

# **Určování autorství**

Authorship determination

Bc. Kamil Řezníček



Univerzita Tomáše Bati ve Zlíně  
Fakulta aplikované informatiky  
akademický rok: 2012/2013

## **ZADÁNÍ DIPLOMOVÉ PRÁCE**

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení: **Bc. Kamil ŘEZNÍČEK**  
Osobní číslo: **A10394**  
Studijní program: **N3902 Inženýrská informatika**  
Studijní obor: **Informační technologie**  
Forma studia: **kombinovaná**

Téma práce: **Určování autorství**

Zásady pro vypracování:

1. Seznamte se s různými metodami umělé inteligence, především pro oblast klasifikace.
2. Vhodně vyberte učící a testovací data pro určování autorství textů.
3. Vhodně vyberte metodu pro klasifikování autorů.
4. Implementujte metody do aplikace.
5. Provedte experimenty úspěšnosti určování autorství.

Rozsah diplomové práce:

Rozsah příloh:

Forma zpracování diplomové práce: tištěná/elektronická

Seznam odborné literatury:


1. VAŠÁK, Pavel. Metody určování autorství. 1. vyd. Praha: Academia, 1980.
2. STAMATATOS, Efstathios. A Survey of Modern Authorship Attribution Methods. Journal of the American Society for Information Science and Technology [online]. 2009, roč. 60(č. 3), 538-556. Přepis dostupný z: <http://www.icsd.aegean.gr/lecturers/stamatatos/papers/survey.pdf>
3. KOPPEL, Moshe, Jonathan SCHLER a Shlomo ARGAMON. Computational methods in authorship attribution. Journal of the American Society for Information Science and Technology [online]. 2009, roč. 60(č. 1), 9-26. Přepis dostupný z: <http://u.cs.biu.ac.il/~koppel/papers/authorship-JASIST-final.pdf>
4. STAMATATOS, Efstathios, George KOKKINAKIS a Nikos FAKOTAKIS. Automatic text categorization in terms of genre and author. Journal Computational Linguistics [online]. 2000, roč. 26(č. 4), 471-495. Dostupné z: <http://acl.ldc.upenn.edu/J/J00/J00-4001.pdf>
5. ZELINKA, I., OPLATKOVÁ, Z., OŠMERA, P., ŠEDA, M., VČELAŘ, F. Evoluční výpočetní techniky – principy a aplikace. BEN – technická literatura, Praha, 2008, ISBN 80-7300-218-3.
6. ŠNOREK M., JIŘINA M.: Neuronové sítě a neuropočítače, ČVUT, 1996, ISBN 80-01-01455-X.
7. BÍLA J.: Umělá inteligence a neuronové sítě v aplikacích, ČVUT, 1996, ISBN 80-01-01275-1.
8. ZELINKA I.: Umělá inteligence I, VUT Brno, 1998, ISBN 80-214-1163-5.

Vedoucí diplomové práce: Ing. Zuzana Komínková Oplatková, Ph.D.  
Ústav informatiky a umělé inteligence

Datum zadání diplomové práce: 22. února 2013

Termín odevzdání diplomové práce: 22. května 2013

Ve Zlíně dne 22. února 2013

  
prof. Ing. Vladimír Vašek, CSc.  
děkan



  
doc. Mgr. Roman Jašek, Ph.D.  
ředitel ústavu

## ABSTRAKT

Tato diplomová práce se zabývá problematikou určování autorství s využitím umělých neuronových sítí. Cílem práce je představit existující a používané techniky, tyto techniky implementovat a experimenty ověřit jejich úspěšnost určování autorství.

Implementované algoritmy jsou testovány na textech v anglickém jazyce, systém je však navržen tak, aby při vypuštění či nahrazení jazykově závislých komponent bylo možné pracovat s dokumenty v libovolném jazyce.

Teoretická část práce seznamuje čtenáře s problematikou určování autorství, její historií a základními metodami. Rovněž je podán pohled na neuronové sítě a jejich struktury. Praktická část zahrnuje implementaci metod a neuronové sítě s dopředným šířením signálu a seznamuje s výsledky testování.

Klíčová slova:

určení autora, rysy autora, stylometrie, umělá neuronová síť

## ABSTRACT

The present study deals with the problem of determining authorship using neural networks. The aim of the thesis is to describe techniques solving the problem in practice and show their success rate based on experiments conducted in the work.

The implemented algorithms are tested on English texts; the replacement of language specific components allows the correct functionality of the program for all languages.

The theoretical part introduces the reader the problem of determining authorship, its history and basic characteristics, together with the analysis of the neural networks and their structure. The practical part contains the implemented methods and feed-forward neural networks and presents the results from the conducted experiments.

Keywords:

authorship determination, authorship attribution, stylometry, artificial neural network

Rád bych poděkoval Ing. Zuzaně Komínkové Oplatkové, Ph.D., za ochotu, cenné rady a trpělivost při vedení mé práce. Děkuji svým nejbližším za jejich oporu a pomoc.

**Prohlašuji, že**

- beru na vědomí, že odevzdáním diplomové/bakalářské práce souhlasím se zveřejněním své práce podle zákona č. 111/1998 Sb. o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších právních předpisů, bez ohledu na výsledek obhajoby;
- beru na vědomí, že diplomová/bakalářská práce bude uložena v elektronické podobě v univerzitním informačním systému dostupná k prezenčnímu nahlédnutí, že jeden výtisk diplomové/bakalářské práce bude uložen v příruční knihovně Fakulty aplikované informatiky Univerzity Tomáše Bati ve Zlíně a jeden výtisk bude uložen u vedoucího práce;
- byl/a jsem seznámen/a s tím, že na moji diplomovou/bakalářskou práci se plně vztahuje zákon č. 121/2000 Sb. o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon) ve znění pozdějších právních předpisů, zejm. § 35 odst. 3;
- beru na vědomí, že podle § 60 odst. 1 autorského zákona má UTB ve Zlíně právo na uzavření licenční smlouvy o užití školního díla v rozsahu § 12 odst. 4 autorského zákona;
- beru na vědomí, že podle § 60 odst. 2 a 3 autorského zákona mohu užít své dílo – diplomovou/bakalářskou práci nebo poskytnout licenci k jejímu využití jen s předchozím písemným souhlasem Univerzity Tomáše Bati ve Zlíně, která je oprávněna v takovém případě ode mne požadovat přiměřený příspěvek na úhradu nákladů, které byly Univerzitou Tomáše Bati ve Zlíně na vytvoření díla vynaloženy (až do jejich skutečné výše);
- beru na vědomí, že pokud bylo k vypracování diplomové/bakalářské práce využito softwaru poskytnutého Univerzitou Tomáše Bati ve Zlíně nebo jinými subjekty pouze ke studijním a výzkumným účelům (tedy pouze k nekomerčnímu využití), nelze výsledky diplomové/bakalářské práce využít ke komerčním účelům;
- beru na vědomí, že pokud je výstupem diplomové/bakalářské práce jakýkoliv softwarový produkt, považují se za součást práce rovněž i zdrojové kódy, popř. soubory, ze kterých se projekt skládá. Neodevzdání této součásti může být důvodem k neobhájení práce.

