spôsob filtrovania nevyžiadanej pošty. Pomocou IP adresy odosielateľa je možné jednoducho zistiť krajinu, z ktorej ten ktorý e-mail prichodí. A následne takto blokovať správy, ktoré prichádzajú napríklad z Ruska či Číny.

5. **Vyžadovanie RFC štandardov** - princíp, na ktorom stojí táto metóda je založený na fakte, že softvér využívaný k odosielaniu nevyžiadanej pošty je často veľmi jednoducho napísaný tak, aby dokázal e-maily odosielať. Často sa potom stáva, že takýto softvér nedodržiava štandardy RFC. Preto pri správnom sledovaní odchýlok od týchto štandardov je možné takýto softvér jednoducho odhaliť a zablokovať.

6. **Greeting delay** - je pomerne jednoduchá technika založená na vkladaní pauzy pred uvítaciu správu od servera. Podľa RFC 5321 (sekcia 3.2), musí klient čakať na túto správu pred začiatkom zasielania ďalších dát. Je pravidlom, že programy určené na rozosielanie nevyžiadanej pošty na túto správu nepočkajú a začnú dáta posielať. Tento stav je pomerne jednoduché zo strany servera identifikovať a takéto spojenie ukončiť.

7. **Nolisting** - každý legitímny SMTP server by mal mať niekoľko alternatívnych MX záznamov. Táto technika pracuje na princípe, že primárny MX záznam ukazuje na neexistujúci SMTP server. V prípade zasielania správy na takýto server, jej prvotné doručenie samozrejme zlyhá. No správne pracujúci a legitímny server odosielateľa sa pokúsi poslať správu na alternatívny server, čo už skončí úspechom. Dôležitým faktom je skutočnosť, že väčšina programov rozosielajúcich nevyžiadanú poštu, správu odošle na prvý záznam, čo sa im samozrejme nepodarí a tým ich snaha končí.

8. **SMTP callback verification** - veľké množstvá nevyžiadanej pošty v SMTP komunikácii často využívajú neexistujúcu From adresu. V prípade tejto techniky sa SMTP server príjemcu pokúsi nadviazať spojenie s odosielateľom. Ak sa toto spojenie nadviazať nedarí, server ho jednoducho odmietne.

9. **Spam-trapping** - je posledná metóda spomínaná v tejto kapitole. Taktiež sa jedná o principiálne veľmi jednoduchý spôsob boja proti nevyžiadanej pošte, kde správca systému alebo užívateľ vygeneruje e-mailovú správu, ktorá nie je bežne používaná. Táto správa je umiestnená na webové stránky tak, že k nej bežný užívateľ nemá prístup, no bot určený na zber adries ju jednoducho nájde. Následne všetky IP či e-mailové adresy, z ktorých na takúto adresu prišli správy, sú automaticky pridané na čiernu listinu a v budúcnosti blokované.

Kapitola 3

Bayesove metódy

Ako už bolo spomínané vyššie v texte, filtre založené na Bayesovej metóde sú veľmi jednoduché, no napriek tomu aj veľmi úspešné v boji proti nevyžiadanej pošte. V nasledujúcich kapitolách sa preto bližšie pozrieme na teoretické pozadie týchto metód. V ďalšom texte potom budú bližšie spomenuté metódy, ktoré tieto teoretické princípy používajú v praxi.

3.1 Bayesova veta

Thomas Bayes [12] v teórii pravdepodobnosti odhalil vzťah [5], [7] medzi pravdepodobnosťou nejakého javu a opačnou podmienenou pravdepodobnosťou. Pre formálny zápis podmienenej pravdepodobnosti javu A za predpokladu výskytu javu B používame zápis $P(A|B)$ a pre zápis opačne podmienenej pravdepodobnosti $P(B|A)$. Jednoduchú Bayesovu vetu môžeme potom zapísať ako 3.1.

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}, \tag{3.1}$$

kde $P(A)$, $P(B)$ sú pravdepodobnosti dvoch náhodných javov A a B, a platí $P(B) > 0$. Ak máme náhodné udalosti A_i, $i = 1, 2, ..., k$ pre ktoré platí:

- A_1, A_2,..., A_i sú navzájom disjunktné a nezlučiteľné,
- v každom pokuse nastáva práve jeden z nich, preto platí $\sum_{i=1}^{n} P(A_i) = 1$,
- $P(A_i) > 0$,
- B je jav s $P(B) > 0$,

pre každé i potom platí obecný vzťah 3.2.

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)} \tag{3.2}$$

3.2 Bayesovo filtrovanie

Bayesovo filtrovanie funguje na vyššie uvedenom princípe, no 3.1 a 3.2 sú nahradené alternatívnym zápisom Bayesovej vety pre jedno slovo 3.3 a pre množinu slov 3.4, ktoré sú pre prácu s e-mailovými správami vhodnejšie.

Spam		Ham	
slovo	počet výskytov	slovo	počet výskytov
viagra	5	dovolenka	10
rolex	12	rolex	2
výhra	18	deti	20

Tabuľka 3.1: Tabuľka počtu výskytu slov v *spame* a *hame*.

$$P(spam|slovo) = \frac{P(spam)P(slovo|spam)}{P(slovo|spam) \cdot P(spam) + P(slovo|ham) \cdot P(ham)}, \tag{3.3}$$

kde:

- $P(spam|slovo)$ je pravdepodobnosť, že správa je spam,
- $P(slovo|spam)$ je pravdepodobnosť, že dané slovo sa vyskytuje v správach typu spam
- $P(slovo|ham)$ je pravdepodobnosť, že dané slovo sa vyskytuje v správach typu ham,
- $P(spam)$ je celková pravdepodobnosť, že akákoľvek správa je spam, ,
- $P(ham)$ je celková pravdepodobnosť, že akákoľvek správa je ham,

$$p = \frac{p_1 p_2 \cdots p_n}{p_1 p_2 \cdots p_n + (1 - p_1)(1 - p_2) \cdots (1 - p_n)} \tag{3.4}$$

kde:

- p je pravdepodobnosť, že správa je spam,
- p_1 je pravdepodobnosť $p(spam|slovo_1)$, že prvé skúmané slovo sa vyskytuje v správach typu spam,
- p_2 je pravdepodobnosť $p(spam|slovo_2)$, že druhé skúmané slovo sa vyskytuje v správach typu spam
- p_n je pravdepodobnosť $p(spam|slovo_n)$, že n-té skúmané slovo sa vyskytuje v správach typu spam.

V praxi štandardne máme dve množiny e-mailových správ. Jedna množina je pošta nevyžiadaná (spam) a druhá množina je naopak pošta legitímna (ham). V prvom kroku sa obidve množiny správ spracujú tak, že sú rozdelené na jednotlivé slová, pričom je výhodou zanechať pôvodnú veľkosť písmen. Z týchto slov sa následne zostavia dve tabuľky 3.1, ktoré obsahujú jednotlivé slová a ich počet výskytov v správach. Z týchto dvoch tabuliek je potom nutné zostaviť jednu tabuľku 3.2, ktorá obsahuje slovo a pravdepodobnosť s akou sa nachádza v spame respektíve hame. Ako je z tejto tabuľky 3.2 možno vidieť, pravdepodobnosť $0,99$ znamená, že sa dané slovo vyskytuje len v pošte, ktorá je nevyžiadaná. Naopak slová s pravdepodobnosťou $0,01$ sa v nevyžiadanej pošte nenachádzajú vôbec. V prípade, keby mala pravdepodobnosť hodnotu $0,5$, znamenalo by to, že sa slovo vyskytovalo v oboch typoch správ v rovnakom počte. Slovo rolex s pravdepodobnosťou $0,86$ sa vo väčšine prípadov nachádzalo v pošte nevyžiadanej, no malo zastúpenie aj v pošte legitímnej. Keď máme

Pravdepodobnosť spam	
slovo	pravdepodobnosť
viagra	0,99
rolex	0,86
výhra	0,99
dovolenka	0,01
deti	0,01

Tabuľka 3.2: Tabuľka slov s pravdepodobnosťou výskytu v *spame*.

takto zostavené tabuľky, klasifikácia príchodzieho e-mailu je jednoduchá a pomerne rýchla. Text takéhoto e-mailu sa rozdelí na slová a v tabuľke sa vyhľadá pravdepodobnosť všetkých slov. Z tejto množiny pravdepodobností sa vyberie určitý počet slov, ktoré majú pravdepodobnosť čo najrôznejšiu od $0,5$. Pomocou takto získaných hodnôt a vzťahu 3.4 jednoducho vypočítame pravdepodobnosť, pomocou ktorej jednoducho určíme, či je príchodzia správa legitímna alebo nevyžiadaná.

Príklad:

Ak príchodzia správa obsahuje slová z tabuľky 3.2 a vyberieme nasledujúce tri: viagra $(0,99)$, výhra $(0,99)$ a deti $(0,01)$. Pravdepodobnosť, že daný e-mail je spam potom dosadením do vzťahu 3.4 jednoducho spočítame ako 3.5.

$$P(Spam) = \frac{(0.99 \cdot 0.99 \cdot 0.01)}{(0.99 \cdot 0.99 \cdot 0.01) + (1 - 0.99) \cdot (1 - 0.99) \cdot (1 - 0.01)} = 0,91 \tag{3.5}$$

Výsledkom je, že daná správa je s 91% pravdepodobnosťou pošta nevyžiadaná.

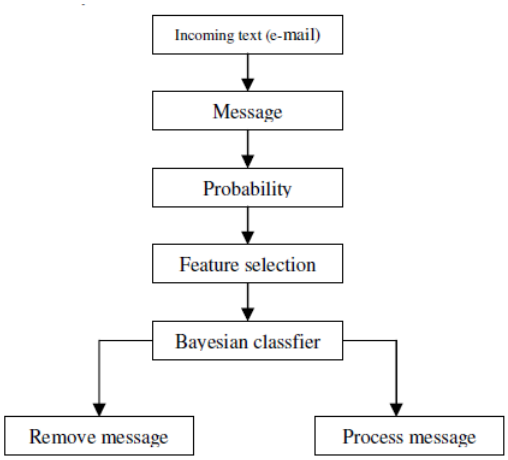
3.3 Metódy využívajúce Bayesovu vetu

Tento matematický princíp, popisovaný vyššie v práci sa v praxi využíva v rôznych metódach a spôsoboch boja proti nevyžiadanej pošte. Niektoré spôsoby ho viac, iné menej upravujú a prispôsobujú s cieľom dosiahnuť čo najlepšie výsledky. V nasledujúcej kapitole si preto z celého spektra takýchto metód niektoré zaujímavé spôsoby popíšeme a viac priblížime.

Bayesov filter spamu využívajúci štatistickú kompresiu dát

Autori článku [9] popisujú možnosť spolupráce v boji proti nevyžiadanej pošte v zmysle zdieľanej databáze spamu. Takáto databáza môže byť zostavovaná decentralizovane, prostredníctvom základne užívateľov e-mailových schránok u veľkých poskytovateľov ako je napríklad Google alebo Yahoo. Problémom u takto zostavovaných databáz je fakt, že obsahuje správy zozbierané od užívateľov, ktoré neraz obsahujú osobné či iné citlivé dáta. Preto užívateľom aj vzhľadom na prínos, takýto zber dát v zásade vadí. V tomto prípade by do úvahy prichádzalo vytváranie odtlačkov správ napríklad pomocou rôznych jednocestných funkcií. Bohužiaľ, ani toto riešenie nie je dokonalé v tom zmysle, že napríklad pri

generovaní hashu vzniká veľmi rozdielny odtlačok v prípade slov `viagra` a `vi@gr@`. Autori našli veľmi kvalitné riešenie vytvorením štatistického modelu ako náhrady za odtlačok e-mailových správ. Pričom predpokladajú, že každá správa x prichádzajúca od zdroja je prirodzene reprezentovatelná ako sekvencia $X = x_1 \dots x_n \in \Sigma$ symbolov nad abecedou Σ, pričom dĺžka $|x|$ môže byť ľubovoľná. Pričom cieľom akéhokoľvek štatistického algoritmu je nájsť takú funkciu zobrazenia $f : \Sigma^* \to [0, 1]$, ktorá odpovedá rozloženiu zdrojovej postupnosti tak presne, ako je to len možné. Ideálne je sekvencia x zakódovaná s $L(x)$ bitmi, kde $L(x) = -logf(x)$. Kompresný algoritmus sa teda musí naučiť čo najlepšie aproximovať správu. Čím lepšia aproximácia, tým kratšie bude výsledné kódové slovo. Tento fakt autorov motivuje k použitiu takéhoto algoritmu v spojení s kategorizáciou textu.
Samotné Bayesovo filtrovanie môže byť potom navrhnuté do niekoľkých modulov, tak ako znázorňuje obrázok 3.1. Príchodzia správa je v prvom rade rozdelená do rysov. Každému rysu je priradená pravdepodobnosť určujúca jeho príslušnosť k správam typu spam. K redukcii vektoru takýchto rysov sa použije algoritmus na výber najvhodnejších rysov, ktorý vytvorý výstupnú množinu rysov. Následne je k výpočtu pravdepodobnosti, že správa prísluší do triedy spam, použitý jednoduchý Bayesov klasifikátor.



Obr. 3.1: Schematické znázornenie spracovanie príchodzej správy, zdroj [9].

Nesporným prínosom popisovanej metódy je zvýšenie súkromia, databáza obsahuje len vybrané rysy. Na druhej strane sa môže ako nevýhoda javiť potreba učenia v prípade Bayesovho klasifikátora, no táto je kompenzovaná veľkou základňou učiteľov v podobe užívateľov.

Algoritmus filtrovania spamu založený na naivnom bayesovi a umelých imunitných systémoch

Ďalším prístupom, popisovaným v [8] je spojenie Bayesovho filtra s umelým imunitným systémom. Kde na jednej strane stojí úspešnosť a presnosť určovania nevyžiadanej pošty pomocou Bayesovej metódy. No na strane druhej stojí slabina tejto metódy, ktorou je nutnosť učenia a taktiež slabá schopnosť adaptovať sa. Tieto dva nedostatky autori článku s úspechom kompenzujú pomocou umelých imunitných systémov, ktoré sú v týchto vlastnostiach veľmi dobré. Základ takejto metódy, spájajúcej umelý imunitný systém s Bayesovou metódou, je možné popísať v nasledujúcich bodoch:

1. *Antigen/Antibody:* Antigény sú vektor rysov novej e-mailovej správy. Množina antigénov je viazaný zoznam. Samotný antigén môžeme definovať nasledovne:

```
typedef struct{
        char[ ] identifyString;
        int gfreqence;          ;počet výskytov antigenu v hame
        int sfrequence;         ;počet výskytov antigenu v spame
}Antigen;
```

Štruktúra protilátky potom môže vyzerať:

```
typedef struct{
        char [ ] identifyString;
        int gfrequence;         ;počet výskytov protilátky v hame
        int sfrequence;         ;počet výskytov protilátky v spame
        float fw;               ;pravdepodovnosť výskytu protilátky v spame
        float pw;               ;pravdepodovnosť výskytu protilátky v hame
        int life;               ;životnosť protilátky
        int numOfCapture;
        int numOfSuccess;
}Antibody;
```

Na základe obyčajného Baesovho algoritmu sú podľa vzťahov 3.6 a 3.7 aktualizované hodnoty `pw` a `fw`, kde N_S, N_H je počet spam, ham správ a N celkový počet správ.

$$pw = \frac{1 + gfrequence}{2 + N_H} \tag{3.6}$$

$$fw = \frac{1 + sfrequence}{2 + N_S} \tag{3.7}$$

2. *Určenie triedy:* určenie, či je príchodzia správa spam alebo naopak ham sa určuje pomocou algoritmu, ktorý vracia hodnotu 0 a 1. Správa je spam ak jeho p_{spam} je väčšia ako stanovený prah 3.8.

$$p = \begin{cases} 1 & p_{spam} > treshold \\ 0 & p_{spam} < treshold \end{cases} \tag{3.8}$$

Pomocou základného Bayesovho algoritmu je p_{spam} možné vypočítať ako 3.9, pričom $p(w_i|c = 1)$ je `fw` a $p(w_i|c = 0)$ je `pw`.

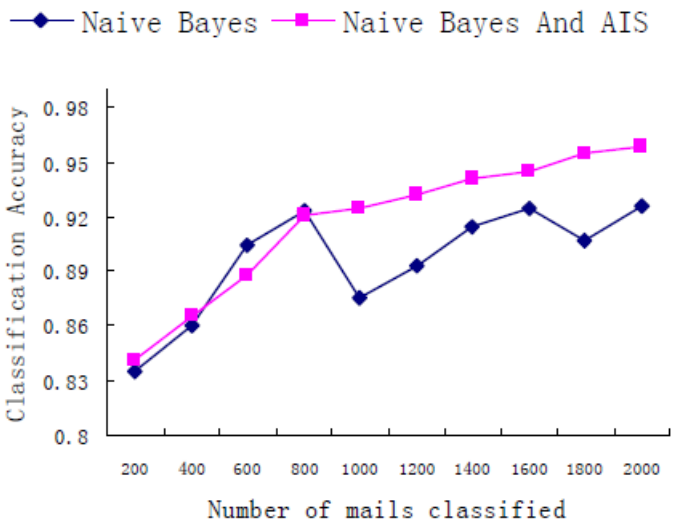
$$p_{spam} = \frac{\frac{N_S}{N} \times \prod_{i=1}^{n} p(w_i|c = 1)}{\frac{N_S}{N} \times \prod_{i=1}^{n} p(w_i|c = 1) + \frac{N_H}{N} \times \prod_{i=1}^{n} p(w_i|c = 0)} \tag{3.9}$$

V jednoduchosti by sa celý postup spracovania príchodzej správy dal popísať v nasledujúcich krokoch.

1. príchod novej e-mailovej správy

2. e-mail vyprodukuje množinu antigénov
3. tieto antigény sú pomocou pamäťových buniek rozpoznané a v prípade zhody je správa označená ako nevyžiadaná
4. v opačnom prípade je vypočítaná podobnosť antigénov a protilátok v nezrelých bunkách
5. v prípade, že je podobnosť vyššia ako stanovený prah, správa je označená ako nevyžiadaná. Inak je označená ako legitímna.
6. na základe spätnej väzby od užívateľov sú pamäťové a nedozreté bunky aktualizované pomocou antigénov

Následné testovanie algoritmu a jeho porovnanie s filtrom, ktorý využíval len Bayesovu metódu ukazuje, že algoritmus s použitím umelých imunitných systémov dosahuje lepšie výsledky. Pre úplnosť treba dodať, že v oboch prípadoch bola použitá rovnaká trénovacia a testovacia množina e-mailov. Ako znázorňuje graf 3.2 pri vyššom počte e-mailov má tento algoritmus vyššiu presnosť a lepšie reaguje na náhle zmeny e-mailov. Z čoho je možné usúdiť, že takýmto spojením Bayesovského filtra s UIS vznikne filter s vyššou robustnosťou.



Obr. 3.2: Graf porovnávajúci dva filtre, zobrazený je počet klasifikovaných správ a presnosť klasifikácie, zdroj [8].

Vylepšený bayesov algoritmus na filtrovanie nevyžiadanej pošty

Autori článku [20] vychádzajú zo základnej Bayesovej vety 3.2, ktorú však pre potreby práce s e-mailovými správami upravujú. Za predpokladu, že máme množinu e-mailových správ $D = \{d_1, d_2, \ldots, d_n\}$, množinu slov $W = \{w_1, w_2, \ldots, w_n\}$ a triedu príslušnosti $C = \{c_1, c_1, \ldots, c_n\}$. Potom $d_i = \{val(w_1), val(w_2), \ldots, val(w_m)\}$ pričom ak je $val(w_i) = 1$ znamená to, že w_i existuje v d_i. S prihliadnutím na Bayesovu vetu, má potom výsledný vzorec tvar 3.10.

$$P(c_i|d_i) = \frac{P(c_i) \prod\limits_{w_i \in d_i} P(w_i|c_i)}{P(d_i)} \tag{3.10}$$

Proces samotného učenia klasifikátor potom pozostáva z nasledujúcich bodov.