**Prohlašuji,**

- že jsem na diplomové práci pracoval samostatně a použitou literaturu jsem citoval. V případě publikace výsledků budu uveden jako spoluautor.
- že odevzdaná verze diplomové práce a verze elektronická nahraná do IS/STAG jsou totožné.

Ve Zlíně

.....  
podpis diplomanta

**OBSAH**

<b>ÚVOD.....</b>	<b>9</b>
<b>I TEORETICKÁ ČÁST.....</b>	<b>10</b>
<b>1 URČOVÁNÍ AUTORSTVÍ .....</b>	<b>11</b>
1.1 VÝVOJ.....	11
1.2 POUŽÍVANÉ METODY .....	11
1.2.1 Délka slov.....	11
1.2.2 Délka vět .....	12
1.2.3 Jednoduché věty vs. souvětí.....	12
1.2.4 Slovní zásoba .....	12
1.2.5 Chyby .....	12
1.3 NEURON .....	12
1.3.1 Umělý neuron.....	13
1.4 NEURONOVÉ SÍTĚ [9].....	14
1.4.1 Rozdělení neuronových sítí.....	15
<b>II PRAKTICKÁ ČÁST .....</b>	<b>17</b>
<b>2 DATA.....</b>	<b>18</b>
2.1 SBĚR DAT .....	18
2.1.1 Počáteční fáze.....	18
2.1.2 Druhá fáze .....	18
2.1.3 Třetí fáze .....	18
2.1.4 Skript v jazyce Python .....	18
<b>3 IMPLEMENTACE .....</b>	<b>20</b>
3.1 PROGRAM V JAZYKU C#.....	20
3.1.1 Věty článku .....	20
3.1.2 Slova článku .....	20
3.1.3 Délka slov.....	21
3.1.4 Délka vět .....	21
3.1.5 Souvětí.....	21
3.1.6 Slovní zásoba .....	21
3.1.7 Chyby a překlepy .....	21
3.1.8 Výstup programu.....	22
3.1.9 Ukázka práce programu .....	22
3.1.9.1 Článek .....	22
3.1.9.2 Vyhodnocení článku .....	22
3.2 NEURONOVÁ SÍŤ V SYSTÉMU MATHEMATICA .....	23
<b>4 TESTOVÁNÍ .....</b>	<b>24</b>
4.1 TRÉNOVACÍ DATA .....	24
4.2 TESTOVACÍ DATA .....	24
4.3 VÝSLEDKY TESTŮ.....	25
4.4 DÍLČÍ VÝSLEDKY .....	29
4.5 VÝSLEDKY OPAKOVANÝCH TESTŮ .....	29
<b>ZÁVĚR .....</b>	<b>34</b>
<b>ZÁVĚR V ANGLIČTINĚ.....</b>	<b>35</b>

<b>SEZNAM POUŽITÉ LITERATURY.....</b>	<b>36</b>
<b>SEZNAM POUŽITÝCH SYMBOLŮ A ZKRATEK.....</b>	<b>38</b>
<b>SEZNAM OBRÁZKŮ .....</b>	<b>39</b>
<b>SEZNAM TABULEK.....</b>	<b>40</b>
<b>SEZNAM PŘÍLOH.....</b>	<b>41</b>



## ÚVOD

Tato práce se věnuje určování autorství. Jejím cílem není zahrnout velmi rozsáhlou oblast autorství textů, ale vytyčit základní známé metody a jejich kombinací a aplikováním otestovat úspěšnost určování autorství. Práce si neklade za cíl být univerzální pro všechny typy textů, ale zaměřuje se převážně na ty kratší až středně dlouhé (jako jsou příspěvky v diskuzních fórech, komentáře k článkům, samotné články, blogy, apod.).

Počátek určování autorství má základ již na konci 18. století a ani dnes se nedá říci, že by tento obor byl obsáhnut. Historickému pozadí a vývoji metod se práce věnuje ve své první kapitole. Zde jsou rovněž představeny vybrané metody, které byly implementovány a testovány.

První kapitola také představuje umělé neuronové sítě. Tyto neuronové sítě zaznamenávají rostoucí popularitu. Jejich aplikace nachází využití v oblastech, ve kterých je potřeba zpracovat velké množství dat. Právě určování autorství ve spojení s neuronovými sítěmi může vytvořit silný nástroj pro klasifikaci.

Pro vyzkoušení aplikovaných metod na reálných datech bylo potřeba sesbírat dostatečné množství učicích a testovacích dat. Tomuto procesu získávání a zpracování dat se věnuje druhá kapitola.

Třetí kapitola informuje o implementaci metod a problémech, které při ní vyvstávaly, a bylo je potřeba řešit. Rovněž nabízí pohled do implementace neuronové sítě.

Čtvrtá kapitola práce seznamuje s procesem a výsledky testování, hodnotí je a podává vysvětlení. Závěr práce navrhuje možnosti jejich zpřesnění a zvýšení úspěšnosti algoritmů.

## **I. TEORETICKÁ ČÁST**

# 1 URČOVÁNÍ AUTORSTVÍ

## 1.1 Vývoj

Jak již bylo řečeno v úvodu, počátek určování autorství sahá již do konce 18. století. Jednalo se o zpochybňování některých her W. Shakespeara, které vedlo k prvotním snahám vytvořit postupy pro ověření autorství. Jedním z průkopníků byl Edmond Malone [1], který v roce 1787 zpochybnil autorství třech částí Shakespearovy historické hry Jindřich IV. Tématem ověření autorství se později začala zabývat řada jazykovědců a následovalo zpochybnění autorství dalších děl[12].

Málo efektivní ruční analýza (počítání statistik) vedla k tomu, že pozornost patřila převážně významným literárním dílům. Na nich pak byly aplikovány a testovány nové metody, aby bylo možné porovnat výsledky.

Následující tabulka ve zkratce zachycuje vývoj studií pro určování autorství.

Jméno	Rok	Typ charakteristické vlastnosti
Mendenhall	1887	Délka vět, délka slov
Mascol	1888	Interpunkce
Yule	1944	Bohatost slovní zásoby
Kjell	1994	Frekvence znaků
Stamatatos a kol.	2000	Kusy syntaxe [2]
Koppel & Schler	2003	Výstřednost [3]
Pavelec a kol.	2007	Typy spojek

Tab. 1: Historie studií pro určování autorství [4]

## 1.2 Používané metody

### 1.2.1 Délka slov

V roce 1901 popsal T. C. Mendenhall v jedné ze svých prvních studií odlišnosti v rozložení četnosti slov různé délky Shakespearova díla od děl Baconových [5]. Metoda je jednoduchá na implementaci. Vyjadřuje počet znaků v jednom slově. Její výsledky ovšem

nejdou příliš přesvědčivé a spíše je vhodná pro rozlišení druhu textu<sup>1</sup>. Akceptovatelnější se stává pro dlouhé texty, pro krátké nemá vypovídající hodnotu. [6]

### 1.2.2 Délka vět

Rovněž popsána T. C. Mendenhallem. Vyjadřuje počet slov ve větě. Oproti délce slov se jedná o metodu náročnější na implementaci zvláště pro jazyky jako angličtina, kde není tolik striktní oddělování vět čárkami.

### 1.2.3 Jednoduché věty vs. souvětí

Jedná se o jednoduchou metodu indikující autorův styl psaní – s převládajícími jednoduchými větami či souvětími.

### 1.2.4 Slovní zásoba

Bohatost slovní zásoby je dalším rysem autorství. Každý autor má určitou slovní zásobu. Je ovšem třeba brát v potaz, že např. u vědeckých článků na stejné téma se bude, byť u různých autorů, slovní zásoba termínů podobat. Metoda je navíc silně závislá na délce textu [7].

### 1.2.5 Chyby

Předpokladem metody je stálý výskyt určitých druhů chyb v autorově písemném projevu. Autor se ovšem může v průběhu doby vyvíjet a vytvářet nové chyby či naopak přestat dělat chyby stávající. Kontrolují se chyby gramatické (např. pro anglický jazyk chybějící koncovka „s“ či „es“ u slovesa v přítomném čase 3. osoby jednotného čísla), překlepy (např. chybějící písmeno ve slově) a chyby formátování (např. všechna slova psaná velkými písmeny). [2]

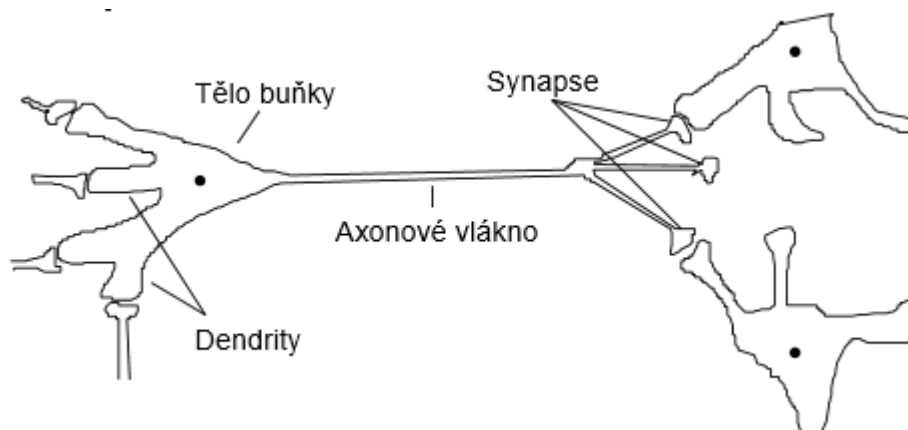
## 1.3 Neuron

Nervová buňka (tzv. neuron) je základním stavebním prvkem nervové soustavy. Mozkovou kůru člověka tvoří asi 15 miliard neuronů. Funkcí neuronů je přenos, uchování a zpracování informací nezbytných pro realizaci životních funkcí organismu.

---

<sup>1</sup> Např. báseň, román,...

Uměle vytvořený neuron je dán svým biologickým vzorem a tvoří základní jednotku neuronové sítě. Zjednodušený biologický neuron je znázorněn na následujícím obrázku.



Obr. 1: Biologický vzor [9]

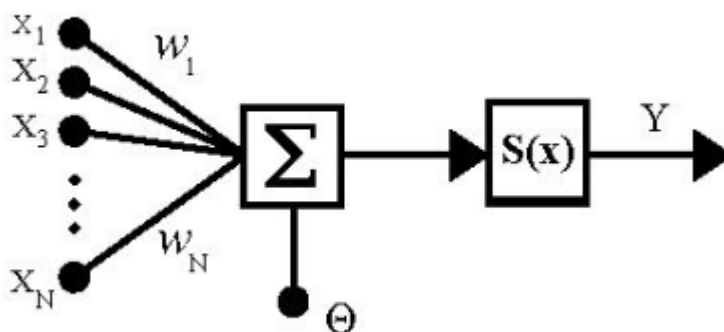
Místo vstupu signálů do těla neuronu (tzv. somatu) reprezentují dendrity. Axonové vlákno přenáší signál daný stupněm vybuzení k synapsím. Ty signál zesilují či zeslabují a předávají ho dalším neuronům. Synapse tvoří výstupní zařízení neuronů. [9]