- *Attitude Analysis Phase* - pri tejto fáze sa na prichádzajúcej správe určí na základe subjektu, odosielateľa, či iných atribútov vážené skóre, s váhou $0,25$ pre každý atribút.
- *Preprocessing* - v tomto momente sú zo správy odstránené všetky nadbytočné prvky ako napríklad HTML elementy. A celá správa je spracovaná do štrukturovanej formy.
- *Tokenization* - správa je rozdelená na jednotlivé slová. Každé slovo je ešte v tomto kroku porovnávané s pozitívnym slovníkom, aby bola určená relevancia obsahu. V prípade, že sa slovo v tomto slovníku nachádza, pokračuje sa ďalším krokom. V opačnom prípade sa slovo porovnáva s negatívnym slovníkom.
- *Post Processing and Final decision making* - konečné rozhodnutie sa určuje pomocou váženého skóre. Ak je skóre menšie ako $0,5$ správa je označená ako spam a vybrané kľúčové slová z nej sú pridané do negatívneho slovníka. V opačnom prípade je správa označená ako ham a kľúčové slová sú priradené do pozitívneho slovníka. Tým, že je vyhodnocovanie rozdelené na dve fázy Attitude Analysis Phase a porovnávanie so slovníkom, sa znižuje možnosť označenia validnej správy za spam.

V praxi sa často stávalo, že legitímna pošta bola označovaná ako spam. Preto sa autori článku rozhodli použiť dva druhy prahov. Za predpokladu, že D je nevyžiadaná pošta s pravdepodobnosťou $P(C1|D)$ a s pravdepodobnosťou $P(C0|D) = 1 - P(C1|D)$ správa legitímna, budú tieto prahy vyzerať nasledovne.

1. *Kritická pravdepodobnosť* t, kde ak $P(C1|D) > t$, potom je správa nevyžiadaná.
2. *Kritický pomer* k, kde ak $(P(C1|D)/P(C0|D)) > k$, potom je správa nevyžiadaná.

Štúdia detekcie SMS spamu na základe výberu rysov kľúčových slov

Posledná popisovaná metóda [3] sa zameriava na použitie bayesovho filtra v SMS. Určité postupy z článku však môžu byť s úspechom použité aj v rámci e-mailovej komunikácie. Autori si c určili ako kategóriu správy, počet slov w_i nachádzajúcich sa v správe c ako N_i a celkový počet slov v c ako n. Potom pravdepodobnosť, že slovo w_i patrí do c počítali podľa vzťahu 3.11.

$$P(w_i) = \frac{N_i}{\sum_{i=1}^{n} N_i} \tag{3.11}$$

Potom mohli pomocou vzťahu 3.12 vypočítať pravdepodobnosť $P(c_2|w_i)$, že text prináleží správe typu spam.

$$P(c_2|w_i) = \frac{P(w_i|c_2)}{P(w_i|c_2) + P(w_i|c_1)} \tag{3.12}$$

Keď sa však slovo w_i vyskytuje len v množine správ spam, pravdepobnosť podľa 3.11 je rovná 0, potom $P(c_2|w_i) = 1$. Na základe tejto podmienky slovo w_i rozhodne, príslušnosť správy k triede bez ohľadu na iné slová. Tento jav sa nazýva *unseen feature words problem.* Autori sa preto rozhodli aplikovať na Bayesov algoritmus zjemňujúci algoritmus výsledkom čoho je vzťah 3.13.

$$\hat{P}(w_i|c_j) = \frac{\lambda + N_{ij}}{\lambda \cdot n_j + \sum_{i=1}^{n} N_{ij}} \tag{3.13}$$

Vo vzťahu 3.13, $\hat{P}(w_i|c_j)$ udáva predchádzajúcu pravdepodobnosť, že slovo w_i existuje v kategórii c_j. N_{ij} značí počet výskytov w_i v triede c_j, zatiaľ čo n_j je počet slov v c_j. λ je

zjemňujúci koeficient a jeho hodnota sa pohybuje v rozmedzí 0 až 1.
V článku tiež do Bayesovho algoritmu začlenili váhu jednotlivých slov. Vzťah na určovanie pravdepodobnosti či je správa spam alebo nie, ktorý zahŕňa váhu každého slova potom vyzerá nasledovne 3.14.

$$P(M|w_1, w_2, \ldots, w_n) = \frac{\prod_{i=1}^{n} P_i^{weight(w_i)}}{\prod_{i=1}^{n} P_i^{weight(w_i} + \prod_{i=1}^{n}(1 - P_i)^{weight(w_i)}} \tag{3.14}$$

Vo vzťahu 3.14 je $weight(w_i)$ váha slova w_i, ktorá je často počítaná ako frekvencia výskytu slova v správe. No autori vychádzali z toho, že vo väčšine prípadov sú dlhšie slová dôležitejšie ako tie krátke. Skombinovali preto dĺžku slov s frekvenciou ich výskytu a tak určili pre každé slovo váhu 3.15.

$$weight(w_i) = frequency(w_i) \times length(w_i) \tag{3.15}$$

Kapitola 4

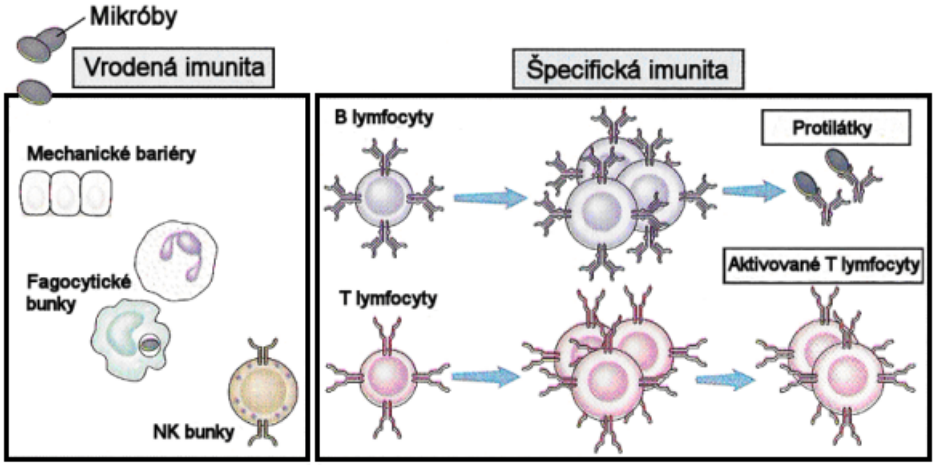
Imunitné systémy

4.1 Biologický imunitný systém

Biologický imunitný systém, ktorého úlohou je zabezpečiť obranu pred cudzími látkami ako sú napríklad baktérie, plesne či vírusy, je jedným z najdôležitejších a zároveň aj najzložitejších systémov v ľudskom organizme. Sám dokáže rozpoznávať a reagovať na látky, s ktorými sa ešte nestretol a vďaka pamäti sa dokáže rýchlejšie vysporiadať s votrelcom, ktorého už pozná. V nasledujúcom texte je čerpané z [2] a [17].

4.2 Rozdelenie imunitného systému

V prípade, že sa budeme bližšie zaoberať biologickým imunitným systémom zistíme, že sa delí na dve časti. Jednou z nich je imunita nešpecifická, známa skôr pod pojmom vrodená imunita. Druhou je potom imunita špecifická známejšia ako imunita získaná 4.1.



Obr. 4.1: Základné rozdelenie biologického imunitného systému. Zdroj [17].

Vrodená imunita

Vrodená (nešpecifická) imunita je evolučne staršia ako imunita získaná, v určitej forme sa vyskytuje u všetkých mnohobunkových živočíchov. Je založená na receptoroch, ktoré sú

schopné rozpoznávať molekulárne vzory spojené s mikrobiologickými patogénmi, ktoré sa nikdy nevyskytujú vo vlastnom organizme. Predstavuje prvú líniu obrany organizmu a na útočníka je schopná zareagovať v priebehu niekoľkých minút. Vrodená imunita zahŕňa fyzikálne (neporušený povrch kože, pohyby rias, chĺpky v nose) a chemické (mastné kyseliny na pokožke, enzýmy v slinách, slzách či pote, alebo kyslé pH v žalúdku a moči) bariéry, fagotické bunky, NK (Natural Killer) bunky, proteíny cytokíny, ktoré sú zodpovedné za reguláciu a koordináciu buniek vrodenej imunity. Patrí sem aj komplementárna kaskádová reakcia, čo je spôsob eliminácie vírusov a baktérií prostredníctvom reťazca aktivovaných bielkovín nachádzajúcich sa v krvi a tiež interferóny. Sú to malé molekuly bielkovín produkované bunkou infikovanou vírusom. Majú schopnosť chrániť iné bunky tela tým, že v nich aktivujú produkciu enzýmov blokujúcich množenie sa vírusu. Interferóny takto tvoria prvú blokujúcu reakciu. Taktiež stimulujú lymfocyty, ktoré ničia bunky infikované vírusom a napríklad aj rakovinové bunky.

Vrodená imunita je taktiež zodpovedná za vyvolanie signálov v APC bunkách (Antigen Presenting Cells), ktoré spôsobujú aktiváciu T-lymfocytov a tým aj aktiváciu špecifickej imunity.

Získaná imunita

Získaná (špecifická) imunita je pomalšia a oproti vrodenej imunite má neskorší nástup, no vďaka reakcii na konkrétny typ antigénu a schopnosti vytvárať pamäťové bunky je tento systém efektívnejší. Získaná imunita nastupuje po aktivácii vrodenej imunity. Nazýva sa špecifická, lebo bunky a mechanizmy patriace do tejto skupiny reagujú vždy na určitý konkrétny antigén, ktorý sa musia v prvom rade naučiť rozoznávať. V prípade špecifickej imunity sú využívané organizmom produkované antigénové receptory, ktoré sú ďalej distribuované do organizmu, pomocou B a T lymfocytov. Tieto receptory sú generované pomocou náhodných procesov ako je napr. mutácia. Týmto je zabezpečená obranyschopnosť organizmu aj pred mikroorganizmami, s ktorými nikdy predtým neprišiel do styku.

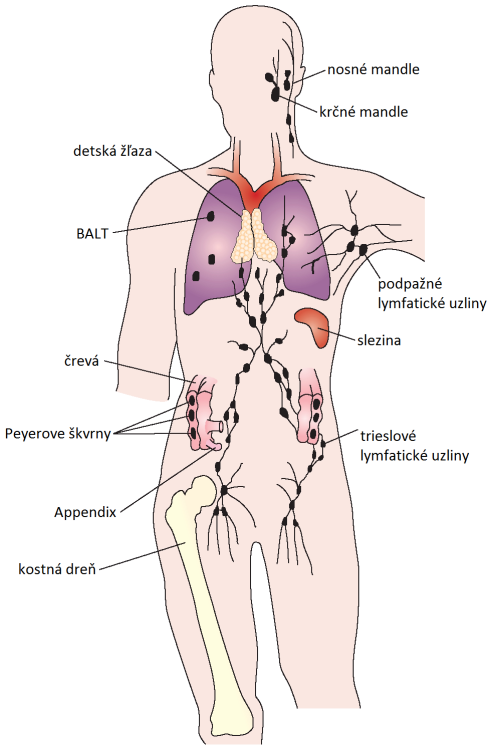
4.3 Komponenty biologického imunitného systému

Imunitný systém môžeme charakterizovať ako systém decentralizovaný. Jeho orgány, tkanivá a bunky sú rozložené po celom organizme. Spolu ich nazývame lymfatické orgány. Sú to miesta, kde lymfocyty interagujú s dôležitými ne-lymfatickými bunkami, či už pri procese dozrievania, alebo pri imunitnej reakcii. Lymfatické orgány môžeme rozdeliť na primárne, zodpovedné za produkciu lymfocytov a sekundárne, kde nastáva samotná imunitná reakcia.

Lymfatická sústava

Lymfatická sústava je neoddeliteľnou súčasťou imunitného systému, preto je dobré si aspoň v stručnosti popísať jej základné prvky 4.2.

- *krčné, nosné mandle*: sú párový orgán, ktorý je prvým miestom, kde sa zachytávajú antigény vnikajúce do organizmu dutinou ústnou a nosnou. Často sú preto prvým miestom množenia sa baktérií.
- *detská žľaza*: je orgán lymfatického systému, ktorý sa postupne mení na nefunkčné tkanivo. Detská žľaza plní hlavnú úlohu pri dozrievaní a novotvorbe T-lymfocytov.



Obr. 4.2: Rozmiestnenie lymfatických orgánov. Zdroj [18].

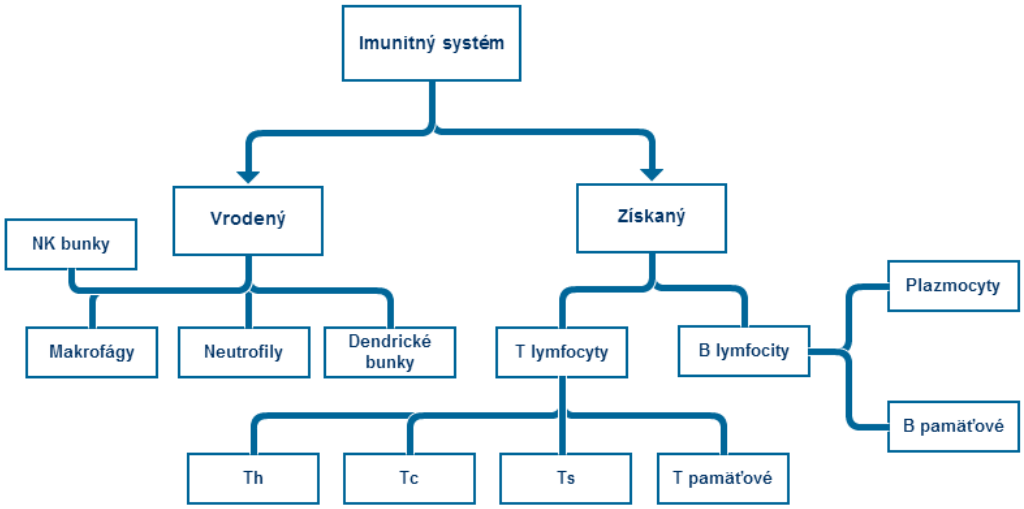
- *BALT(Bronchus-Associated Lymphoid Lissue)*: sú sliznice imunitného systému, konkrétne sliznice v dýchacej sústave. Tieto sliznice obsahujú veľké množstvo pamäťových buniek, aktivovaných T-lymfocytov či makrofágov.
- *črevo*: je zodpovedné nielen za trávenie potravy, ale rovnako aj za obranu proti vonkajším vplyvom. Obrannú funkciu tvorí črevná mikroflóra, črevná sliznica.
- *Peyerove škvrny(plaky)*: sú útvary obsahujúce nahromadenú lymfatickú tkaninu v sliznici tenkého čreva.
- *Appendix*: je výbežok slepého čreva. Jeho stena obsahuje lymfatické tkanivo.
- *kostná dreň*: je mäkké tkanivo v dlhých kostiach zodpovedné za produkciu lymfocytov.
- *podpažné, trieslové lymfatické uzliny*: tvoria spojenia lymfatických ciev, sú to miesta kde nastáva špecifická imunitná reakcia.
- *slezina*: funguje podobne ako lymfatické uzliny, na rozdiel od nich ale kontroluje prítomnosť antigénov priamo v krvi.

4.4 Bunky imunitného systému

Imunitný systém je tvorený zložitým systémom orgánov, buniek a molekúl. Imunitných buniek, ktoré tvoria základ imunitnej obrany, sa v ľudskom organizme nachádza okolo 10^{12}. Každý deň je veľké množstvo týchto buniek nahradených novými. Ich konkrétne počty sa ale

líšia aj od stavu, v ktorom sa organizmus nachádza. Napríklad v prípade infekcie môže počet leukocytov narásť aj niekoľkonásobne. Môžeme ich nájsť roztrúsené po celom organizme - nachádzajú sa v krvi, lymfe, ale aj v jednotlivých orgánoch a tkanivách.

Existuje veľké množstvo rôznych druhov imunitných buniek 4.3, preto v tejto časti budú popísané len tie najdôležitejšie.



Obr. 4.3: Rozdelenie buniek imunitného systému.

Neutrofily, makrofágy a príbuzné bunky

Do tejto skupiny buniek patria takzvané fagocytycké bunky (doslova pojedači buniek), čo sú vlastne biele krvinky, schopné pojedať iné bunky. Patria k nešpecifickej imunite a sú to najpočetnejšie lymfocyty.

Neutrofily tvoria asi 60 až 70% celkového množstva bielych krviniek v organizme a sú tak základnou silou nešpecifickej imunity. Spolu s makrofágmi vznikajú v kostnej dreni. Sú to asi $12 - 15\mu m$ veľké guľovité bunky. V tele dospelého človeka ich denne vzniká okolo 10^{11}, pričom majú veľmi krátku životnosť - v priemere okolo 6 hodín. Ak sa v tomto čase nestretne neutrofil s antigénom, umiera. V lymfatických uzlinách sa môžu vytvárať zásoby neutrofilov, ktoré sa v prípade infekcie dajú veľmi rýchlo použiť.

Makrofágy sú evolučne najstarší sprostredkovatelia vrodenej imunitnej reakcie. Aj keď tvoria iba $5 - 10\%$ populácie bielych krviniek, hrajú centrálnu úlohu nielen vrodenej ale aj špecifickej imunity. Po vzniku prechádzajú niekoľkými štádiami a na rozdiel od neutrofilov sú pri infekcii schopné sa deliť. Majú tiež vyššiu životnosť a keďže ich odpoveď na infekciu je pomalšia, dominujú v neskorších fázach vrodenej imunitnej reakcie.

Ďalším rozdielom medzi neutrofilmi a makrofágmi je v schopnosti prezentácie antigénu. Makrofágy (a tiež dendritické bunky), na rozdiel od neutrofilov, majú schopnosť na svojom povrchu prezentovať antigény. Pohltený antigén upravia a takto upravený antigén vystavia na svojom povrchu v spojení s MHC molekulami. Lymfocyty s touto schopnosťou sa označujú ako APC bunky (antigen presenting cell - bunky prezentujúce antigén). T lymfocyty totiž nie sú schopné rozpoznať antigén priamo, ale len vo väzbe s MHC molekulami.

Dendritické bunky sa vyskytujú najmä v slezine a lymfatických uzlinách. Aj keď nemajú schopnosť fagocytózy, sú príbuzné makrofágov. Ich úlohou je prezentácia antigénu T lymfocytom a predstavujú najúčinnejšie APC bunky. Časť dendritických buniek, nazývajú

sa Langerhansove bunky, sa nachádza v koži, odkiaľ transportujú antigény do lymfatických uzlín a tam spúšťajú imunitnú reakciu.

NK bunky

NK bunky sa radia medzi lymfocyty, ktoré sú schopné rýchlo ničiť niektoré nádorové a vírusmi napadnuté bunky. Aj keď je ich funkcia podobná ako funkcia Tc lymfocytov, na rozdiel od nich NK bunky nepotrebujú pre svoju činnosť stimuláciu a aktiváciu od konkrétneho antigénu a Th lymfocytov. Na rozpoznanie toho, či má NK bunka zabíjať sa využívajú MHC I receptory, ktoré slúžia na identifikáciu buniek. Niektoré nádorové a vírusové bunky sa bránia proti Tc bunkám tým, že na svojom povrchu majú vystavenú buď nekompletnú, alebo žiadnu MHC I molekulu. NK bunky práve tieto bunky vyhľadávajú a ničia. Na zabíjanie používajú (podobne ako Tc lymfocyty) granule s cytotoxickými látkami.

NK bunky sa vyskytujú hlavne v krvi a slezine, kde tvoria asi 10% populácie všetkých lymfocytov.

B lymfocyty

Špecifickú imunitnú odpoveď zabezpečujú dva rozdielne typy lymfocytov - lymfocyty T a B. B lymfocyty zodpovedajú za protilátkovú odpoveď, po interakcii s antigénom sa diferencujú na plazmatické bunky, ktoré vo veľkom množstve produkujú protilátky.