Soma i axonové vlákno jsou obaleny membránou, která za jistých okolností generuje elektrické signály – tím je umožněno šíření informace. Synaptickými branami jsou impulsy přenášeny na dendrity jiných neuronů. Podrážděné neurony při dosažení tzv. prahu (určité hraniční meze) sami generují impuls a zajišťují tak šíření příslušné informace. Předpokladem paměťové schopnosti neuronů je změna synaptické propustnosti po každém průchodu signálu. Během života organismu prodělává propojení neuronů svůj vývoj – vytváří se nové, nebo přerušují (při zapomínání) staré synaptické spoje. [10]

### 1.3.1 Umělý neuron

Umělý neuron je základem matematického modelu neuronové sítě. Získáme jej přeformulováním zjednodušené funkce neuronu do matematické řeči. V podstatě se jedná o jednoduchou jednotku, která vynásobí všechny vstupy jejich váhami a takto získané hodnoty sečte. Výslednou hodnotu dosadí do přenosové funkce a výsledek této funkce je výstupem neuronu, který slouží jako vstup do dalších neuronů. Hodnoty vah se během učení mění [13].

Následující obrázek popisuje model umělého neuronu.



Obr. 2: Model umělého neuronu [11]

Popis modelu

- $x_i$  - vstupy neuronu
- $w_i$  – synaptické váhy
- $\Theta$  – práh
- $S(x)$  – přenosová (aktivační) funkce neuronu
- $Y$  – výstup neuronu

## 1.4 Neuronové sítě [9]

Standardní algoritmy nepracují dobře s porušenými nebo nekompletními daty. V reálném světě jsou ovšem právě tato jediným dostupným druhem dat. Řešení nabízí použití neuronových sítí – výpočetního modelu, který je schopný se sám učit.

Neuronové sítě jsou jedním z výpočetních modelů používaných v umělé inteligenci. Jsou inspirovány biologickými neuronovými sítěmi. Určitým způsobem je tato vlastnost předurčuje k tomu, aby byly, z hlediska základních principů, schopny se chovat stejně nebo alespoň podobně jako jejich biologické vzory. Vytvoření umělého lidského mozku se všemi jeho schopnostmi je jen velmi těžce řešitelná věc, nicméně skýtá se šance alespoň některé funkce lidského myšlení simulovat a tyto pak implementovat.

Ukládání, zpracování a předávání informace probíhá prostřednictvím celé neuronové sítě spíše než pomocí určitých paměťových míst. Paměť a zpracování informace je tedy ve své podstatě spíše globální než lokální.

Znalosti jsou ukládány prostřednictvím síly vazeb mezi jednotlivými neurony. Vazby vedoucí ke správné odpovědi jsou posilovány a naopak.

Základní a podstatná vlastnost neuronových sítí je učení. Tímto se neuronové sítě základně odlišují od dosud běžného použití počítačů, kdy transformaci vstupní množiny dat na množinu výstupních dat určuje právě fáze učení.

#### 1.4.1 Rozdělení neuronových sítí

Všechny typy neuronových sítí se skládají ze stejných stavebních jednotek – neuronů. O jaký typ sítě se jedná, určují různé přenosové funkce, spojení mezi neurony a učící algoritmus.

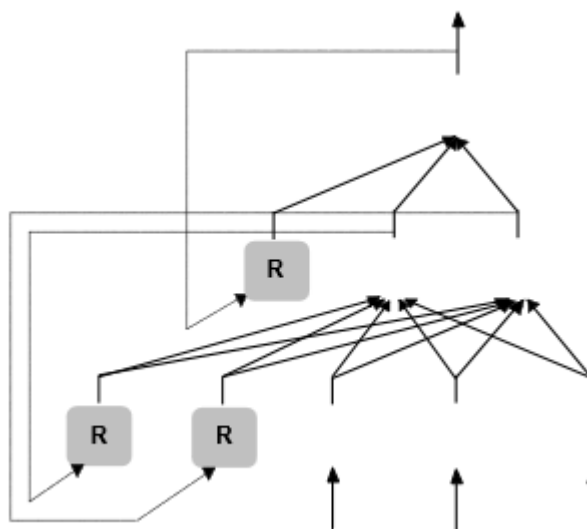
Umělé neuronové sítě se dělí podle několika hlavních kritérií.

##### Podle počtu vrstev

- Jednovrstvé neuronové sítě
- Vícevrstvé neuronové sítě

##### Podle metody šíření informace

- Dopředné neuronové sítě – existují v nich pouze dopředné spojení mezi neurony. Každý neuron jedné vrstvy vysílá signály na každý neuron vrstvy následující. Neexistují spojení do předcházející vrstvy, ani spojení v rámci jedné vrstvy.
- Rekurentní neuronové sítě – mají složitější rozdělení vrstev. Některé neurony jsou zároveň vstupní i výstupní. Signál v nich může putovat oběma směry.



Obr. 3: Rekurentní vícevrstvá neuronová síť [9]

**Podle přítomnosti „učitele“**

- s učitelem – síti je poskytována informace, jak má na dané vstupy reagovat. Podle toho se síť snaží přizpůsobit svoje váhy tak, aby se aktuální výstupní informace co nejvíce podobala požadovanému originálu. Úlohou učícího algoritmu je minimalizovat učící chybu – při dosažení jejího určitého minima učení končí.
- bez učitele – síti není definováno, jak má reagovat na daný vstup, tzn. není poskytnuta informace o požadovaném výstupu.



## **II. PRAKTICKÁ ČÁST**

## 2 DATA

### 2.1 Sběr dat

#### 2.1.1 Počáteční fáze

Počáteční získávání dat probíhalo manuálně. Z internetových blogů<sup>2</sup> psaných v anglickém jazyce bylo staženo nejprve osm článků od každého autora. Pět článků bylo určeno pro trénovací množinu a zbylé tři články pro testování.

#### 2.1.2 Druhá fáze

Pro velkou chybu při rozpoznávání dvou autorů ve vzorcích dat z prvotní fáze bylo učiněno rozhodnutí, že vzorek dat je příliš úzký a dojde k jeho rozšíření na osm článků pro nastavení neuronové sítě (trénování) a pět článků pro samotné testování schopnosti neuronové sítě na základě vstupních parametrů správně určit jednoho ze dvou autorů.