B lymfocyty vznikajú v kostnej dreni. Ich vývoj možno rozdeliť na dve štádiá - na štádium nezávislé a na štádium závislé od prítomnosti antigénu v organizme. Štádium nezávislé od antigénu sa odohráva v kostnej dreni a predstavuje dozrievanie nezrelých B lymfocytov na zrelé, plne funkčné bunky. Odhaduje sa, že každý deň sa vyprodukuje približne $5 \cdot 10^7$ nových B lymfocytov, pričom do obehu sa dostáva len niečo okolo 10% z nich. Zbytok je eliminovaný negatívnou a pozitívnou selekciou, pričom väčšinou ide o lymfocyty ktoré sú citlivé na vlastné bunky.

Zrelé B lymfocyty, označované aj ako naivné, opúšťajú kostnú dreň, krvou a lymfou sa dostávajú do sekundárnych lymfatických orgánov, najmä sleziny a lymfatických uzlín. Ak sa B lymfocyt stretne s antigénom, pre ktorý nesie špecifický receptor, aktivuje sa a začne sa diferencovať na plazmatické bunky. Ich úlohou je samotná produkcia protilátok. Súčasne sa diferencuje aj na pamäťové bunky. Tie neprodukujú protilátky, ale majú predĺženú životnosť a imunitnému systému umožňujú rýchlejšiu reakciu na opakovanú infekciu rovnakým typom antigénu. Diferenciácia B lymfocytov na plazmocyty či pamäťové B lymfocyty je úplne závislá od prítomnosti antigénu v organizme, preto toto obdobie označujeme ako štádium závislé od antigénu. Ak sa naivný B lymfocyt so svojím antigénom nestretne, hynie, približne za 4 až 8 týždňov.

T lymfocyty

Podobne ako B lymfocyty, aj T lymfocyty vznikajú v kostnej dreni. Na rozdiel od nich, ale dozrievajú v detskej žľaze. Medzi ich funkcie patrí regulácia aktivity iných buniek a tiež zabíjanie infikovaných buniek. T lymfocyty môžeme rozdeliť na tri hlavné skupiny: T pomocné bunky (Th - helper), T cytotoxické bunky (Tc) a T supresorové, alebo regulačné bunky (Ts). Okrem nich môžu, podobne ako B lymfocyty, vytvárať T lymfocyty aj pamäťové bunky.

Na rozdiel od B lymfocytov T lymfocyty nevedia rozpoznať antigén priamo, ale len v podobe MHC II, ktorý je prezentovaný APC bunkami.

T pomocné bunky sú nevyhnutné na aktiváciu B lymfocytov, ostatných T lymfocytov, makrofágov a NK buniek. Th takto tvoria základ celého imunitného systému. Pri ich poškodení alebo zničení, napr. vírusom HIV, sa celý imunitný systém zrúti.

T cytotoxické bunky sú schopné eliminovať vírusy, nádorové bunky a iné typy útočníkov. Po aktivácii sa napoja na nepriateľskú bunku a pomocou cytotoxínov ju zničia.

Imunitný systém nás chráni pred množstvom potenciálne patogénnych mikroorganizmov, zatiaľ čo sa vyhýba reakcie s vlastnými zložkami tela. Zlyhanie tejto tolerancie vedie k vývoju autoimúnnych chorôb, ktoré postihujú približne 5% populácie. T supresorové lymfocyty majú za úlohu potláčať takéto nežiadúce reakcie imunitného systému. Ich úlohou je brzdiť činnosť ostatných buniek zodpovedných za imunitnú reakciu a takto chrániť organizmus pred alergickými reakciami a autoimúnnymi chorobami

4.5 Imunitná reakcia

Úplne prvou ochrannou bariérou medzi okolitým svetom a ľudským organizmom sú koža, tráviaci a dýchací systém. Tiež sú to cesty, ktorými do nášho organizmu najčastejšie vstupujú škodlivé cudzie látky. V prípade, že sa im podarí prekonať tieto primárne bariéry, ako ďalšia prekážka v ich ceste sú mechanizmy vrodenej imunity. Na imunitnej reakcii sa ako prvé začínajú podieľať neutrofily a makrofágy, ktoré začnú s elimináciou patogénov v priebehu niekoľkých hodín.

Okrem toho má vrodená imunita na starosti aj šírenie signálov (pomocou APC buniek), ktoré spúšťajú mechanizmy špecifickej obrany. V rámci vrodenej imunity existujú ešte aj ďalšie mechanizmy, ako je napríklad činnosť NK buniek, komplementárna kaskádová reakcia (eliminácia patogénov pomocou reťazca bielkovín nachádzajúcich sa v krvi), alebo činnosť interferónov. Interferóny sú molekuly bielkovín produkované bunkou infikovanou vírusom. Majú schopnosť chrániť iné bunky organizmu tým, že v nich aktivujú produkciu enzýmov blokujúcich množenie sa vírusu. Ich ďalšou funkciou je stimulácia lymfocytov.

Špecifická imunitná odpoveď je zahájená v lymfatických uzlinách. Patogény, ktoré vstupujú do organizmu napr. cez kožu, sú odchytené APC bunkami, ktoré ich spracujú a vystavia pomocou MHC komplexu na svojom povrchu. APC bunky potom takto upravené antigény transportujú do lymfatických uzlín.

Aby T a B lymfocyty mohli začať plniť svoju funkciu, je potrebné aby sa aktivovali. B lymfocyty sa môžu aktivovať aj priamo kontaktom s antigénom, T lymfocyty potrebujú na svoju aktiváciu rozpoznať antigén naviazaný na MHC molekulu. Tc lymfocyty potrebujú na svoju plnú aktiváciu 2 signály: prvý od Th lymfocytu, ktorý je aktivovaný pomocou APC bunky. Druhý signál je tvorený kontaktom Tc lymfocytu s nakazenou bunkou.

Výsledkom aktivácie lymfocytov je ich diferenciácia na efektorové (plazmocyty u B lymfocytov a aktivované T lymfocyty) a pamäťové bunky. Aktivované T lymfocyty opúšťajú lymfatické uzliny a sú schopné lokalizovať a zničiť antigény na ľubovoľnom mieste organizmu. Pamäťové bunky takisto opúšťajú lymfatické uzliny a zaisťujú rýchlejšiu reakciu v prípade ďalšieho napadnutia rovnakým antigénom. Organizmus využíva bunkovú imunitu hlavne na ničenie intracelulárnych mikróbov (vírusy). Plazmocyty ostávajú v lymfatických uzlinách a produkujú protilátky, ktoré sa potom pohybujú organizmom a napádajú antigén. Protilátková imunita sa využíva hlavne ako obrana proti extracelulárnym mikróbom, ako sú napríklad baktérie.

4.6 Umelý imunitný systém

Umelé imunitné systémy sú inšpirované svojím biologickým obrazom. Skupina týchto systémov, ktorá slúži na riešenie výpočtových problémov, sa nesnaží biologické systémy dokonale kopírovať, ale skôr čerpá z prírody inšpiráciu a snaží sa využívať len niektoré vlastnosti.

Vlastnosti imunitného systému

- *Jedinečnosť* - Imunitný systém každého organizmu je jedinečný, dokonca aj v rámci živočíšneho druhu. Jednoznačná identifikácia vlastných (zdravých) buniek je zabezpečená pomocou MHC molekúl. Táto vlastnosť umožňuje organizmu rozpoznať každú cudziu bunku a teda reagovať aj na podnety s ktorými sa nikdy doteraz nestretol.
- *Distribuovanosť, decentralizovanosť* - V IS neexistuje žiadny centrálny riadiaci prvok a systém je riadený interakciou medzi jednotlivými agentmi (molekuly, bunky). Bunky imunitného systému sú rovnomerne distribuované po celom organizme a tým je zabezpečená rýchla odozva.
- *Paralelita* - Vďaka absencii centrálneho riadiaceho prvku a distribuovanosti je systém schopný súčasne spracovávať množstvo signálov naraz.
- *Schopnosť učenia, pamäť* - Po aktivácii sa niektoré B a T lymfocyty menia na pamäťové bunky. Tieto bunky majú vyššiu životnosť a umožňujú pri ďalšom podobnom útoku rýchlejšiu reakciu IS.
- *Robustnosť* - Imunitná reakcia sa odohráva na niekoľkých úrovniach, pričom akcie na jednotlivých úrovniach sa čiastočne prekrývajú a dopĺňajú. Takto je aj pri zlyhaní niektorého mechanizmu zabezpečená obranyschopnosť organizmu.
- *Odolnosť voči šumu* - Pre rozpoznanie určitého antigénu nie je potrebná úplná zhoda medzi antigénom a receptorom imunitnej bunky. Tieto receptory majú určitú pružnosť, ktorá umožňuje imunitným bunkám vyrovnať sa s malými zmenami u antigénov.
- *Detekcia anomálií* - Vďaka schopnosti rozpoznávať vlastné a nevlastné bunky pomocou negatívnej selekcie, dokáže organizmus reagovať aj na patogény s ktorými sa doposiaľ nestretol a tiež aj na vlastné bunky, ktoré sa nechovajú korektne (napr. nádorové bunky).

4.7 Algoritmy umelého imunitného systému

Aby mohli umelé imunitné systémy napodobovať tie biologické, boli vytvorené základné algoritmy, ktoré modelujú správanie biologického imunitného systému. V nasledujúcej kapitole budú preto tieto algoritmy popísané bližšie.

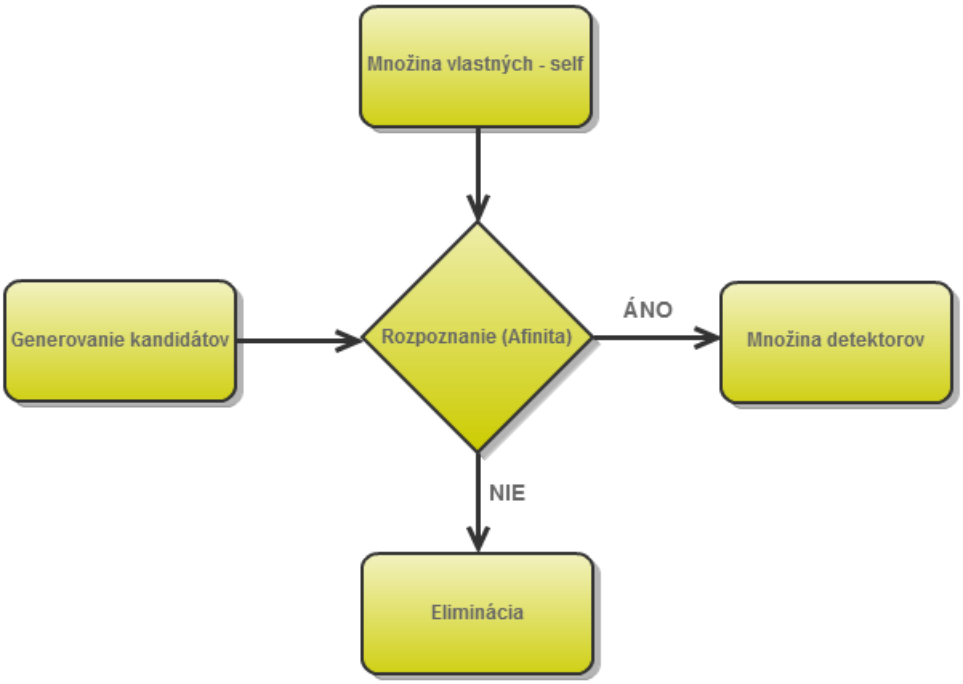
Pozitívna selekcia

Algoritmus pozitívnej selekcie vychádza z dozrievania T-lymfocytov v detskej žľaze. Aplikuje sa na elimináciu T-lymfocytov bez receptorov, či tých ktoré sú pre organizmus zbytočné 4.4.

Algoritmus

Majme množinu vlastných prvkov *self* a definovanú veľkosť množiny detektorov n. Potom samotný algoritmus vyzerá nasledovne:

1. *Inicializovanie* - náhodne vygenerujem množinu kandidátov na detektory.
2. *Cenzúra* - pokiaľ nebola vyprodukovaná množina detektorov o veľkosti n
   (a) Vyhodnotenie afinity medzi každým vlastným prvkom a kandidátom.
   (b) Ak kandidát rozpozná niektorý element množiny *self*, je tento kandidát pridaný do množiny detektorov. V opačnom prípade je eliminovaný.



Obr. 4.4: Pozitívna selekcia

Negatívna selekcia

Tento algoritmus je taktiež inšpirovaný dozrievaním T-lymfocyotv v detskej žľaze. Po vyprodukovaní nových T-lymfocytov je každý testovaný na schopnosť rozlišovať vlastné bunky. Ak to lymfocyt dokáže, je eliminovaný, dôsledkom čoho je dosiahnuté, že prežijú len lymfocyty ktoré budú útočiť len na cudzie bunky 4.5.

Algoritmus

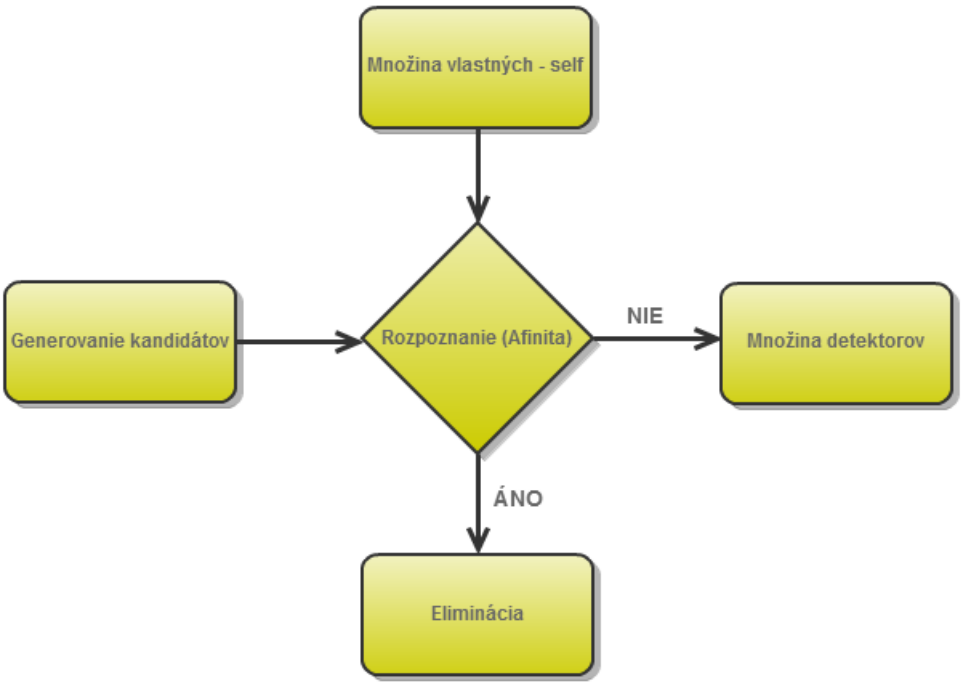
Majme množinu vlastných prvkov *self*, ktoré chceme chrániť a definovanú veľkosť množiny detektorov n. Potom samotný algoritmus vyzerá nasledovne:

1. *Inicializovanie* - náhodne vygenerujem množinu kandidátov na detektory.

2. *Cenzúra* - pokiaľ nebola vyprodukovaná množina detektorov o veľkosti n

   (a) Vyhodnotenie afinity medzi každým vlastným prvkom a kandidátom.

   (b) Ak kandidát rozpozná niektorý element množiny *self*, je tento kandidát eliminovaný. V opačnom prípade je umiestnený do množiny detektorov.



Obr. 4.5: Negatívna selekcia

Algoritmus detekcie

V prípade tohto algoritmu ide v podstate o beh umelého imunitného systému. V tomto prípade je množina detektorov porovnávaná s kontrolovanou množinou. A ak je prvok z kontrolovanej množiny rozpoznaný, môžeme o ňom prehlásiť, že je nevlastný *non-self* 4.6.

Algoritmus klonálnej selekcie

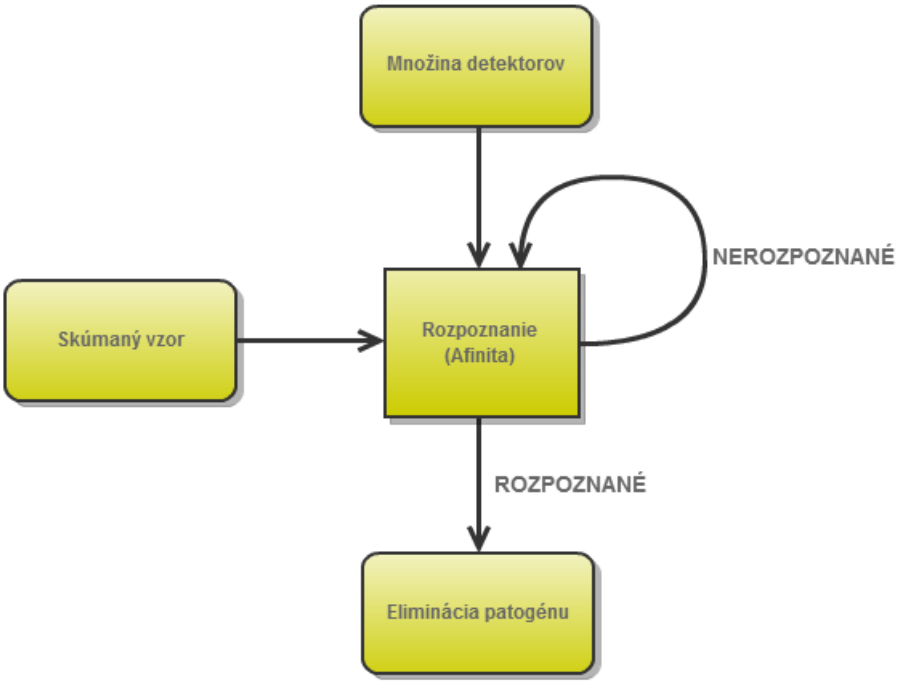
Algoritmus klonálnej selekcie vznikol na základe teórie, ktorá sa snaží vysvetliť spôsob aktivácie imunitných buniek. Tento algoritmus v svojich rôznych úpravách patrý v oblasti umelých imunitných systémov k najčastejšie používaným technikám.

Algoritmus

Majme množinu antigénov, ktoré chceme rozpoznávať a veľkosť množiny protilátok n, ktoré chceme vyprodukovať 4.7.

1. *Inicializovanie* - náhodne vygenerujem populáciu imunitných buniek.

2. *Generovanie populácie* - pre každý antigén



Obr. 4.6: Algoritmus detekcie

(a) Vyberieme len tie bunky, ktoré majú najvyššiu afinitu k antigénu.

(b) Generovanie klonov - čím lepšie daná bunka antigén rozpoznáva, tým viac kópií bunky vyprodukujeme.

(c) Mutácia - každú novú bunku zmutujeme podľa vyššie spomenutého pravidla - čím je afinita väčšia, tým budú mutácie danej bunky menšie.