#### 2.1.3 Třetí fáze

Kvůli časově náročnému získávání potřebných dat a pro jednoduchost dalšího získávání byl napsán skript v jazyce Python, který stahuje články z daného blogu a ukládá jejich obsah do souboru na lokálním disku. Tímto byl vyřešen problém s časově náročným manuálním získáváním článků a celkový sběr dat se zúžil pouze na problém vyhledání a předání URL<sup>3</sup> blogu. Skript již sám získá odkazy na konkrétní články i texty těchto článků.

#### 2.1.4 Skript v jazyce Python

Jak již bylo uvedeno výše, pro získání článků autorů byl napsán skript v jazyce Python. Pro získání celého obsahu webové stránky blogu byl použit modul urllib<sup>4</sup>. Pro úpravu zdrojového kódu stránky a získání seznamu odkazů na jednotlivé články byla použita knihovna BeautifulSoup<sup>5</sup>. Zdrojový kód stránek jednotlivých článků byl stažen a získán

---

<sup>2</sup> **Blog** neboli také weblog je druh internetové prezentace, jejíž obsah se skládá především z článků rozdělených do kategorií. – převzato z <http://it-slovník.cz/pojem/blog>

<sup>3</sup> Unique Resource Locator = jednoznačné určení zdroje. Je to způsob, jak jednoznačně zapsat umístění souboru na internetu nebo na intranetu. – převzato z <http://www.jakpsatweb.cz/html/url.html>

<sup>4</sup> Dokumentace dostupná na <http://docs.python.org/2/library/urllib.html>

<sup>5</sup> Dokumentace dostupná na <http://www.crummy.com/software/BeautifulSoup/bs4/doc/>

z nich autorův text. Texty autora byly uloženy do textového dokumentu na lokální disk. Jednotlivé články byly od sebe odděleny pro jejich následné snadné zpracování.

### 3 IMPLEMENTACE

Zjištěné metody pro určování autorství byly implementovány v jazyku C#<sup>6</sup> v nástroji Visual Studio Express 2012 od společnosti Microsoft. Vstupem programu byl textový dokument obsahující sadu článků jednoho autora. Výstupem programu byl rovněž textový dokument obsahující však číselné hodnoty pro jednotlivé atributy autorství z daných článků. Tento výstupní textový dokument se použil jako vstup pro neuronovou síť implementovanou v systému Mathematica<sup>7</sup> od společnosti Wolfram. Výstupem této sítě byla schopnost rozpoznat správného autora článku.

#### 3.1 Program v jazyku C#

Objektem reprezentujícím článek autora je v programu třída Article, jejíž metody vrací číselné vyjádření konkrétního atributu článku. Konkrétní implementace atributů je popsána v kapitolách 5.1.3 – 5.1.7.

##### 3.1.1 Věty článku

Jako první byl ze článku získán seznam všech vět, případně souvětí. Větou je chápána ta část textu, která je ukončena interpunkčním znaménkem (konkrétně tečkou, vykřičníkem a otazníkem), za kterým následuje velké písmeno. Tato metoda získání vět textu není optimální, protože do vět zahrne i řetězce jako jsou zkratky (např. V.I.P. – very important person budou tři samostatné věty), případně zkrácené názvy (např. univerzita T. Bati budou dvě věty). Souvětí, tedy věty oddělené čárkou, středníkem či spojkou, byly brány jako jedna věta. Samotné interpunkční znaménka byla z vět vyjmuta.

##### 3.1.2 Slova článku

K získání seznamu všech slov v článku se dospělo rozdělením vět podle mezer. Mezi slova tak byly zahrnuty i anglické určité a neurčité členy<sup>8</sup>, které samostatně většinou netvoří žádné slovo a mohly proto zkreslovat statistiku, avšak jejich výskyt ve větách byl ponechán jako jeden z prvků určení autora.

---

<sup>6</sup> Jedná se o objektově orientovaný programovací jazyk pro .NET Framework. Tutoriál dostupný na <http://msdn.microsoft.com/en-us/library/aa287558%28v=vs.71%29.aspx>.

<sup>7</sup> Dokumentace dostupná na <http://www.wolfram.com/mathematica/>.

<sup>8</sup> Jako „a“, „an“, „the“.

### 3.1.3 Délka slov

Metoda `AverageWorldLength` ve třídě `Article` přijímá jako parametr seznam všech slov článku. Pro každé slovo je spočítána jeho délka (počet znaků). Metoda vrací aritmetický průměr všech těchto počtu znaků.

### 3.1.4 Délka vět

Metoda `AverageSentenceLength` přijímá jako parametr seznam vět. Pro každou větu je spočítán počet slov v této větě. Slova jsou získána rozdělením věty podle mezer. Metoda vrací aritmetický průměr počtu slov.

### 3.1.5 Souvětí

Metoda `MultipleSencences` přijímá jako parametr seznam vět. Pro každou větu zjišťuje, jestli se nejedná o souvětí. Pokud ano, označí větu jako souvětí. Metoda vrací poměr počtu souvětí k počtu jednoduchých vět.

### 3.1.6 Slovní zásoba

Jedná se o bohatost autorova slovníku v článku. Metoda `Vocabulary` přijímá seznam všech slov a kontroluje unikátnost těchto slov, takže duplicitní slova se započítají jako jedno slovo. Metoda vrací počet unikátních slov.

### 3.1.7 Chyby a překlepy

Metoda `GetErrors` vrací počet gramatických chyb a počet překlepů v článku. K tomu využívá webovou službu `After the Deadline`<sup>9</sup>. Program používá třídu `AfterTheDeathline` implementovanou Arikem Poznanskim<sup>10</sup>. Článek je přes webového klienta odeslán na server a z odpovědi ve formátu XML<sup>11</sup> je získán počet gramatických chyb a překlepů. Metoda vrací tyto dva údaje.

---

<sup>9</sup> Popis služby dostupný na <http://afterthedeathline.com/>.

<sup>10</sup> Implementace třídy dostupná na <http://www.codeproject.com/Articles/209514/Csharp-Library-for-Grammar-and-Spell-Checking>

<sup>11</sup> Extensible Markup Language. Dokumentace dostupná na <http://afterthedeathline.com/>.