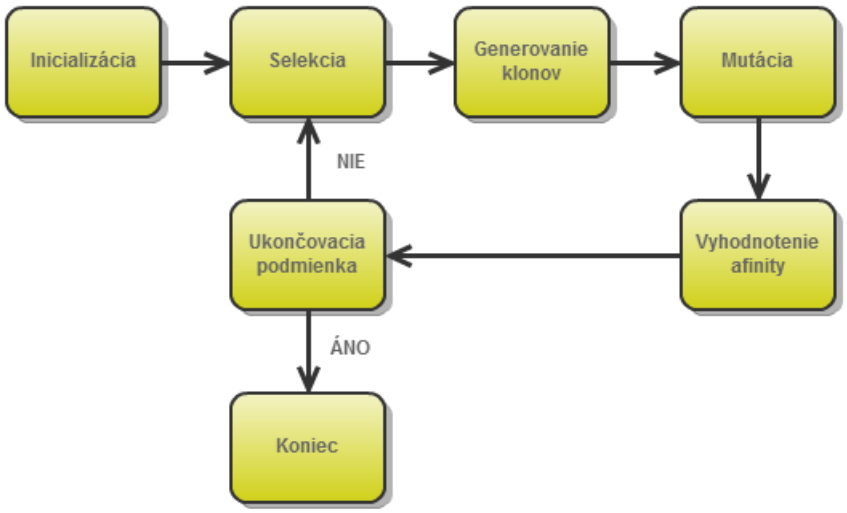
(d) Vyhodnotenie afinity - pre každú zmutovanú bunku vyhodnotíme afinitu k antigénu

3. Krok 2 opakujeme až do splnenia ukončovacieho kritéria (miera afinity, počet cyklov...)

4.8 Využitie umelých imunitných systémov k detekcii nevyžiadanej pošty

Algoritmus na filtrovanie nevyžiadaných SMS správ na princípe umelého imunitného systému

Autori článku sa učia z princípov biologického imunitného systému. V tomto prípade využívajú výhody teórie imunitných systémov k filtrovaniu informácií. Tabuľka 4.1 predstavuje analógiu biologického imunitného systému a systému na filtrovanie spam SMS. Ak by sme samotný algoritmus chceli podrobnejšie skúmať, dal by sa popísať v nasledujúcich krokoch.



Obr. 4.7: Klonálna selekcia

Biologický imunitný systém	Systém na filtrovanie spam SMS heightself
normal short message	
non-self	spam short message
antibody	detecter
antigen	unknown type of message
T-cells costimulation	user feedback
affinity	message similary
Recognise antigen	classify message

Tabuľka 4.1: Tabuľka analógie biologického a umelého IS

Predspracovanie SMS

Vzhľadom k faktu, že SMS správa je v textovej forme, predspracovanie a prečistenie správ je nevyhnutné. Dokument využíva vektorový model priestoru (VSM) k reprezentácii textu. Využíva vektor $(w_1, w_2, \cdots, w_n)$ kde w_i je váha rysu danej položky a n je dimenzia rysov. Sú použité tri kroky k vektorizácii správy:

1. predspracovanie textu správy - správa je rozdelená na slová a sú odobrané stop slová ako sú napríklad the, about, and a niektorých slov ktoré nemajú zmysel pri klasifikácii.
2. Redukovanie rysov - po predspracovaní je správa rozdelená na množstvo slov. Ak je k trénovaniu použité veľké množstvo správ, tréningový súbor môže byť veľký, preto je nutná redukcia príznakov.
3. Výpočet váh

Generovanie Antigénu a protilátok

Výstupom predspracovania správ je antigén v podobe vektoru. Bunky antigénu zodpovedajú vektoru príznakov textu. Protilátky sú predstavované knižnicou vygenerovanou počas procesu trénovania.

Definovanie a výpočet podobnosti

Podobnosť (afinita) je schopnosť protilátky naviazať sa na odpovedajúci antigén. Podobnosť je možné vyjadriť rôznymi spôsobmi. Napríklad Hammingovou vzdialenosťou či Euklidovskov vzdialenosťou.

Rozpoznanie spam SMS

V prvom rade je vypočítaná podobnosť antigénu a charakteristického prvku v pamäti protilátok. Ak je táto podobnosť väčšia ako stanovený prah, je správa označená ako nevyžiadaná. V opačnom prípade správa pokračuje na ďalšie spracovanie. Časť protilátok je uložených v knižnici protilátok, aby bola ďalšia detekcia urýchlená.

Kapitola 5

Návrh spam filtra

Pri základnom návrhu algoritmu bola inšpirácia čerpaná z článku [8]. Táto práca je bližšie popísaná v 3.3. Pre pripomenutie, autori článku vytvárajú UIS pomocou lymfocytov, ktoré reprezentujú jednotlivé slová. Každý lymfocyt si pritom uchováva hodnoty reprezentujúce jeho väzbu k hamu či spamu. V systém si udržiaval zhruba 200 lymfocytov a úspešnosť systému bola umocnená využitím pamäťových buniek. V rámci testovania s takto navrhnutým systémom dosahovali presnosť (accuracy) v rozsahu 82 - 96%
Cieľom bolo na túto prácu istým spôsobom nadviazať a rozšíriť o spôsoby, ktoré by už dosiahnuté výsledky vylepšili. Niektoré z týchto viac či menej úspešných metód si priblížime v nasledujúcom texte.

Priama selekcia lymfocytov

V rámci procesu učenia systému je vytváraná databáza slov z trénovacích sád. V databáze sú popri týchto slovách ukladané aj rôzne štatistiky popisujúce ich vlasnosti. Následne v procese výberu kandidátov na lymfocyty a pri samotnom generovaní lymfocytov sa potom podľa týchto vlastností vyberajú len najvhodnejší kandidáti. Tento výber je napríklad možné podmieniť veľkosťou pravdepodobnosti, že sa dané slovo nachádza v spame alebo naopak v hame. Rôzna voľba kandidátov na lymfocyty spôsobuje, že systém dosahuje viac, či menej uspokojivé výsledky v rámci samotnej detekcie. Samotná voľba kandidátov na lymfocyty a jej vplyv na výsledky systému bude podrobnejšie diskutovaná v ďalšom texte.

Rozdelenie množín lymfocytov

Ak vychádzam z predpokladu, že lymfocyt, ktorý sa viaže na slová nachádzajúce sa s veľkou pravdepodobnosťou výskytu v spame, upozorňuje na príslušnosť samotnej správy k nevyžiadanej pošte. Je možné predpokladať, že podobné chovanie bude mať lymfocyt, ktorý má naopak veľkú pravdepodobnosť výskytu v hame bude upozorňovať na skutočnosť, že správa patrí do pošty vyžiadanej. Preto som sa rozhodol rozdeliť množinu lymfocytov na skupinu lyfocytov identifikujúce spam a skupinu identifikujúcu ham. Týmto spôsobom sa podarilo navýšiť úspešnosť systému odhalovať nevyžiadanú poštu a zároveň znížiť počet správ z triedy ham, ktoré systém označí ako spam. Podrobnejší popis rozdelenia množiny lymfocytov a jeho vplyv na výsledky systému budú popisované ďalšej časti práce.

Štatistika dĺžky slov

Na rozdiel od základnej metódy tento doplňujúci UIS používa ako hlavnú črtu lymfocytov dĺžku slova. Každý lymfocyt si udržiava dĺžku slova a hodnotu, ktorá vyjadruje výskyt slov takejto dĺžky v správach typu spam alebo ham. Vychádza sa pri tom z priemerného počtu slov takejto dĺžky. V jednom prípade bol tento UIS trénovaný už počas procesu trénovania celého systému. V druhom prípade bol tento sekundárny UIS trénovaný až počas behu (procesu detekcie) primárneho UIS, predpokladalo sa totiž, že sa takto lepšie prispôsobí štýlu správ daného užívateľa. V obidvoch prípadoch však výsledky boli značne zmetočné a prinášali do systému skôr chybovosť ako zlepšenie kvality výsledkov.

Dvojslová

Táto metóda v podstate pracuje veľmi podobne ako metóda popisovaná v práci už vyššie spomínanej. V tomto prípade sa vychádzalo z predpokladu, že ak môže byť pre spam signifikantné jedno slovo, tak skupina (v našom prípade dvojica) slov môže niesť rovnakú, prípadne vyššiu informáciou o tom, či je správa ham alebo spam. Táto metóda sa osvedčila a výsledky boli potešujúce. Vo finále táto metóda dopĺňa hlavný UIS. Jej podrobnejší popis je možné nájsť v nasledujúcich kapitolách 6.4.

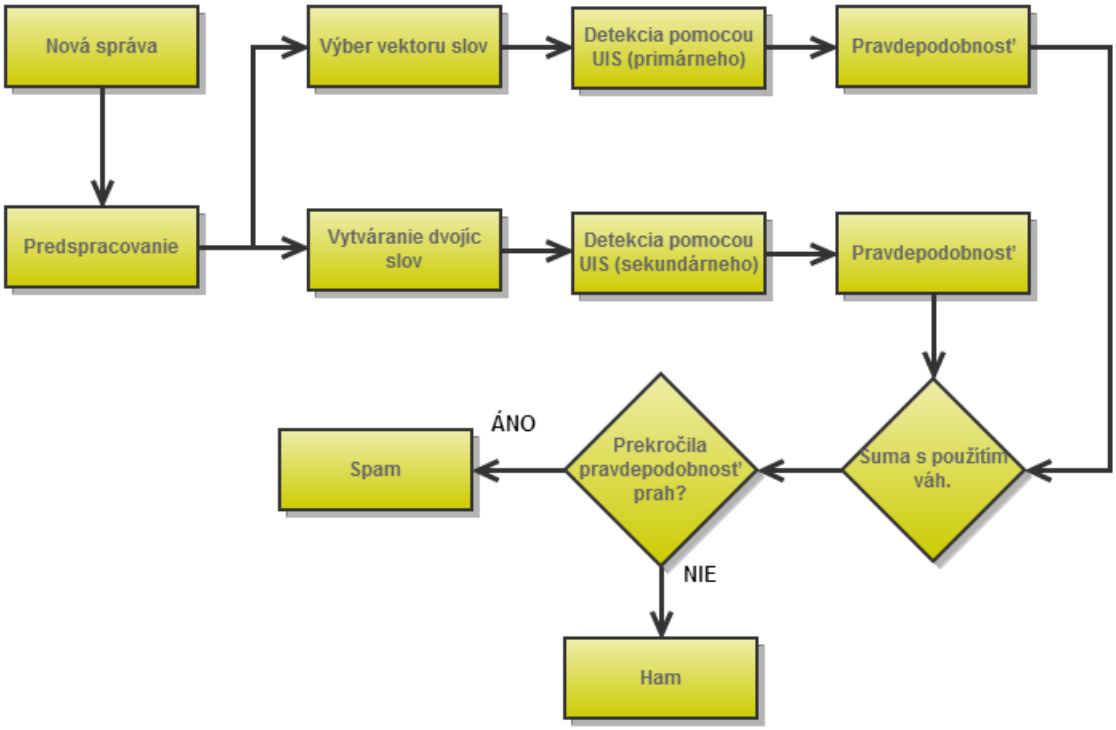
5.1 Navrhnutý systém

Vo finále je systém navrhovaný ako hybridné spojenie spájajúce výhody jednoduchého bayesovho systému a umelých imunitných systémov. Vo výsledku systém pracuje so štyrmi množinami vhodne volených lymfocytov, ku ktorým je pristupované ako k dvom samostatne pracujúcim umelým imunitným systémom. Prímárny imunitný systém pracuje dve množiny lymfocytov (lymfocyty pre spam a lymfocyty pre ham), ktoré sa viažu na jednotlivé slová. Druhý umelý imunitný systém taktiež používa dve množiny lymfocytov ako systém primárny, no na rozdiel od neho sa tieto lymfocyty viažu na dvojslová. V prípade správneho naviazania lymfocytu na správu sa hodnoty lymfocytu aktualizujú. V prípade ak ide o správu z triedy spam, aktualizujú sa hodnoty lymfocytov v množine spam lymfocytov a naopak v prípade, že išlo o správu typu ham, aktualizujú sa hodnoty lymfocytov z ham množiny. Každý z dvoch imunitných systémov pomocou bayesovho vzťahu vypočíta pravdepodobnosť príslušnosti testovanej správy k nevyžiadanej pošte. Výstupom samotnej detekcie sú potom dve pravdepodobnosti získané z primárneho a sekundárneho imunitného systému. Tieto pravdepodobnosti sú pomocou vzťahu 5.1 prepočítané na výslednú pravdepodobnosť, ktorá definitívne určuje, či testovaná správa patrí do množiny pošty vyžiadanej alebo nevyžiadanej. Celý navrhovaný systém by potom bolo možné pomocou diagramu znázorniť nasledovne 5.1.

$$Score = P_{spam} = \frac{P_{primarny\,UIS} + (2 \cdot P_{sekundarny\,UIS})}{3} \tag{5.1}$$

5.2 Návrh aplikácie

Aplikácia bola od počiatku navrhovaná s dôrazom na rozčlenenie do logických a funkčných blokov. Prínosom takéhoto návrhu by v prvom rade mala byť lepšia správa a údržba zdrojových kódov a v neposlednej rade tiež možnosť testovania logických celkov. Návrh bol



Obr. 5.1: Diagram navrhnutého systému.

realizovaný bez ohľadu na programovací jazyk, v ktorom je samotná aplikácia implementovaná. Celok je rozčlenený do nasledujúcich blokov:

- Lexikálny blok - spracovanie správ na úrovni raw podoby správ, odstránenie pre ďalšiu činnosť aplikácie zbytočných častí správ, práca s HTML telom správ.
- Ukladanie a práca s dátami - manipulácia, spracovávanie a ukladanie dát v perzistentnej forme.
- Nastavenia - uchovávanie nastavení aplikácie.
- Výsledky - uchovávanie a spracovanie výsledkov a výstupov aplikácie.
- Umelý imunitný systém - implementuje UIS, jeho súčasti a algoritmy.
- Grafické užívateľské rozhranie - spracováva vstupy od užívateľa a sprostredkúva mu výstupy behu aplikácie.

Kapitola 6

Implementácia aplikácie

Pre samotnú implementáciu bol zvolený programovací jazyk C# a .Net Framework verzie 4. Ako vývojové prostredie bolo pre svoju funkčnosť a komplexnosť volené Visual Studio 2010. Vrámci implementácie boli použité rôzne podporné knižnice na vykresľovanie grafov (WPFToolkit), na prácu s databázou (SQLite adapter, SQLite Expert Personal) či v prípade GUI, framework s možnosťou vytvárania dokovacích okien podobných ako poznáme z Visual Studia 2010 (AvalonDock 2.0).

6.1 Predspracovanie správ

Aplikácia, či už v režime učenia (trénovania), alebo počas jej samotného behu (testovania), na vstupe očakáva e-mailové správy v ich surovej (raw) forme. Takéto správy, obsahujú množstvo nadbytočných informácií, ktoré pre samotný beh aplikácie nie sú kľúčové.

```
Message-ID: <31690903.1075854043652.JavaMail.evans@thyme>
Date: Tue, 19 Sep 2000 04:28:00 -0700 (PDT)
From: john.griffith@enron.com
To: daren.farmer@enron.com
Subject: Cornhusker
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: John Griffith
X-To: Daren J Farmer
X-cc:
X-bcc:
X-Origin: Farmer-D
X-FileName: dfarmer.nsf

Darren,

How are things going?

Thanks.
John
```

Výpis 6.1: E-mailová správa v raw forme

Preto je pred každým spracovaním nutné zo správy získať len pre nás užitočné informácie. Tieto informácie sú v systéme prezentované štruktúrou dipMessage. Táto štruktúra

uchováva napríklad odosielateľa, príjemcu, HTML časť tela e-mailu, telo e-mailu v textovej prezentácii a jej definíciu v systéme môžeme vidieť nižšie.

```
public struct dipMessage
{
    public String From      { get; set; }
    public String To        { get; set; }
    public String Subject   { get; set; }
    public String TextBody  { get; set; }
    public String HtmlBody  { get; set; }
    public String SentDate  { get; set; }
}
```

Výpis 6.2: Štruktúra dipMessage

V mnohých prípadoch je telo správy reprezentované ako HTML kód. Takéto telo je v priebehu predspracovania interpretované a prevedené na normálny text. Takto sa eliminujú pokusy útočníka o zatemnenie textu vo forme HTML, tak ako to ukazuje nasledujúci príklad, kde je takéto zatemnenie vykonané pomocou komentárov.

```
V<!--THYME-->ia<!--ONION-->gr<!--ALMOND-->a
```

Výpis 6.3: Zatemňovanie textu pomocou HTML komentárov.

Následne je na text aplikovaná haxorovacia funkcia, ktorá má za úlohu upraviť slová typu `V@iagr@` na slová `Viagra` a teda odstrániť špeciálne znaky z textu. Takto upravená a spracovaná správa je pripravená buď to na uloženie do databázy, alebo k ďalšiemu spracovaniu.

Uchovávanie dát

Je správnym predpokladom, že aplikácia počas svojho behu operuje s pomerne veľkým množstvom dát. V tomto prípade by klasické ukladanie, či už formou textových alebo XML súborov, bolo pomerne neefektívne a neúnosné. Napriek tomu, je systém navrhnutý s ohľadom na obecnosť a možnosť úpravy podľa budúcich požiadaviek. Preto implementuje triedu dipMiddleWare. Táto medzivrstva sa stará o ukladanie a načítanie dát. Úplne tak odtieňuje túto činnosť od dát a štruktúr používaných aplikáciou. Preto je pomerne jednoduché vykonať zmenu ukladania dát na iný spôsob, napríklad už spomínané XML súbory alebo MS SQL či iné.

SQLite

V aktuálnej verzii aplikácia pre ukladanie dát využíva SQLite. Tento spôsob bol zvolený pre nesporné prínosy oproti textovým či XML súborom a zároveň pre jednoduchosť nasadenia, napríklad voči MS SQL. Zároveň umožňuje využitie jazyka SQL, ktorý bol priamo navrhnutý a optimalizovaný na prácu s veľkým objemom dát. Okrem výberu a ukladania dát je SQLite aplikáciou využívaná aj k výpočtom. Napríklad pomocou triggerov spúšťaných pri vkladaní a aktualizácii záznamov. V tomto prípade je databázou počítaná pravdepodobnosť, že dané slovo sa vyskytuje v správach patriacich do triedy spam. Ukážku tabuľky s dátami môžeme vidieť na 6.1.

RecNo	Word	HamCount	SpamCount	MsgMatched	MsgSpam	SpamProbability	Evolute	Spam_w	Ham_w	IsDw
1	Even	31	17	48	17	35	35	0.0179640718562874	0.031936127744511	☐
2	better	66	35	101	35	34	34	0.0359281437125748	0.0668662674650699	☐
3	http	449	323	772	323	41	41	0.323353293413174	0.449101796407186	☐
4	ridiculopathy	1	0	1	0	0	0	0.00099800399201596	0.00199600798403194	☐
5	news_detail	1	0	1	0	0	0	0.00099800399201596	0.00199600798403194	☐
6	668White	1	0	1	0	0	0	0.00099800399201596	0.00199600798403194	☐
7	House	10	10	20	10	50	50	0.0109780439121756	0.0109780439121756	☐
8	President	22	17	39	17	43	43	0.0179640718562874	0.0229540918163673	☐
9	Boner	1	0	1	0	0	0	0.00099800399201596	0.00199600798403194	☐
10	Must	3	12	15	12	80	80	0.0129740518962076	0.00399201596806387	☐

Obr. 6.1: Ukážka výpisu z DB.

6.2 Učenie - trénovanie

V tomto prípade sú vstupom procesu učenia dve čo do počtu rovnaké množiny. Jedna obsahujúca správy patriace do triedy spam a druhá naopak správy spadajúce do triedy ham. Každá správa z týchto množín je spracovaná pomocou procesu predspracovania a uložená vo forme štruktúry dipMessage. Z tejto štruktúry je následne použitá položka TextBody, obsahujúca predspracovaný text. Následne je tento text upravovaný a sú z neho odstránené príliš krátke a príliš dlhé slová. Tento text je potom rozdelený na jednotlivé slová a tie sú uložené nasledujúcim spôsobom do histogramu.

```
// Vytvaranie histogramu slov.
UpdateDictionary(Dictionary<String, Occurrence> dWords, String[]
    sWords, bool isSpam)
{
        if (isSpam) // V pripade aj je sprava typu spam.
        {
                // Pre kazde slovo v tejto sprave.
                foreach (var word in sWords)
                {
                        // Ak sa uz v histograme nachadza, navys
                            jeho pocet vykytov o jedno.
                        if (dWords.ContainsKey(word))
                        {
                                var occ = dWords[word];
                                occ.oSpam++;
                                occ.MsgSpam++;
                                occ.MsgMatched++;
                                dWords[word] = occ;
                        }
                        // Ak sa slovo v histograme este neobjavilo
                            pridaj ho do histogramu.
                        else
                                dWords.Add(word, new
                                    Occurrence(0,1,1));
                }
        }
        // Ak je sprava typu ham.
        else
        {
                // Pre kazde slovo v tejto sprave.
```

```
                foreach (var word in sWords)
                {
                        // Ak sa uz v histograme nachadza, navys
                           jeho pocet vykytov o jedno.
                        if (dWords.ContainsKey(word))
                        {
                                var occ = dWords[word];
                                occ.oHam++;
                                occ.MsgMatched++;
                                dWords[word] = occ;
                        }
                        // Ak sa slovo v histograme este neobjavilo
                           pridaj ho do histogramu.
                        else
                                dWords.Add(word, new
                                   Occurrence(1,0,0));
                }
        }
}
```

Výpis 6.4: Vytváranie histogramu slov.

Takto získaný histogram je uložený do databázy. Počas ukladania sú pomocou vzťahov 6.1 a 6.2, 6.3 vychádzajúcich z 3.6 a 3.7 vypočítané doplňujúce informácie o týchto slovách. Týmto je proces trénovanie úspečne ukončený a užívateľ je pomocou GUI o tejto skutočnosti informovaný.

$$pravdepodobnost\ v\ spame = \frac{pocet\ v\ spame}{(pocet\ v\ hame + pocet\ v\ spame)} \cdot 100 \tag{6.1}$$

$$spam\_w = \frac{1 + SpamFreq}{2 + n\_spam} \tag{6.2}$$

$$ham\_w = \frac{1 + HamFreq}{2 + n\_ham} \tag{6.3}$$

Kde SpamFreq predstavuje počet výskytov daného slova v množine správ spam. HamFreq je počet výskytov daného slova v množine správ ham. Premenná n_spam značí veľkosť trénovacej množiny spam správ a n_ham predstavuje veľkosť trénovacej množiny ham správ.

Výber lymfocytov

Ako sa predpokladá umelý imunitný systém pre svoju správnu činnosť potrebuje lymfocyty. Jeden lymfocyt je v aplikácii reprezentovaný triedou dipLymphocyte, ktorá obsahuje nasledujúce položky

```
public class dipLymphocyte
{
        /// Time to live, doba zivota daneho lymfocytu.
        protected Int64 iTtl = 2000;
        /// Pocet sprav, ktore dany lymfocyt testoval.
        protected Int64 iMsgMatched = 1;
        /// Pocet sprav, ktore lymfocyt identifikoval ako spam.
```

```
        protected Int64 iMsgSpam = 0;
        /// Uspesnost daneho lymfocytu pri detekcii spamu.
        protected Double fSuccessRate = 0;
        /// Frekvencia vyskytu slova reprezentujuceho lymfocyt v
           spame.
        protected Int64 SpamFreq = 0;
        /// Frekvencia vyskytu slova reprezentujuceho lymfocyt v
           hame.
        protected Int64 HamFreq = 0;
        /// Hodnota s akou sa viaze dany lymfocyt na Spam.
        protected Double Spam_w;
        /// Hodnota s akou sa viaze dany lymfocyt na Ham;
        protected Double Ham_w;
        /// Prah pri ktorom sa lymfocyt meni na pamatovu bunku.
        protected Double MemoryCellTreshold = 0.97;
}
```

Výpis 6.5: Trieda reprezentujúca lymfocyt.

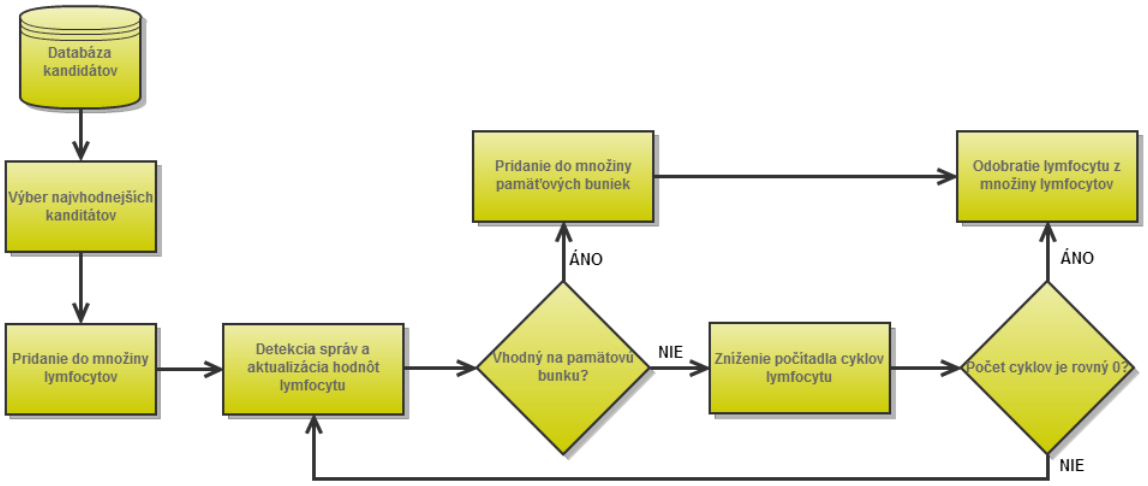
Dôležitou úlohou je však vytvorenie lymfocytov. V prípade, že by sme za lymfocyt pokladali každé slovo uložené počas procesu učenia do databázy, mohlo a aj by sa veľmi často stávalo, že by počet lymfocytov presiahol hranicu 100000. Priamym a pomerne nežiaducim by bol dôsledok jednak veľkej pamäťovej náročnosti a zároveň nárast časovej náročnosti v prípade prehľadávania tak rozsiahlej množiny lymfocytov. Preto je nutná správna voľba slov, ktoré budú reprezentovať lymfocyty. Pre potreby aplikácie sú vyberané dve množiny. Množina kandidátov určených pre spam kde platí že sa vyberá n kandidátov s najvyšším $spam\_w$ získaným pomocou vzťahu 6.2 a pravdepodobnosťou, že sa slovo nachádza v spame získanou vzťahom 6.1, väčšou ako 80%. V prípade množiny pre ham sa naopak volia kandidáti, ktorí majú najvyššiu hodnotu ham_w, získanú vzťahom 6.3. Taktiež v prípade tohto výberu platí, že sa vyberajú len kandidáti, ktorých pravdepodobnosť výskytu v spame nepresiahne 40%. Veľkosť týchto množín je v aplikácii možné upravovať pomocou nastavení v GUI.

Životný cyklus lymfocytu z množiny spam

Počas behu systému môže lymfocyt prechádzať niekoľkými fázami životného cyklu. Prvou a pre každý lymfocyt najbežnejšou fázou je stav vyčkávania na detekciu a samotná detekcia. Druhou fázou, ktorú nemusí dosiahnuť každý lymfocyt je premena na pamäťovú bunku. Ďalšou a v tomto prípade aj poslednou fázou môže byť, keď lymfocyt presiahne počet cyklov detekcie a je zo systému odstránený. Tento životný cyklus je zachytený na diagrame 6.2

Životný cyklus lymfocytu z množiny ham

Podobne ako lymfocyt z množiny spam, aj tento lymfocyt sa môže nachádzať vo fázy čakania na detekciu a detekcie. Druhou a v tomto prípade posladnou je fáza, kedy vypršal počet detekčných cyklov lymfocytuje a je zo systému odstránený. Keďže pamäťové bunky striktne detekujú, či je daná správa triedy spam, nie je vhodné, aby sa lymfocyt z množiny ham transformoval na pamäťovú bunku.



Obr. 6.2: Cyklus lymfocytu v systéme.

Pamäťové bunky

Pamäťové bunky vo veľkej miere prispievajú k nielen vyššej rýchlosti vyhodnocovania, ale aj k vyššej presnosti celého systému. Typickou črtou týchto buniek je oproti lymfocytom omnoho väčší počet cyklov detekcie. V zásade platí, že keď sa slovo z príchodzej správy detekuje pomocou pamäťovej bunky, je celá správa označená ako spam. Je zrejmé a žiaduce aby sa pamäťovou bunkou stali len lymfocyty s vysokou úspešnosťou v prípade detekovania spamu. Toto sa dosahuje pomocou nasledujúcich dvoch krokov.

1. Každý lymfocyt si udržiava hodnotu fSuccessRate, ktorá je vypočítaná pomocou vzťahu 6.4
2. Pamäťovou bunkou sa môže stať jedine lymfocyt, ktorého hodnota fSuccessRate presiahne predom nastavený prah MemoryCellTreshold.

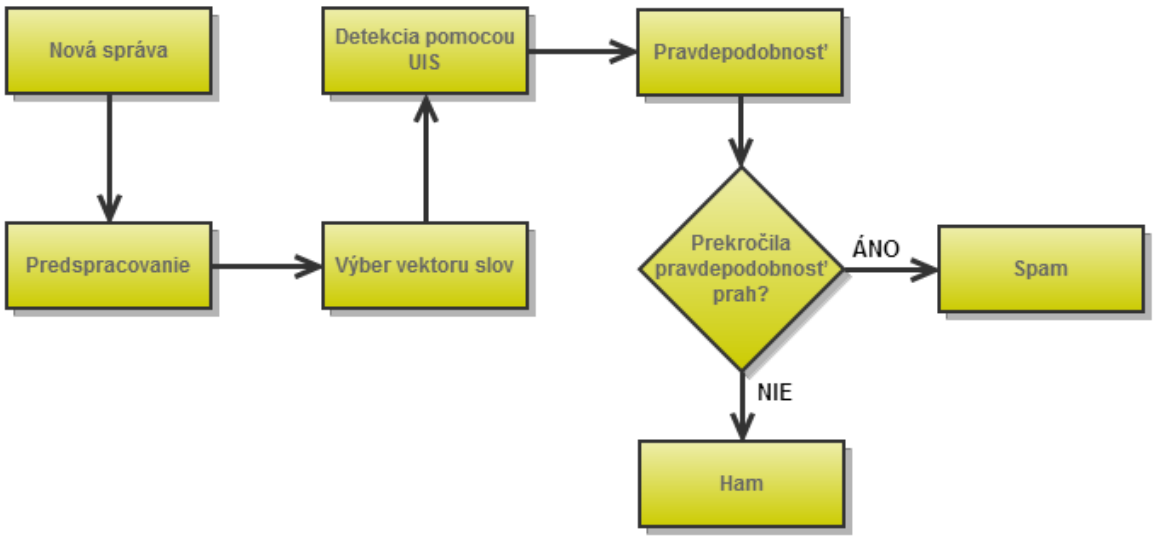
$$fSuccessRate = \frac{iMsgSpam}{iMsgMatched} \tag{6.4}$$

6.3 Testovanie - proces detekcie

Od chvíle, keď je pomocou procesu trénovania vytvorená báza dát, nič nebráni tomu, aby bola aplikácia používaná k určovaniu či ten ktorý e-mail patrí do triedy spam alebo naopak do triedy ham. Samotný priebeh detekcie je znázornený diagramom 6.3 a jeho jednotlivé kroky sú podrobnejšie popísané v nasledujúcom texte.

Normalizácia vstupnej správy

Vstupom podobne ako v prípade procesu trénovania sú správy v ich surovej (raw) forme. Tieto správy je nutné taktiež spracovať a dostať do podoby reprezentovanej štruktúrou dipMessage. Z tejto štruktúry sa potom aj v tomto prípade pracuje len s položkou TextBody.



Obr. 6.3: Priebeh procesu detekcie.

Následne sú opäť vybrané len slová s vyhovujúcou dĺžkou. No na rozdiel od procesu trénovania sa v tomto prípade zo vzniknutej množiny slov odstránia viacnásobné výskyty jednotlivých položiek. Takto vzniknutý vektor príznakov je vstupom samotného algoritmu detekcie. Možný príklad implementácie výberu slov obmedzenej dĺžky pomocou jazyka LINQ (Language Integrated Query) je možné vidieť v nasledujúcej ukážke.

```
public String[] RemoveShortWords(String[] sWords, int iMinLength =
    4 , int iMaxLength = 15)
{
        var WordsQuery =
                from word in sWords
                where word.Length >= iMinLength && word.Length <=
                    iMaxLength
                select word;

        return WordsQuery.ToArray();
}
```

Výpis 6.6: Výber slov žiadanej dĺžky pomocou jazyka LINQ.

Naviazanie lymfocytu na slovo

Všetky získané slová zo vstupnej e-mailovej správy sú postupne porovnávané s množinou lymfocytov. V prípade pozitívnej reakcie a teda zhody a naviazania lymfocytu na dané slovo sa lymfocyt aktualizuje a zároveň sú z neho získané hodnoty Spam_w a Ham_w. Tieto hodnoty sú uložené vo vektore, ktorý obsahuje aj hodnoty ostatných pozitívne detekovaných slov.

```
public void DecectLymphocytes(String[] sUniqWords, List<_w_> W_List)
{
```

```
        if (dLymphocytes.Count > 0)
        {
                var sorted = from s in sUniqWords
                                        orderby s.Length descending
                                        select s;
                // Detekcia.
                foreach (var word in sorted)
                {
                        var l = dLymphocytes.Find(word);
                        if (l != null)
                                l.GetAffinity(W_List);
                }

                // Aktualizacia hodnot lymfocytov.
                dLymphocytes.UpdateLymphocytes();

                // Premena lymfocytov na pametove bunky.
                var toMemoryCells =
                   dLymphocytes.GetMemoryCellsCandidate();
                var toRemove = dMemoryCells.Add(toMemoryCells,
                   m_settings.MaxMemoryCellCycles);

                // Odstranenie lymfocytov, ktore sa zmenili na
                   pametove bunky
                // alebo presiahli pocet cyklov detekcie.
                dLymphocytes.Remove(toRemove);
        }
}
```

Výpis 6.7: Vyhľadávanie slov v množine lymfocytov.

Výpočet pravdepodobnosti

Výsledkom behu aplikácie je určenie príslušnosti danej správy k triede ham alebo triede spam. V princípe platí vzťah 6.5.

$$p = \begin{cases} 1 & p_{spam} > treshold \\ 0 & p_{spam} < treshold \end{cases} \tag{6.5}$$

Z vektorov hodnôt spam_w a ham_w nie je problém určiť hodnotu p_spam pomocou vzťahu 6.6.

$$p_{spam} = \frac{\frac{N_S}{N} \times \prod_{i=1}^{n} Spam\_w}{\frac{N_S}{N} \times \prod_{i=1}^{n} Spam\_w + \frac{N_H}{N} \times \prod_{i=1}^{n} Ham\_w} \tag{6.6}$$

V programe by potom tento výpočet mohol vyzerať nasledovne.

```
Double p_spam = 1.0;
Double SpamW_sum = 1.0, HamW_sum = 1.0,

foreach (var res in W_list)
{
```

```
        SpamW_sum *= res.w_spam;
        HamW_sum *= res.w_ham;
}

var a = (Double)CurrentSpamCount / CurrentMsgCount;
var b = (Double)CurrentHamCount / CurrentMsgCount;

var c = (a * SpamW_sum);
var d = (b * HamW_sum);

p_spam = c / (c + d);
```

Výpis 6.8: Výpočet hodnoty p_spam.

Aktualizácia hodnôt lymfocytov

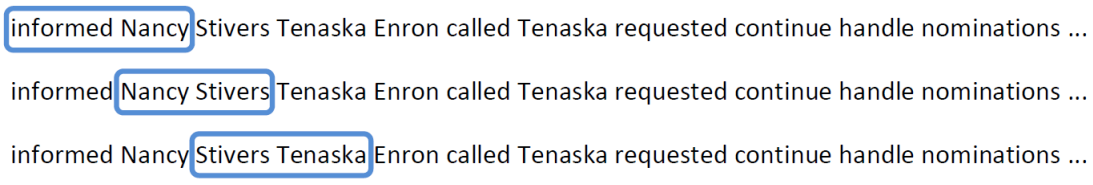
V prípade aktualizácie hodnôt lymfocytov môžu nastať dva prípady. Prvým je, že správa bola označená ako spam. Potom sú aktualizované, prípadne doplnené nové lymfocyty do množiny spam lymfocytov. Ak lymfocyty v tejto množine len aktualizujeme, je inkrementovaná hodnota SpamFreq a za použitia vzťahu 6.2 je vypočítaná aktuálna hodnota spam_w. Druhým prípadom je označenie správy ako ham. Tu sa aktualizujú, prípadne pridávajú nové lymfocyty do množiny ham lymfocytov. V prípade aktualizácie lymfocytov je inkrementovaná hodnota HamFreq a použitím vzťahu je prepočítaná hodnota ham_w.

6.4 Dvojslovné lymfocyty

Predchádzajúci spôsob detekcie dosahuje pomerne uspokojivé výsledky. Aj napriek tomu je snaha tieto výsledky ešte vylepšiť a urobiť ich uspokojivejšie. Z mnohých, aj osobne testovaných, spôsobov padla voľba na použitie metódy využívajúcej nielen jednoslovné ale aj dvojslovné lymfocyty. Kroky v ktorých sa táto metóda odlišuje od vyššie popisovanej jednoslovnej metódy budú popísané v nasledujúcom texte.

Učenie

Učenie prebieha podobne ako v prípade metódy vystavanej na použití jednotlivých slov. Vstupný text je však pomocou okna o šírke dvoch slov delený na dvojice, podobne ako znázorňuje obrázok 6.4. Z takto získaných dvojíc sa opäť vytvorí histogram, ktorý sa uloží do databázy.



Obr. 6.4: Vytváranie dvojslov zo vstupného textu.

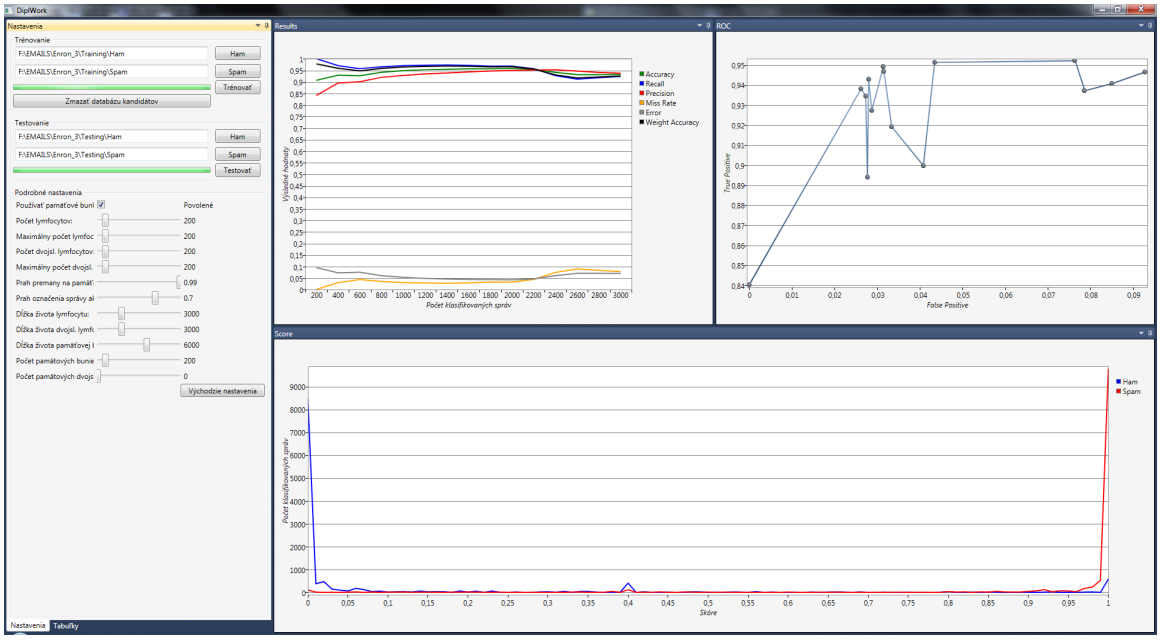
Proces detekcie

Proces detekcie prebieha veľmi obdobne ako v prípade systému s jednoslovnými lymfocytmi. V tomto prípade je však vynechaný krok s odstránením viacnásobných výskytov slova v texte. Po takomto spracovaní sa vektor príznakov, vytvorený zo vstupnej správy, prechádza a presne ako v prípade učenia sa vytvárajú dvojslová. Takto vzniknuté dvojslová sa následne porovnávajú s množinou dvojslovných pamäťových buniek a lymfocytov. Množina lymfocytov vznikla na základe databázy a dvojslov spĺňajúcich podmienku pravdepodobnosti výskytu týchto dvojíc v správach typu spam. Pravdepodobnosť príslušnosti do triedy spam je v prípade príchodzej správy počítaná rovnakým spôsobom ako v prípade jednoslovných lymfocytov.