### 3.1.8 Výstup programu

Program vytváří nový textový dokument s výsledky metod. Pro každý článek je vypsán jeden řádek. Řádek obsahuje výstupy metod oddělené tabulátory.

### 3.1.9 Ukázka práce programu

Jak již bylo uvedeno, program načte textový soubor s články jednoho autora oddělenými od sebe řetězcem znaků „#####“<sup>12</sup>. Řetězec byl vybrán takový, aby byl zajištěn jeho co nejméně pravděpodobný výskyt v samotném textu článku. Po načtení článku je vytvořen seznam všech vět a seznam všech slov. Následně je u textu článku zkontrolována gramatika a překlepy webovou službou After the Deadline. Používání webové služby pro kontrolu gramatiky a překlepů s sebou nese fakt, že články nemohou být kontrolovány bezprostředně po sobě, aby server nepřestal reagovat, proto je mezi odesíláním textu jednoho a druhého článku ke kontrole program na 0,7 sekundy „uspán“<sup>13</sup>. Tento údaj byl zjištěn testováním jako nejkratší možná doba mezi kontrolami článků, kdy server ještě odesílá odpovědi.

#### 3.1.9.1 Článek

Ukázkový článek<sup>14</sup> byl vybrán od autora Daniela Coopera<sup>15</sup>.

#### 3.1.9.2 Vyhodnocení článku

Pro článek program vyhodnotil následující data:

- Průměrná délka vět: 27,75
- Poměr souvětí k větám jednoduchým: 1
- Průměrná délka slova: 4,83783783783784
- Slovní zásoba vzhledem k počtu vět: 21
- Počet překlepů vzhledem k počtu vět: 1
- Počet gramatických chyb vzhledem k počtu vět: 0,25

---

<sup>12</sup> Viz. Příloha 2 : Ukázka souboru s články autora

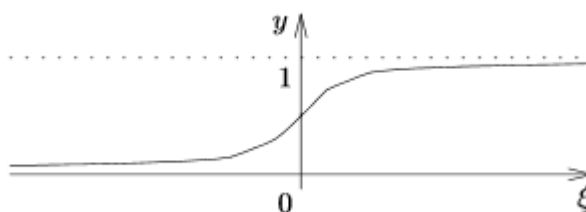
<sup>13</sup> Jedná se o pozdržení vykonávání dalších instrukcí.

<sup>14</sup> Viz. Příloha 1: Ukázkový článek

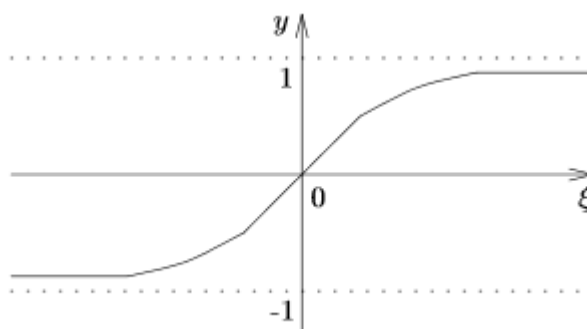
<sup>15</sup> Článek dostupný na <http://www.engadget.com/2013/05/08/microsoft-whitespace-tanzania/>

### 3.2 Neuronová síť v systému MATHEMATICA

Jako síť byla implementována dopředná neuronová síť se šesti neurony ve skryté vrstvě a jedním výstupním neuronem. Počet skrytých neuronů byl v průběhu testování navyšován z šesti na 10 a následně 15. Testy probíhaly pro aktivační funkce tvaru sigmoidy i hyperbolického tangentu.



Obr. 4: Aktivační funkce tvaru sigmoidy



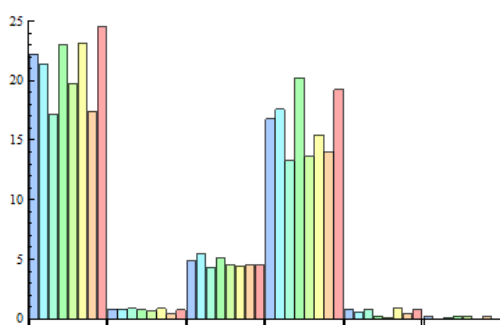
Obr. 5: Aktivační funkce hyperbolický tangens

## 4 TESTOVÁNÍ

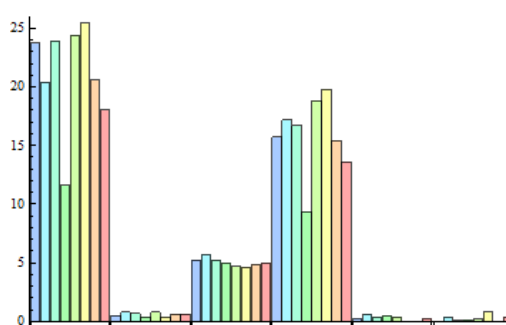
Testování probíhalo na datech z druhé fáze sběru<sup>16</sup>.

### 4.1 Trénovací data

Pro naučení sítě posloužilo 8 článků každého autora.



Obr. 6: Graf trénovací množiny  
autora1



Obr. 7: Graf trénovací množiny  
autora2

Z obrázků Obr. 6 a 7 je patrné, že testovací data prvního i druhého autora se značně podobají.

Prvních osm sloupců vyjadřuje průměrnou délku vět, dalších osm poměr jednoduchých vět k souvětím, dalších osm průměrnou délku slova, dalších osm bohatost slovní zásoby vzhledem k počtu vět, dalších osm počet překlepů vzhledem k počtu vět a posledních osm sloupců počet gramatických chyb vzhledem k počtu vět.

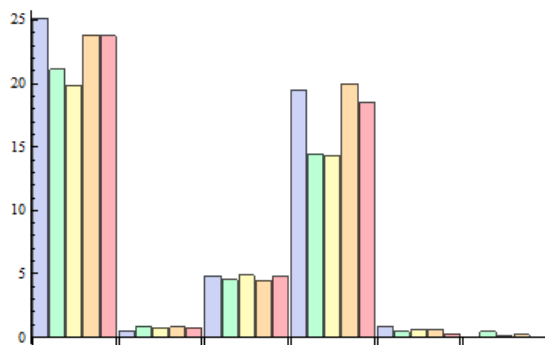
### 4.2 Testovací data

Pro testovací účely bylo určeno pět článků od každého autora.