Kapitola 7

Popis a použitie aplikácie

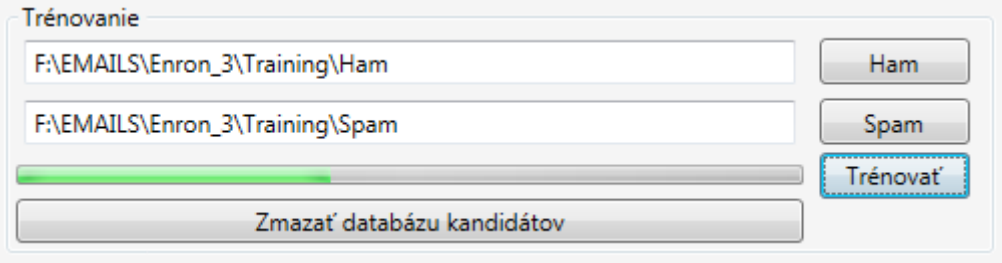
Grafické užívateľské rozhranie (GUI) umožňuje s aplikáciou jednoducho pracovať a meniť jej prevádzkové nastavenia. Podobne ako je rozdelený beh aplikácie na proces učenia (trénovania), detekcie (testovania) a výstupy, má aj GUI aplikácie podobné členenie 7.1. Podrobný prácu s GUI popisuje nasledujúca kapitola.



Obr. 7.1: Hlavné okno aplikácie.

7.1 Trénovanie

Aj napriek tomu, že proces trénovania je pomerne zložitý a pre samotný beh aplikácie v podstate nevyhnutný, je jeho nastavenie v GUI pomerne jednoduché. Ako vidno na obrázku 7.2 je nutné určiť cesty k trénovacím sadám ham a spam. Následne je možné samotný proces trénovania spustiť tlačítkom Trénovať. O priebehu procesu trénovania užívateľa informuje progressbar. O ukončení behu procesu je užívateľ informovaný pomocou informačného dialógu. V prípade použitia novej trénovacej sady je nutné databázu zmazať použitím tlačidla Zmazať databázu kandidátov.



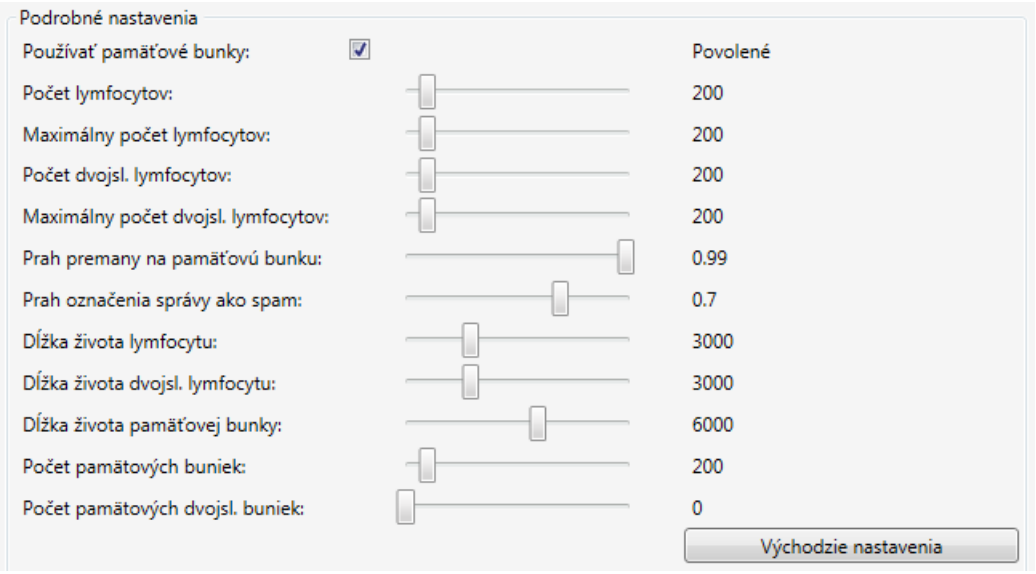
Obr. 7.2: Nastavenie trénovania.

7.2 Testovanie

Podobne ako v prípade trénovania, aj v tomto prípade je nutné určiť množiny správ, ktoré chceme testovať. Samotné testovanie a najmä jeho výsledky je možné ovplyvňovať pomocou nastavení 7.3. O priebehu je užívateľ opäť informovaný pomocou progressbaru a na jeho konci informačným dialógom.

7.3 Nastavenia

Úpravou parametrov nastavení je možné výrazne meniť a ovplyvňovať správanie aplikácie a tým aj samotných výsledkov. Pomocou nastavení je možné meniť nasledujúce parametre 7.3:



Obr. 7.3: Nastavenia.

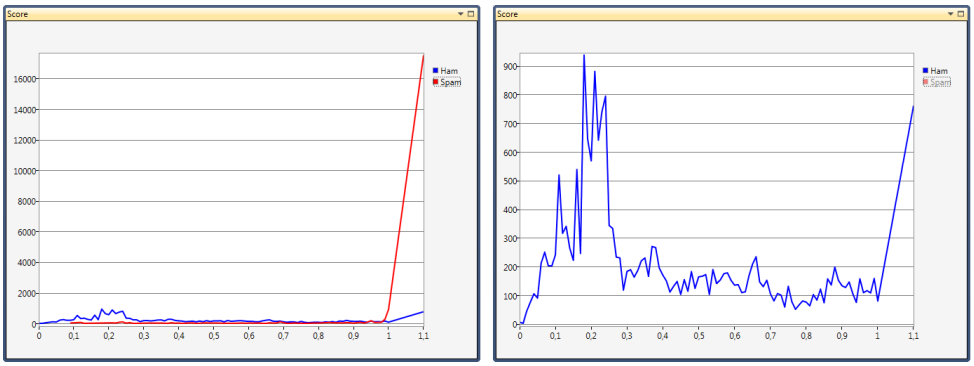
- *Používať pamäťové bunky* - určuje, či sa v priebehu detekcie budú používať pamäťové bunky.
- *Počet lymfocytov* - stanovuje počet lymfocytov, ktoré budú vytvorené na základe databázy.

- *Maximálny počet lymfocytov* - určuje maximálny počet lymfocytov v systéme.
- *Počet dvojsl. lymfocytov* - stanovuje počet dvojslovných lymfocytov, ktoré budú vytvorené na základe databázy.
- *Maximálny dvojsl. lymfocytov* - určuje maximálny počet dvojslovných lymfocytov v systéme.
- *Prah premeny na pamäťovú bunku* - parameter stanovujúci prah, pri ktorom sa môže lymfocyt transformovať na pamäťovú bunku.
- *Prah označenia správy ako spam* - je nastavenie určujúce prah pravdepodobnosti, od ktorého je správa označená ako spam.
- *Dĺžka života lymfocytu* - určuje maximálny počet cyklov detekcie lymfocytu.
- *Dĺžka života dvojsl. lymfocytu* - určuje maximálny počet cyklov detekcie dvojslovného lymfocytu.
- *Dĺžka života pamäťovej bunky* - určuje maximálny počet cyklov detekcie pamätovej bunky.
- *Počet pamätových buniek* - určuje maximálny počet pamäťových buniek v systéme.
- *Počet dvojsl. pamätových buniek* - určuje maximálny počet dvojslovných pamäťových buniek v systéme.
- *Východzie nastavenia* - nastaví systém na pôvodné nastavenia.

7.4 Výstupy

Samozrejme je nutné mať možnosť si zobraziť výstupy systému. Tie sú aplikáciou zobrazované jednak v podobe tabuliek a samozrejme aj vo forme grafov.

Grafy



Obr. 7.4: Práca s grafom.

- V prípade grafov, je tieto možné uberať či pridávať pomocou pripináčiku, rovnako ako samotné okná.

- Taktiež je možné vypínať a zapísať pomocou klikania na názov série v legende grafu jednotlivé série zobrazované v grafoch. V tomto prípade, ak je to možné graf automaticky prepočíta mierku na svojich osiach.

Tabuľky

Aplikácie zobrazuje dosiahnuté výsledky aj pomocou tabuliek. V prvej tabuľke 7.5 je možné nájsť priemerné hodnoty testov. Druhá tabuľka 7.6 potom predstavuje podrobný výpis ohodnotenia jednotlivých súborov predstavujúcich e-mailové správy.

Tabuľka výsledkov

Name	Value
Average accuracy	0,9403
Average recall	0,9564
Average precision	0,9279
Average miss rate	0,0436
Average error	0,0597
Average ham score	0,0993
Average spam score	0,9537

Obr. 7.5: Zobrazenie výsledkov v podobe priemerov.

Podrobný výpis hodnotenia správ

Path	Probability	Result	Real
F:\EMAILS\Enron_3\Testing\Spam\g35	0,0815	INCORRECTLY	SPAM
F:\EMAILS\Enron_3\Testing\Ham\msg2037.eml	0,8023	INCORRECTLY	HAM
F:\EMAILS\Enron_3\Testing\Spam\h2	0,3329	INCORRECTLY	SPAM
F:\EMAILS\Enron_3\Testing\Spam\q22	0,4	INCORRECTLY	SPAM
F:\EMAILS\Enron_3\Testing\Spam\q77	0,0021	INCORRECTLY	SPAM
F:\EMAILS\Enron_3\Testing\Spam\q205	0,4	INCORRECTLY	SPAM
F:\EMAILS\Enron_3\Testing\Spam\q222	0,663	INCORRECTLY	SPAM
F:\EMAILS\Enron_3\Testing\Ham\msg2343.eml	1	INCORRECTLY	HAM
F:\EMAILS\Enron_3\Testing\Ham\msg2342.eml	1	INCORRECTLY	HAM
F:\EMAILS\Enron_3\Testing\Ham\msg648.eml	0	CORRECTLY	HAM
F:\EMAILS\Enron_3\Testing\Ham\msg2484.eml	0,1949	CORRECTLY	HAM
F:\EMAILS\Enron_3\Testing\Spam\l20	1	CORRECTLY	SPAM
F:\EMAILS\Enron_3\Testing\Spam\k6	1	CORRECTLY	SPAM
F:\EMAILS\Enron_3\Testing\Ham\msg2482.eml	0	CORRECTLY	HAM

Obr. 7.6: Podrobný výpis výsledkov.

Kapitola 8

Testovanie aplikácie

Nasledujúca kapitola popisuje samotné testovanie implementovanej aplikácie. Kapitola podrobne popisuje množinu navrhovaných a vykonaných testov. Taktiež sú tu podrobne definované veličiny, ktoré boli počas testov skúmané. Samozrejme je tu obsiahnutý aj podrobný popis dát, na ktorých aplikácia bola testovaná.

8.1 Testované veličiny

Pre potreby testov v kapitole je nutné definovať si základné pojmy. Systémom spracované správy môžu byť na jeho výstupe ohodnotené jedným z nasledujúcich spôsobov:

- true positive (TP) - systémom správne rozoznaná správa z triedy spam
- false positive (FP) - správa z triedy spam označená systémom ako ham
- true negative (TN) - systémom správne rozoznaná správa z triedy ham
- false negative (FN) - správa z triedy ham systémom označená ako spam

V prípade, že máme e-mailové správy klasifikované do jednej zo štyroch vyššie uvedených skupín. Môžeme tak definovať veličiny popisujúce vlastnosti systému.

- Accuracy (správnost) - vyjadruje pomer všetkých správne označený ham a spam správ, ku všetkým správam ktoré systém spracoval. Ohodnocuje tak správnosť systému.
$$Accuracy = \frac{TN + TP}{TP + FP + TN + FN} \tag{8.1}$$
- Precision (presnosť) - vyjadruje pomer správ označených ako spam, ku všetkým správam označeným ako spam. Ohodnocuje tak schopnosť systému správne identifikovať spam.
$$Precision = \frac{TP}{TP + FN} \tag{8.2}$$
- Recall - vyjadruje pomer správ označených ako spam, ku správam označeným ako spam a správam, ktoré boli zo skupiny spam, ale systém ich označil ako ham. Určuje tak schopnosť systému rozpoznať spam.
$$Recall = \frac{TP}{TP + FP} \tag{8.3}$$

- Miss rate - je pomer správ z triedy spam označených systémom ako ham, ku všetkým správam spam a správam triedy spam označených ako ham. Vyjadruje tak množstvo správ, ktoré neboli správne identifikované ako spam.

$$Missrate = \frac{FP}{TP + FP} \tag{8.4}$$

- Error - je pomer všetkých nesprávne označených správ (správy typu ham systém vyhodnotil ako spam, a správy typu spam označil ako ham), ku všetkým správam, ktoré systém spracoval.

$$Error = \frac{FN + FP}{TP + FP + TN + FN} \tag{8.5}$$

8.2 Testovacie dáta

Dáta, na ktorých bola implementovaná aplikácia testovaná boli získané z korpusu Enron a SpamAssassin. Každá sada obsahuje iné množstvo správ typu ham, spam určených buď to na trénovanie aplikácie alebo na jej testovanie. Podrobné rozdelenie správ popisuje tabuľka 8.1.

Dátová sada	**Trénovanie**		**Testovanie**	
	ham	spam	ham	spam
Enron	3000	2000	1500	1500
SpamAssassin	2000	2000	500	500

Tabuľka 8.1: Tabuľka popisujúca testovacie sady dát.

Na takto pripravených dátach boli následne vykonané testy. Postup týchto testov bol rovnaký a je možné ho zhrnúť v nasledujúcich krokoch.

1. V prípade novej dátovej sady Zmazať databázu kandidátov a pokračuj krokom 2. inak krokom 3.
2. Pomocou dát z dátovej sady určených na trénovanie bol systém trénovaný a vytvorená tak nová databáza kandidátov.
3. Úprava nastavení pre potreby testu. Pri každom teste je v texte popísaná konfigurácia systému počas testu. V prípade zmeny testovacích dát pokračuj krokom 4. inak 5.
4. Výber žiadaných testovacích dát z dátovej sady na ktorú bol systém trénovaný.
5. Samotné testovanie a získanie výsledkov.

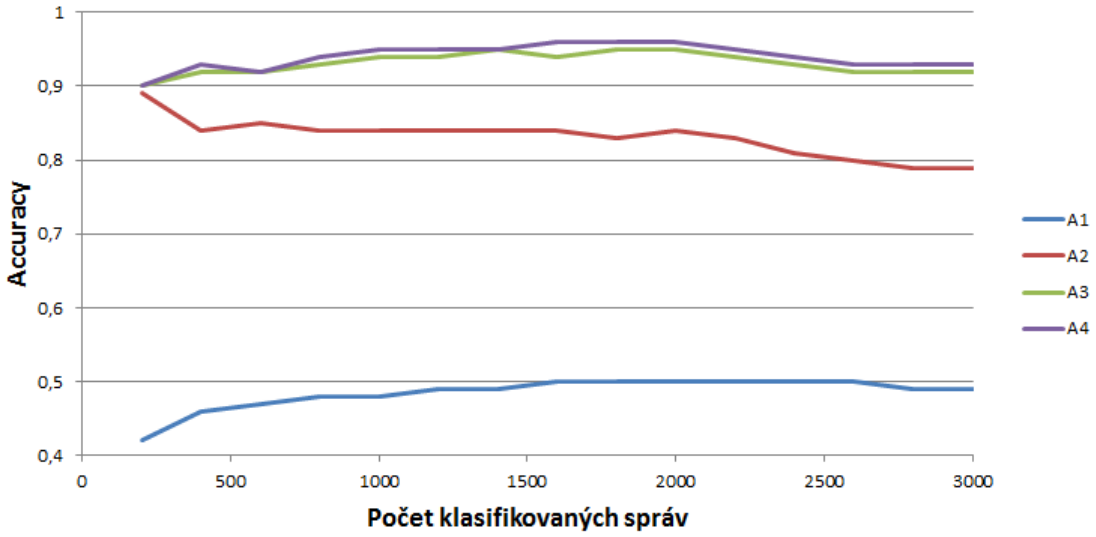
Pomocou takto stanoveného postupu boli vykonané rôzne testy, ich podrobnejší popis a výsledky sú zachytené v nasledujúcom texte.

8.3 Vykonané testy - korpus Enron

Test vplyvu spôsobu výberu kandidátov na lymfocyty.

V nasledujúcom teste bolo cieľom názorne ukázať vplyv spôsobu výberu kandidátov na výslednú kvalitu výsledkov systému. V teste boli tvorené dve množiny lymfocytov (ham, spam) a boli vyskúšané štyri spôsoby voľby kandidátov, do týchto množín.

- (A1) Náhodný výber lymfocytov - tento výber bol v réžii databázy bez akéhokoľvek radenia výsledkov či určovania podmienok, ktoré by museli kandidáti spĺňať.
- (A2) Výber na základe pravdepodobnosti výskytu v spame (hame) - v tomto prípade bola množina kandidátov usporiadaná podľa pravdepodobnosti výskytu v spame (hame) od najväčšej po najmenšiu a z takto usporiadanej množiny bolo vybraných n kandidátov.
- (A3) Výber pomocou hodnoty spam_w (ham_w) - tu boli opäť kandidáti usporiadaní zostupne, no podľa hodnoty spam_w (ham_w). Z takto usporiadanej množiny bolo vybraných n kandidátov.
- (A4) Výber pomocou hodnoty spam_w (ham_w) s použitím pravdepodobnosti výskytu v spame (hame) - v tomto prípade boli kandidáti vyberaní podobne ako v predchádzajúcom bode. No bolo pridané obmedzenie že, museli spĺňať pravdepodobnosť výskytu v spame väčšiu ako 80%, prípadne menšiu ako 40%.



Obr. 8.1: Graf úspešnosti systému pri rôznom spôsobe výberu lymfocytov.

Podľa výsledkov v tabuľke 8.2 či grafe 8.1 je možné usúdiť, že správny výber voľby kandidátnych slov na lymfocyty má pomerne značný význam na výsledné správanie systému. Aj na základe tohto testu je v aplikácii použitá posledná voľba výberu - (A4).

Test	Accuracy	Recall	Precision	Miss rate	Error	Spam score
A1	0,4847	0,0289	0,3643	0,9711	0,5153	0,0943
A2	0,8306	0,7224	0,9255	0,2776	0,1694	0,7251
A3	0,9314	0,9559	0,9122	0,0441	0,0686	0,9583
A4	0,9403	0,9564	0,9279	0,0436	0,0597	0,9537

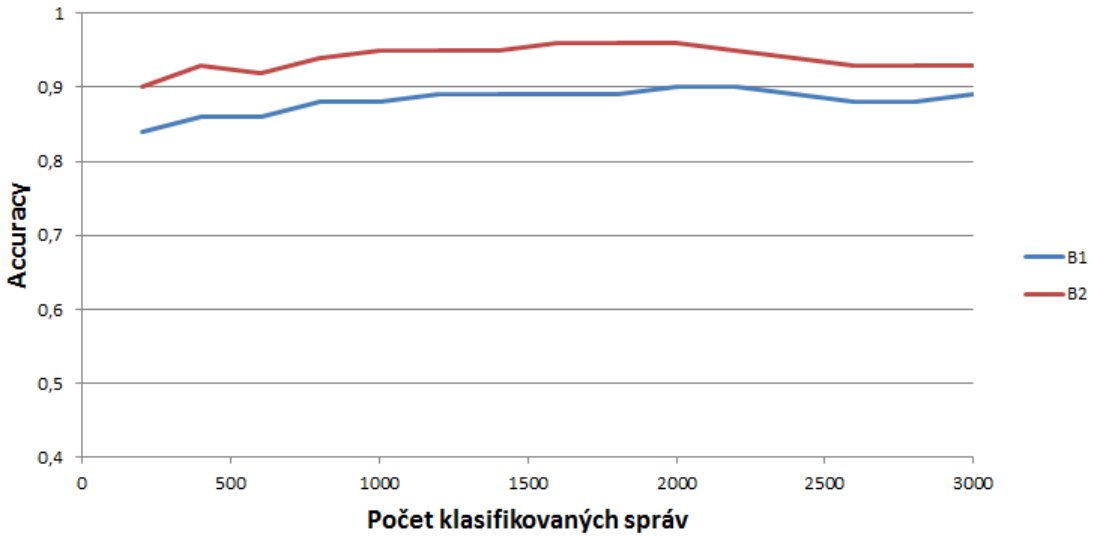
Tabuľka 8.2: Tabuľka priemerných hodnôt nameraných počas testu spôsobov výberu lymfocytov.