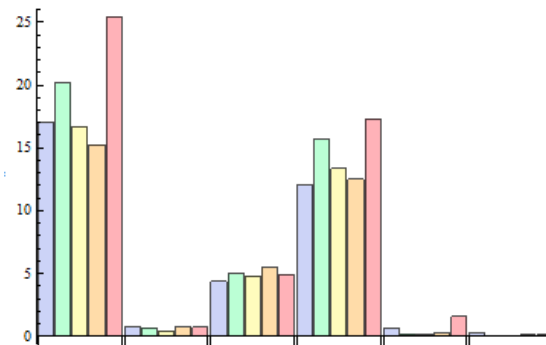
---

<sup>16</sup> Viz. kapitola 2.1.2





Obr. 8: Graf testovací množiny  
autora1



Obr. 9: Graf testovací množiny  
autora2

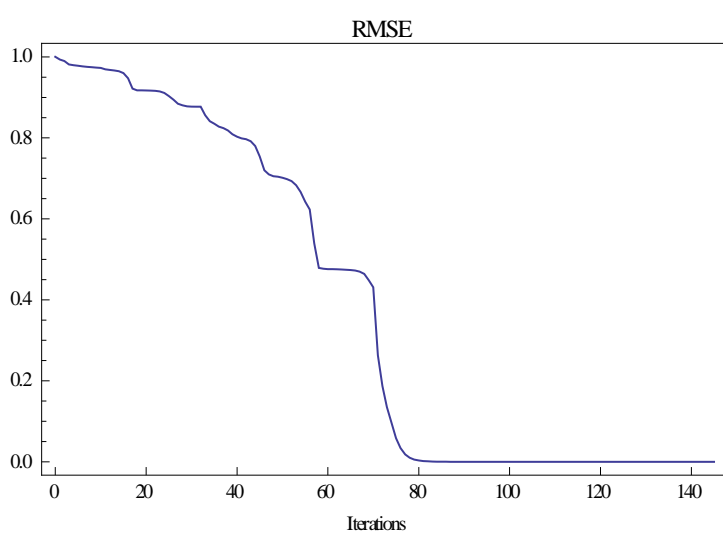
### 4.3 Výsledky testů

Pro přehlednost se testy uvádějí v tabulkách. Graf RMSE u iterací zobrazuje snižování chyby v průběhu iterací.

Název:	Test 1: 6 neuronů, sigmoida
Popis:	Neuronová síť o šesti skrytých neuronech a aktivační funkci tvaru sigmoidy.
Iterací k naučení: 528	<p>RMSE</p> <p>Iterations</p>

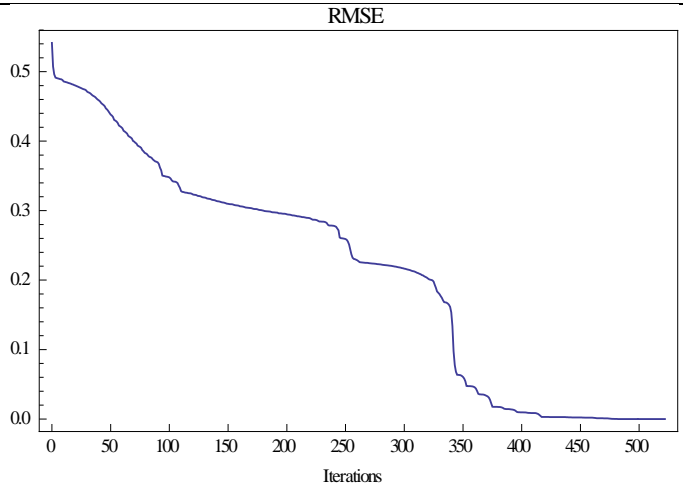
Chyba:	40%
Chybně určené články:	3, 5, 6, 10

Tab. 2: Test 1: 6 neuronů, sigmoida

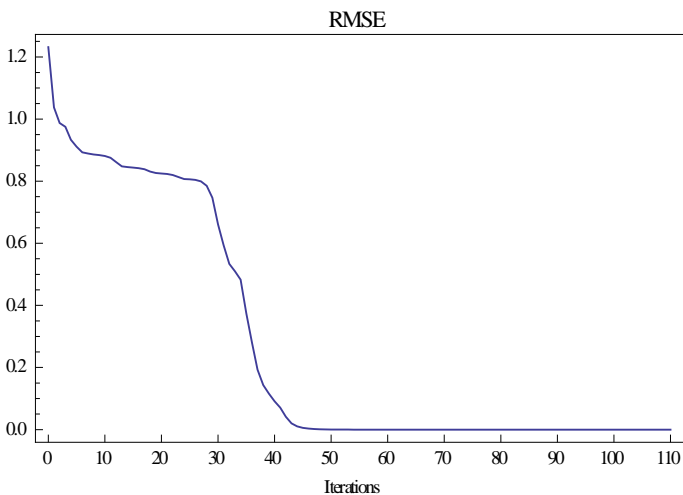
Název:	Test 2: 6 neuronů, hyperbolický tangens
Popis:	Neuronová síť o šesti skrytých neuronech a aktivační funkci tvaru hyperbolického tangentu
Iterací k naučení: 145	
Chyba:	40%
Chybně určené články:	5, 6, 7, 10

Tab. 3: Test 2: 6 neuronů, hyperbolický tangens

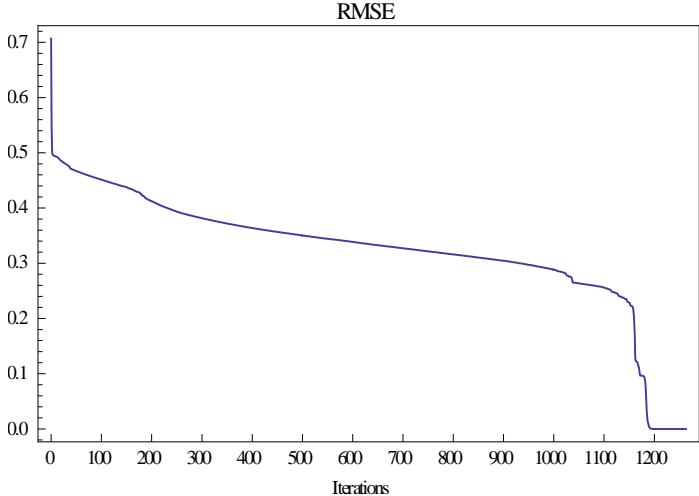
Název:	Test 3: 10 neuronů, sigmoida
Popis:	Neuronová síť o deseti skrytých neuronech a aktivační funkci tvaru sigmoidy.