Test vplyvu rozdelenia množiny lymfocytov.

Týmto testom bol skúmaný vplyv rozdelenia množiny lymfocytov na dve množiny a teda ham a spam lymfocytov.

- (B1) Jedna množina - v tomto prípade bola použitá jediná množina lymfocytov v ktorej sa nachádzali kandidáti s najväčšou hodnotou spam_w spĺňajúci podmienku, že pravdepodobnosť ich výskytu v spame je väčšia ako 80%.
- (B2) Dve oddelené množiny - tu boli vytvorené dve množiny. Prvá, tvorená slovami s najväčším spam_w a zároveň mali pravdepodobnosť výskytu v spame väčšiu ako 80%, bola množina spam lymfocytov. Druhá, naopak tvorená slovami s najväčším ham_w a pravdepodobnosťou výskytu v spame menšou ako 40% tvorila množinu ham lymfocytov.

Poznámka: V obidvoch prípadoch bol vytvorený rovnaký počet lymfocytov. Ak v prvom teste (B1) vznikla množina s počtom 200 lymfocytov, tak v druhom (B2) teste vznikli dve množiny o veľkosti 100 lymfocytov.



Obr. 8.2: Graf úspešnosti systému pri použití jednej a dvoch množín lymfocytov.

Ako je možné vidieť z 8.3 či grafu 8.2 rozdelenie množiny lymfocytov na spam a ham lymfocyty prinieslo zvýšenie správnosti a presnosti systému. Na základe tohto testu je v aplikácii použité ako ďalšie rozšírenie práve delenie množiny lymfocytov.

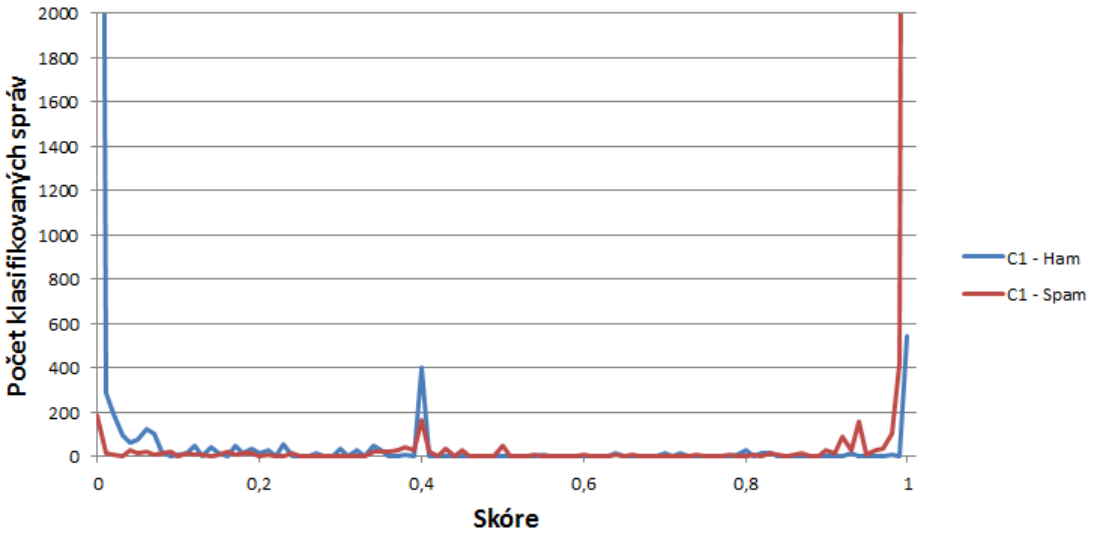
Test	Accuracy	Recall	Precision	Miss rate	Error	Spam score
B1	0,8809	0,9799	0,8192	0,0201	0,1191	0,9738
B2	0,9403	0,9564	0,9279	0,0436	0,0597	0,9537

Tabuľka 8.3: Tabuľka priemerných hodnôt nameraných počas testu delenia množiny lymfocytov.

Vplyv použitého algoritmu na skóre.

Tento test poukazuje na vplyv použitého algoritmu na rozloženie skóre na množine správ. Boli testované nasledujúce kombinácie konfigurácií.

- (C1) Použitá množina jednoslovných lymfocytov.
- (C2) Kombinácia množín jednoslovných a dvojslovných lymfocytov doplnená o použitie pamäťových buniek.

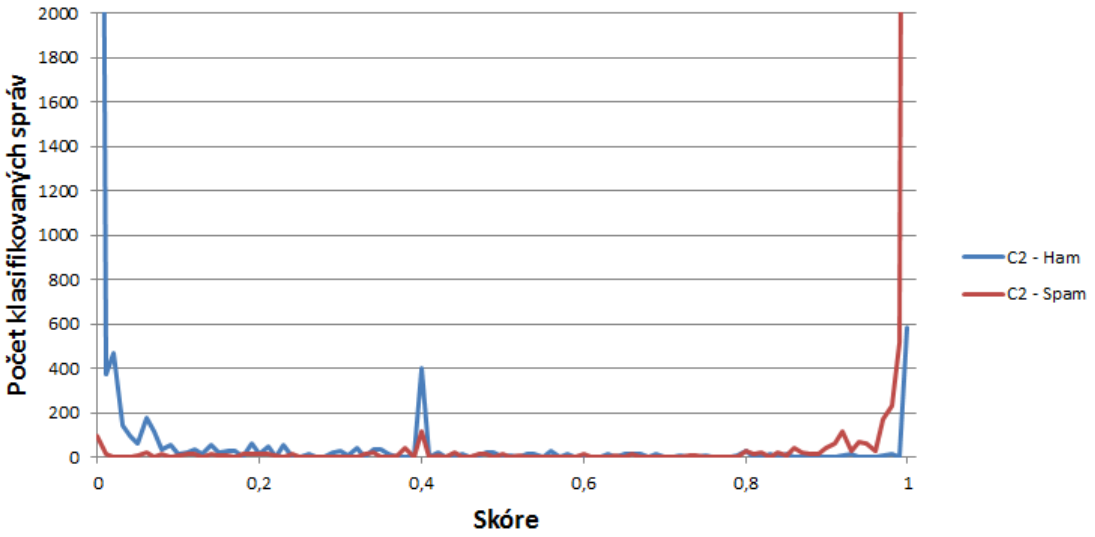


Obr. 8.3: Graf rozloženia skóre pri teste C1.

Test	Accuracy	Recall	Precision	Miss rate	Error	Spam score
C1	0,9312	0,9329	0,9315	0,0671	0,0688	0,9361
C2	0,9403	0,9564	0,9279	0,0436	0,0597	0,9537

Tabuľka 8.4: Tabuľka priemerných hodnôt nameraných počas testov C1 a C2.

V teste sa potvrdilo, že použitie dvojslovných lymfocytov v spojení s pamäťovými bunkami vylepšuje správanie systému. Ako je možné vidieť na grafoch 8.9 a 8.10 v prípade použitia dvojslovných lymfocytov a pamäťových buniek je ohodnotenie spamu lepšie ako v prípade použitia jednoslovných lymfocytov. Toto sa prejaví aj na celkovej úspešnosti systému, ktorú je možné vidieť na grafe 8.11 či ostatných výsledkoch v tabuľke 8.4.



Obr. 8.4: Graf rozloženia skóre pri teste C2.

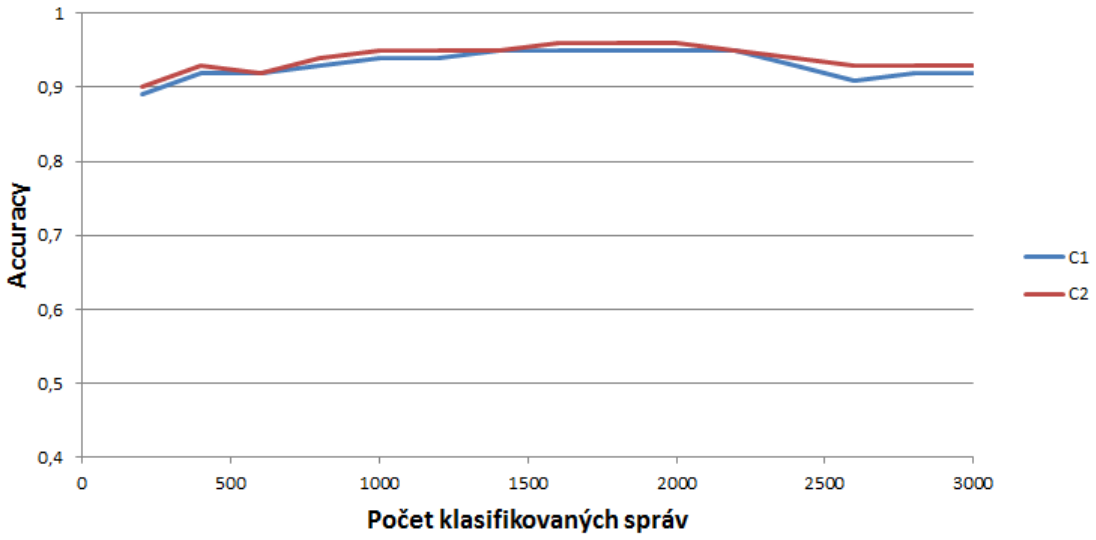
Vplyv veľkosti množiny lymfocytov na systém.

V tomto teste je snaha prezentovať vplyv veľkosti množiny lymfocytov na samotnú funkčnosť systému. V priebehu celého testu bolo zakázané používanie pamäťových buniek. Množina lymfocytov bola postupne navyšovaná. Jej minimálna hodnota bola 100 lymfocytov a maximálna potom 900 lymfocytov. Výsledky testu je možné nájsť v grafe 8.6 a v tabuľke 8.5.

Test	Accuracy	Recall	Precision	Miss rate	Error	Spam score
100	0,9279	0,9183	0,9377	0,0817	0,0721	0,9187
200	0,9356	0,9422	0,9314	0,0578	0,0644	0,9402
300	0,9426	0,9570	0,9316	0,0430	0,0574	0,9505
400	0,9462	0,9651	0,9314	0,0349	0,0538	0,9571
500	0,9463	0,9661	0,9306	0,0339	0,0537	0,9595
600	0,9490	0,9716	0,9310	0,0284	0,0510	0,9632
700	0,9497	0,9742	0,9301	0,0258	0,0503	0,9655
800	0,9482	0,9745	0,9272	0,0255	0,0518	0,9645
900	0,9463	0,9745	0,9240	0,0255	0,0537	0,9585

Tabuľka 8.5: Tabuľka priemerných hodnôt nameraných pri rôznom počte lymfocytov.

Z výsledkov 8.5 vyplýva, že navyšovaním počtu lymfocytov sa zvyšuje aj úspešnosť systému odhaliť nevyžiadanú poštu. Taktiež je možné pozorovať, že tento nárast je len do určitej hranice počtu lymfocytov. Taktiež je nutné počítať s faktom, že čím väčší je počet lymfocytov, tým pomalšie je správa spracovaná.



Obr. 8.5: Graf úspešnosti systému pri testoch C1 a C2.

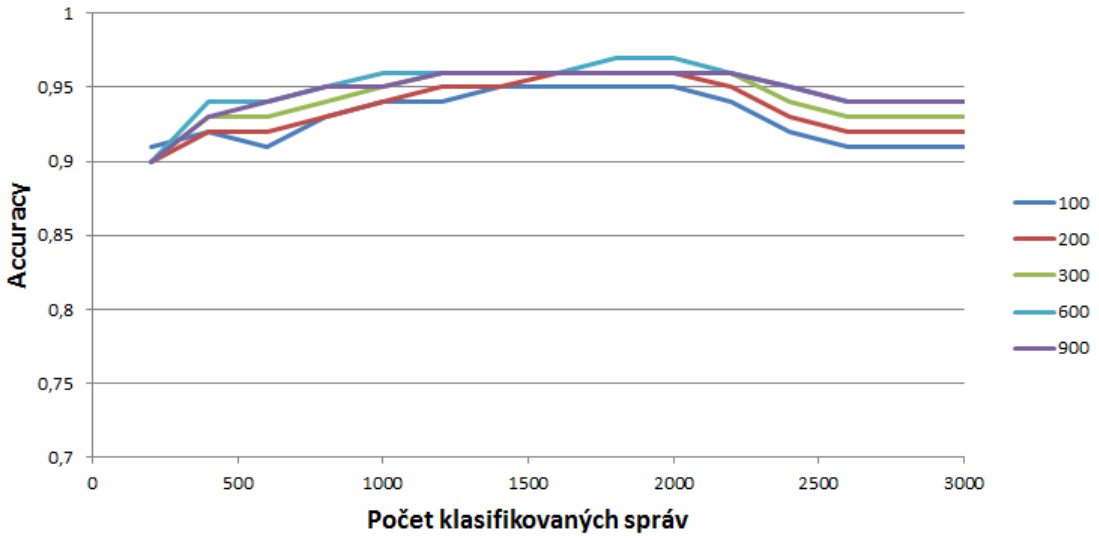
8.4 Vykonané testy - korpus SpamAssassin

Test vplyvu spôsobu výberu kandidátov na lymfocyty.

- (A1) Náhodný výber lymfocytov - tento výber bol v réžii databázy bez akéhokoľvek radenia výsledkov či určovania podmienok, ktoré by museli kandidáti spĺňať.
- (A2) Výber na základe pravdepodobnosti výskytu v spame (hame) - v tomto prípade bola množina kandidátov usporiadaná podľa pravdepodobnosti výskytu v spame (hame) od najväčšej po najmenšiu a z takto usporiadanej množiny bolo vybraných n kandidátov.
- (A3) Výber pomocou hodnoty spam_w (ham_w) - tu boli opäť kandidáti usporiadaní zostupne, no podľa hodnoty spam_w (ham_w). Z takto usporiadanej množiny bolo vybraných n kandidátov.
- (A4) Výber pomocou hodnoty spam_w (ham_w) s použitím pravdepodobnosti výskytu v spame (hame) - v tomto prípade boli kandidáti vyberaní podobne ako v predchádzajúcom bode. No bolo pridané obmedzenie že museli spĺňať pravdepodobnosť výskytu v spame väčšiu ako 80% prípadne menšiu ako 40%.

Test	Accuracy	Recall	Precision	Miss rate	Error	Spam score
A1	0,5930	0,1861	0,2861	0,8139	0,4070	0,2109
A2	0,4127	0,7797	0,4497	0,2203	0,5873	0,8890
A3	0,9276	0,9025	0,9506	0,0975	0,0724	0,9290
A4	0,9263	0,8950	0,9551	0,105	0,0737	0,9184

Tabuľka 8.6: Tabuľka priemerných hodnôt nameraných počas testu spôsobov výberu lymfocytov.



Obr. 8.6: Graf úspešnosti systému pri rôznom počte lymfocytov.

Aj v prípade tohto testu na sade SpamAssasin vychádzajú veľmi podobné výsledky 8.7, 8.6 ako v prípade sady Enron. A preto vyzerá byť najvhodnejšou voľbou na výber kandidátov voľba (A4).

Test vplyvu rozdelenia množiny lymfocytov.

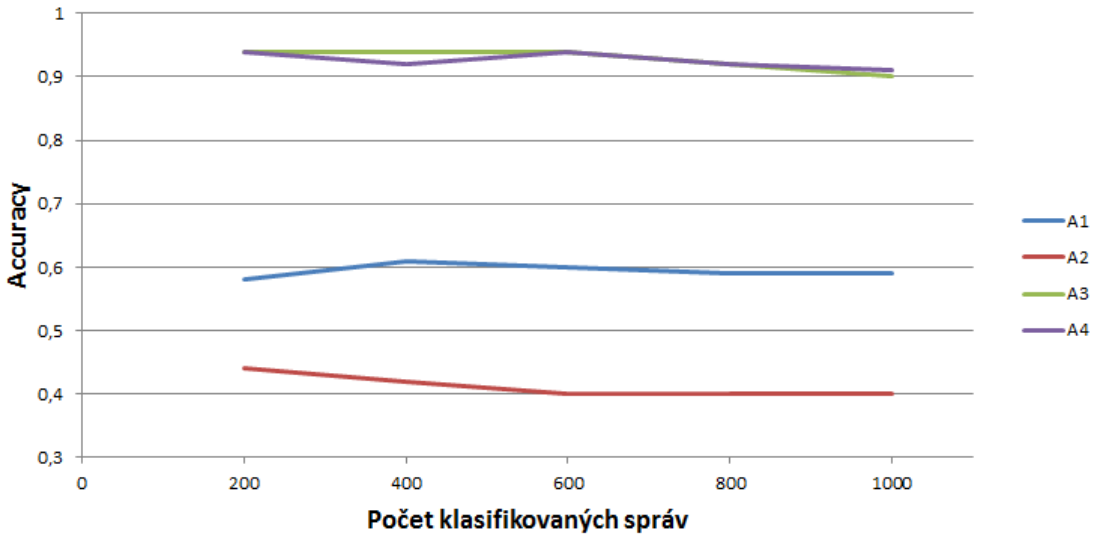
- (B1) Jedna množina - v tomto prípade bola použitá jediná množina lymfocytov v ktorej sa nachádzali kandidáti s najväčšou hodnotou spam_w spĺňajúci podmienku, že pravdepodobnosť ich výskytu v spame je väčšia ako 80%.
- (B2) Dve oddelené množiny - tu boli vytvorené dve množiny. Prvá, tvorená slovami s najväčším spam_w a zároveň mali pravdepodobnosť výskytu v spame väčšiu ako 80%, bola množina spam lymfocytov. Druhá, naopak tvorená slovami s najväčším ham_w a pravdepodobnosťou výskytu v spame menšou ako 40% tvorila množinu ham lymfocytov.

Poznámka: V obidvoch prípadoch bol vytvorený rovnaký počet lymfocytov. Ak v prvom teste (B1) vznikla množina s počtom 200 lymfocytov, tak v druhom (B2) teste vznikli dve množiny o veľkosti 100 lymfocytov.

Test	Accuracy	Recall	Precision	Miss rate	Error	Spam score
B1	0,8735	0,8328	0,9066	0,1672	0,1265	0,8990
B2	0,9263	0,8950	0,9551	0,1050	0,0737	0,9340

Tabuľka 8.7: Tabuľka priemerných hodnôt nameraných počas testu delenia množiny lymfocytov.

Ako je možné vidieť z 8.7 či v grafe 8.8 aj v tomto prípade podobne ako v rovnakom teste na sade Enron sú dosahované lepšie výsledky s použitím ham a spam množím lymfocytov.



Obr. 8.7: Graf úspešnosti systému pri rôznom spôsobe výberu lymfocytov.

Vplyv použitého algoritmu na skóre.

- (C1) Použitá množina jednoslovných lymfocytov.
- (C2) Kombinácia množín jednoslovných a dvojslovných lymfocytov doplnená o použitie pamäťových buniek.

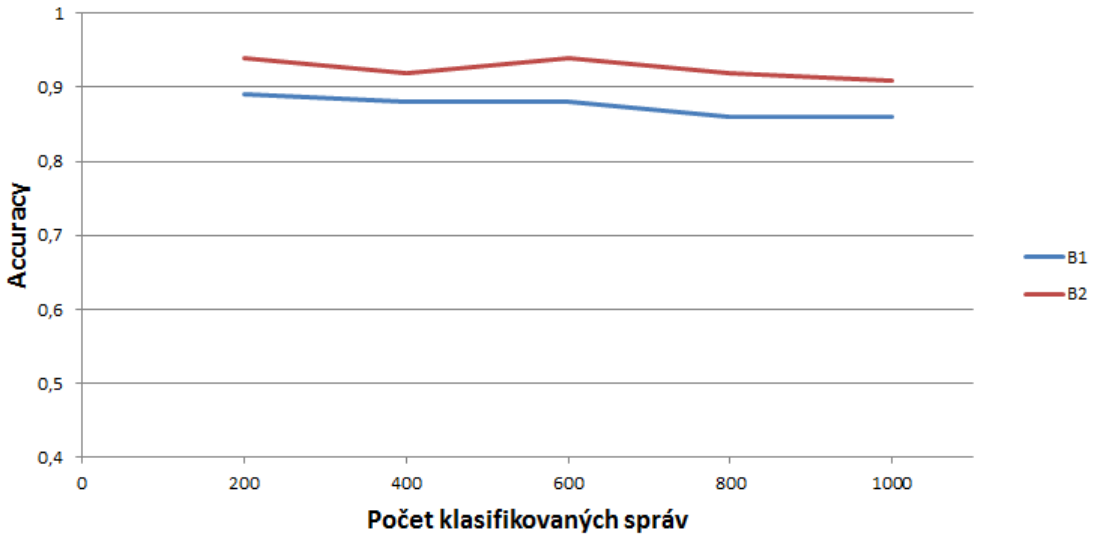
Test	Accuracy	Recall	Precision	Miss rate	Error	Spam score
C1	0,9545	0,9271	0,9810	0,0729	0,0455	0,9635
C2	0,9081	0,9373	0,8857	0,0627	0,0919	0,9102

Tabuľka 8.8: Tabuľka priemerných hodnôt nameraných počas testov C1 a C2.

V tomto teste podľa výsledkov 8.11, 8.8 na rozdiel od testu v prípade sady Enron boli výsledky lepšie v prípade voľby (C1). Aj to môže napovedať, že pre rôzne sady správ je vhodná rôzna konfigurácia systému.

Vplyv veľkosti množiny lymfocytov na systém.

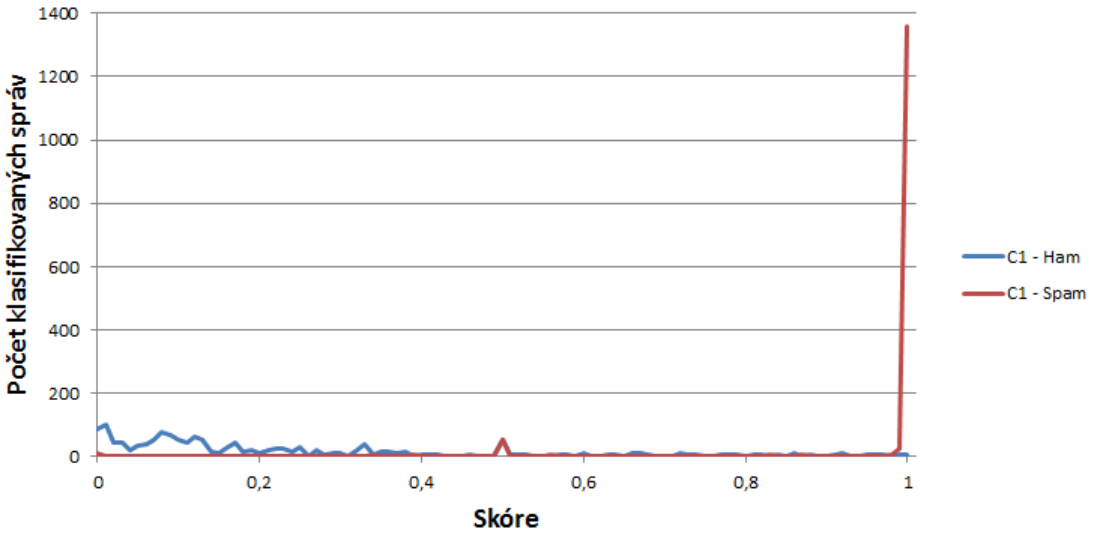
V tomto prípade, podobne ako v predchádzajúcom teste, výsledky 8.9, 8.12 vychádzajú pomerne rozdielne od rovnakého testu na sade Enron. V tomto prípade najlepšie výsledky dosahujúca konfigurávia s množinou o veľkosti 100 lymfocytov.



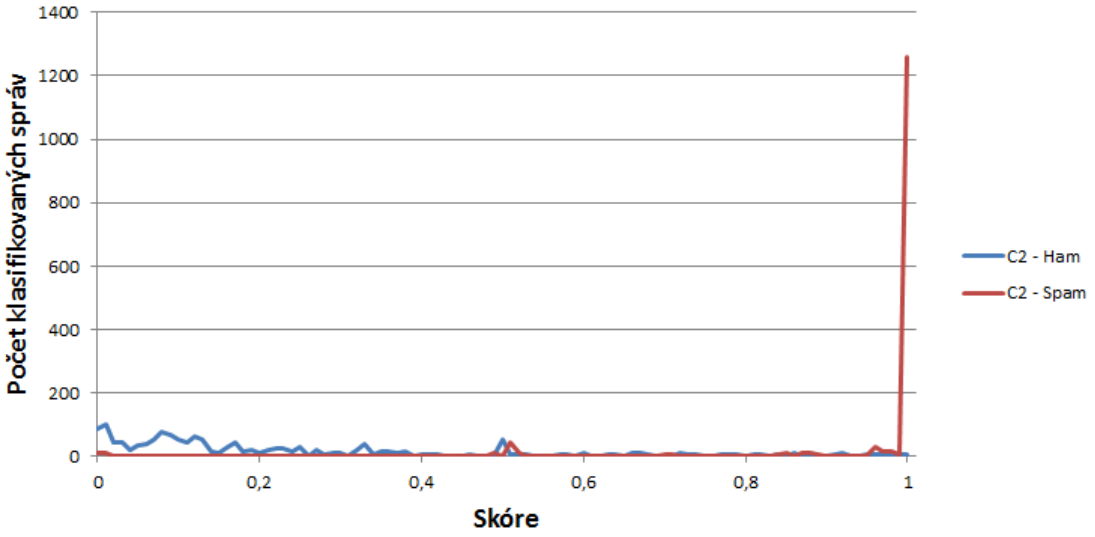
Obr. 8.8: Graf úspešnosti systému pri použití jednej a dvoch množín lymfocytov.

Test	Accuracy	Recall	Precision	Miss rate	Error	Spam score
100	0,9026	0,9164	0,8921	0,0836	0,0974	0,8579
200	0,8570	0,9293	0,8122	0,0707	0,1430	0,8377
300	0,8395	0,9436	0,7815	0,0564	0,1605	0,7768
400	0,8294	0,9432	0,7689	0,0568	0,1707	0,7982
500	0,8132	0,9436	0,7491	0,0564	0,1868	0,7198
600	0,8047	0,9401	0,7406	0,0599	0,1953	0,7328
700	0,7962	0,9383	0,7317	0,0617	0,2038	0,7425
800	0,7914	0,9379	0,7267	0,0621	0,2086	0,6503
900	0,7773	0,9101	0,7202	0,0899	0,2227	0,6513

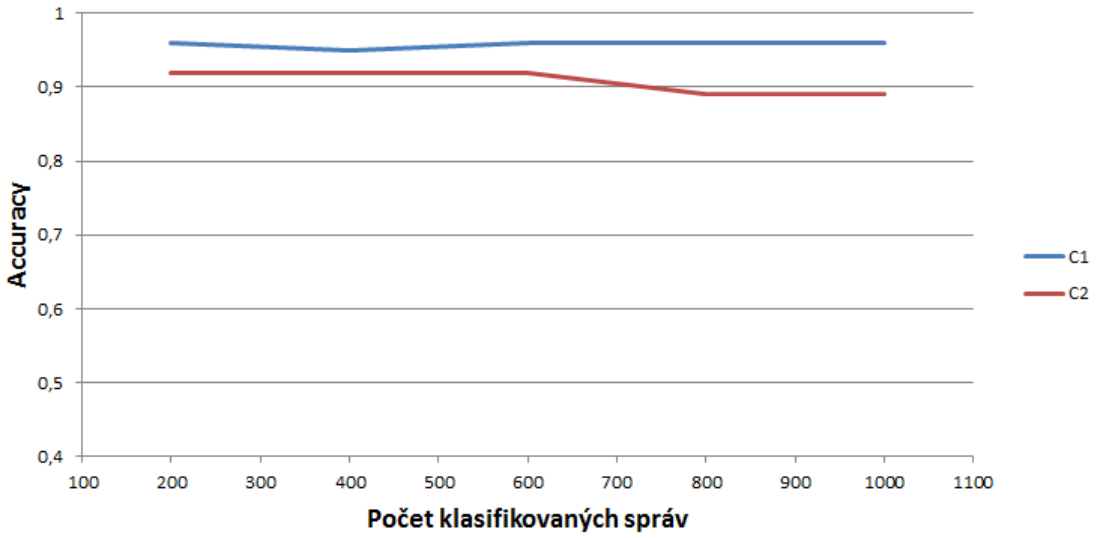
Tabuľka 8.9: Tabuľka priemerných hodnôt nameraných pri rôznom počte lymfocytov.



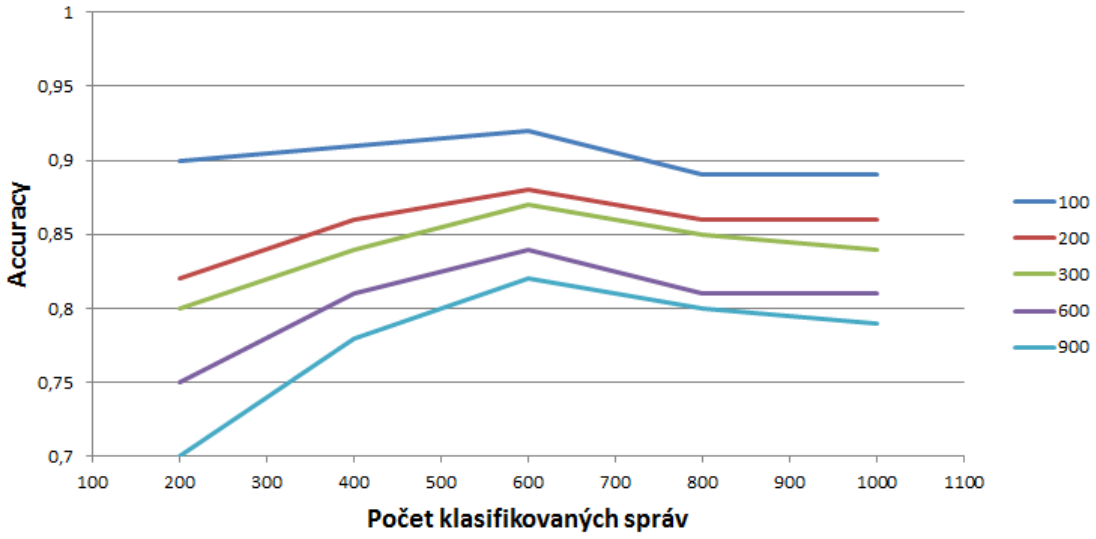
Obr. 8.9: Graf rozloženia skóre pri teste C1.



Obr. 8.10: Graf rozloženia skóre pri teste C2.



Obr. 8.11: Graf úspešnosti systému pri testoch C1 a C2.



Obr. 8.12: Graf úspešnosti systému pri rôznom počte lymfocytov.

Kapitola 9

Zhodnotenie

Po zhodnotení testov vykonaných na sadách e-mailov Enron a SpamAssassin je možné vyvodiť viacero záverov. V prípade rozšírenia o priamy výber lymfocytov na základe ich vlastností sa na obidvoch testovacích sadách dá tvrdiť, že aplikácia dosahovala lepšie výsledky ako boli dosahované v referenčnej práci [8]. Podobné tvrdenie platí aj v prípadee použitia rozdelenej množiny lymcofytov na ham a spam lymfocyty. Ak sa však pozrieme na testy v prípade použitia jednoslovných lymfocytov a jednoslovných lymfocytov v spojení s dvojslovnými lymfocytmy je možné tvrdiť, že v rámci jednej sady boli dosahované lepšie výsledky ako v rámci druhej sady.

Z celkového pohľadu je možné povedať, že aplikácia oproti referenčnému článku, v prípade použitia kombinácie všetkých vylepšení dosahuje lepšie výsledky a to aj na sadách s menším počtom testovacích e-mailov.

9.1 Možné kroky do budúcna

V rámci rozšírenia aplikácie sa ponúka viacero možností. Je tu možnosť použiť iné vzťahy pri výpočte dielčích pravdepodobností, taktiež je tu priestor k použitiu iného váženého vzťahu k výpočtu výslednej pravdepodobnosti, že správa je spam. Samozrejme tu existuje možnosť získať a následne použiť iné štatistické informácie o slovách použitých na vytváranie lymfocytov.

Z aplikačného hľadiska tu potom vidím rozšírenie napríklad v doplnku spolupracujúcom s e-mailovými klientami a beh aplikácie v reálnych podmienkach. S tým je samozrejme spojená možnosť doplnenia spätnej väzby od užívateľa.

Kapitola 10

Záver

Cieľom práce bolo v jej úvode čitateľa oboznámiť s rôznymi viac či menej známymi metódami, ktoré sa v súčasnosti využívajú na boj proti nevyžiadanej pošte.

Druhá kapitola preto robila prierez takýmito metódami aj so stručným princípom ich fungovania. Oboznámili sme sa s metódami, ktoré boli viac zamerané na užívateľa a jeho gramotnosť v tejto oblasti, no nezabudli sme ani na metódy založené na rôznych zoznamoch či na princípe spracovávania obsahu správ. Záver druhej kapitoly sa potom niesol v duchu metód pracujúcich na rôznych iných princípoch.

Tretia kapitola potom vo svojej prvej časti podrobnejšie čitateľa oboznámila s princípom Bayesovej metódy. A v časti druhej boli popísané rôzne úpravy tejto metódy s cieľom dosiahnuť v praxi čo najlepšie výsledky v boji s nevyžiadanou poštou. V práci tiež bolo spomenuté spojenie Bayesovej metódy s UIS, pričom výsledná metóda spája pozitíva obidvoch prístupov. Následne boli popísané vlastnosti biologického imunitného systému a ním inšpirovaného umelého imunitného systému.

Z týchto poznatkov je potom v práci ťažené pri návrhu vlastného systému využívaného k identifikovaniu nevyžiadanej pošty. Tento návrh bol potom implementovaný v jazyku C#. Vzniknutá aplikácia bola testovaná na dvoch sadách e-mailových správ. Na každej sade bola vykonaná séria testov, zameraných na otestovanie mnou navrhovaných vylepšení. Z testov vyplynulo, že všetky tri implementované vylepšenia sú pre aplikáciu prínosné.

Už počas práce vznikali rôzne nápady na vylepšenie aplikácie. Preto je tu do budúcnosti priestor na ich možné zapracovanie.

Literatúra

[1] Adámek, M.: *Spam -jak nepřivolávat, nepřijímat a nerozesílat nevyžádanou poštu.* Průvodce (Grada), Grada, 2009, ISBN 9788024726380.
URL http://books.google.cz/books?id=d-ZMVJEGV0cC

[2] Buc, M.: *Imunológia.* Veda, Bratislava, 2001, ISBN 80-224-0667-8.

[3] Chen, C.-J.; Cui, Y.-D.; Xie, T.: Study of Spam Short Message Filtering Based on Features Selection of Key Words. In *Pattern Recognition, Communications in Computer and Information Science*, ročník 321, editace C.-L. Liu; C. Zhang; L. Wang, Springer Berlin Heidelberg, 2012, ISBN 978-3-642-33505-1, s. 646–654, doi:10.1007/978-3-642-33506-8˙79.
URL http://dx.doi.org/10.1007/978-3-642-33506-8_79

[4] Duan, Z.; Dong, Y.; Gopalan, K.: DMTP: Controlling spam through message delivery differentiation. *Computer Networks*, ročník 51, č. 10, 2007: s. 2616 – 2630, ISSN 1389-1286, doi:10.1016/j.comnet.2006.11.015.
URL http://www.sciencedirect.com/science/article/pii/S1389128606003471

[5] Eric W. Weisstein: Bayes' Theorem [online]. [cit. 2012-12-20].
URL http://mathworld.wolfram.com/BayesTheorem.html

[6] Gudkova, D.; Shcherbakova, T.: Spam in November 2012 [online]. 2012-12-19 [cit. 2012-12-28].
URL
http://www.securelist.com/en/analysis/204792258/Spam_in_November_2012

[7] Jaromír Baštinec: Statistika, operační výzkum, stochastické procesy. 2010-10-11 [cit. 2012-12-28].
URL http://mathworld.wolfram.com/BayesTheorem.html

[8] Luo, Q.; Liu, B.; Yan, J.; aj.: Research of a Spam Filtering Algorithm Based on Na x0EF;ve Bayes and AIS. In *Computational and Information Sciences (ICCIS), 2010 International Conference on*, dec. 2010, s. 152 –155, doi:10.1109/ICCIS.2010.43.

[9] Sudhakar, V.; Rao, C.; Somayajula, S.: Bayesian Spam Filtering using Statistical Data Compression. *International Journal of Computer Science and Information Security*, ročník 9, 2011: s. 157–159, ISSN 1947-5500.
URL http://www.docstoc.com/docs/105065051/
Bayesian-Spam-Filtering-using-Statistical-Data-Compression

[10] WWW stránky: 10 Minute Mail [online]. [cit. 2012-12-20].
URL http://10minutemail.com/10MinuteMail/index.html

[11] WWW stránky: Anti-spam techniques [online]. [cit. 2012-12-20].
URL http://en.wikipedia.org/wiki/Anti-spam_techniques

[12] WWW stránky: Bayes' theorem [online]. [cit. 2012-12-20].
URL http://en.wikipedia.org/wiki/Bayes'_theorem

[13] WWW stránky: Komodia [online]. [cit. 2012-12-20].
URL http://www.komodia.com/contact-us

[14] WWW stránky: SORBS (Spam and Open-Relay Blocking System) [online]. [cit. 2012-12-20].
URL http://www.sorbs.net/lookup.shtml

[15] WWW stránky: Tagged Message Delivery Agent (TMDA) [online]. [cit. 2012-12-20].
URL http://tmda.net/

[16] WWW stránky: Types of Spam Filters [online]. [cit. 2012-12-20].
URL http://www.clearmymail.com/guides/spam_filter_types.aspx

[17] WWW stránky: Umelé imunitné systémy [online]. [cit. 2012-12-20].
URL http://alife.tuke.sk

[18] WWW stránky: The Immune System (Structure and Function) (Nursing) [online]. [cit. 2012-12-22].
URL http://http://what-when-how.com/nursing/the-immune-system-structure-and-function-nursing-part-1/

[19] WWW stránky: Spam Statistics and Facts [online]. [cit. 2012-12-28].
URL http://www.spamlaws.com/spam-stats.html

[20] Yin, H.; Chaoyang, Z.: An Improved Bayesian Algorithm for Filtering Spam E-Mail. In *Intelligence Information Processing and Trusted Computing (IPTC), 2011 2nd International Symposium on*, oct. 2011, s. 87 –90, doi:10.1109/IPTC.2011.29.

Dodatok A

Obsah CD

- aplikácia v podobe setup.exe
- sady e-mailov
- zdrojové kódy aplikácie
- zdrojové kódy textu práce