Iterací k naučení: 522	
Chyba:	30%
Chybně určené články:	5, 6, 10

Tab. 4: Test 3: 10 neuronů, sigmoida

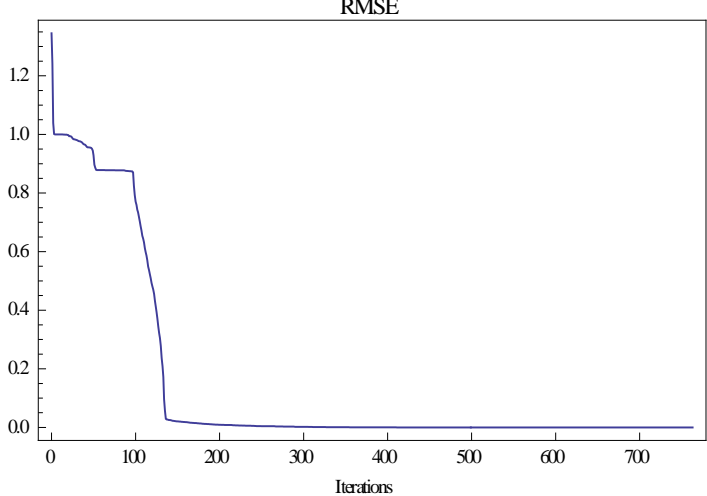
Název:	Test 4: 10 neuronů, hyperbolický tangens
Popis:	Neuronová síť o deseti skrytých neuronech a aktivační funkci tvaru hyperbolického tangentu
Iterací k naučení: 110	
Chyba:	40%
Chybně určené články:	5, 6, 7, 10

Tab. 5: Test 4: 10 neuronů, hyperbolický tangens

Název:	Test 5: 15 neuronů, sigmoida
Popis:	Neuronová síť o patnácti skrytých neuronech a aktivační funkci tvaru sigmoidy.
Iterací k naučení: 1263	
Chyba:	20%
Chybně určené články:	5, 10

Tab. 6: Test 5: 15 neuronů, sigmoida

Název:	Test 6: 15 neuronů, hyperbolický tangens
Popis:	Neuronová síť o patnácti skrytých neuronech a aktivační funkci tvaru hyperbolického tangentu

Iterací k naučení: 763	
Chyba:	40%
Chybně určené články:	5, 6, 7, 10

Tab. 7: Test 6: 15 neuronů, hyperbolický tangens

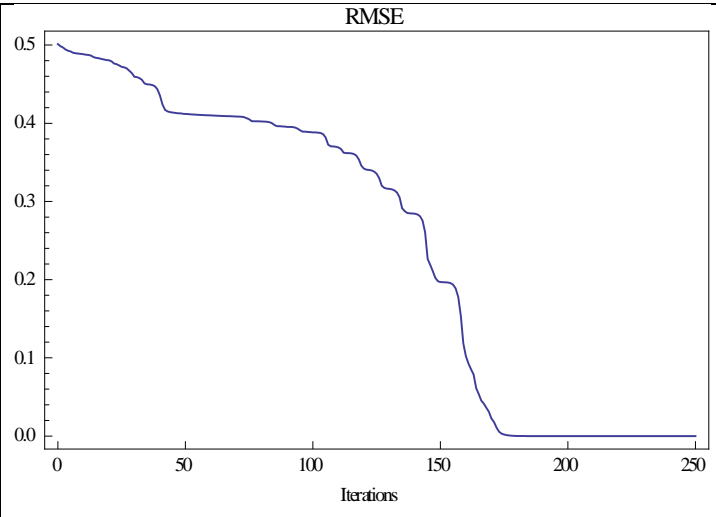
#### 4.4 Dílčí výsledky

Zatímco neuronová síť s aktivační funkcí tvaru sigmoidy se zvyšujícím se počtem neuronů snižuje chybu v určování autora, pro hyperbolický tangens nemá nárůst neuronů ve skryté vrstvě o nic pozitivnější vliv na přiřazení textu ke správnému autorovi.

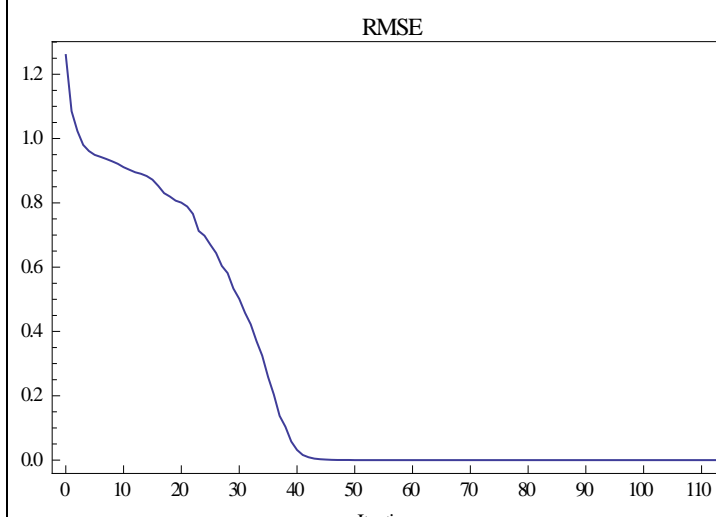
Během testování nemohl zůstat bez povšimnutí fakt, že nesprávně byly k autorovi přiřazovány pouze články s čísly 5, 6, 7 a 10. Proto budou tyto články přesunuty do trénovací množiny a nahrazeny jinými články z této množiny. Nové testování bude probíhat stejně jako původní.

#### 4.5 Výsledky opakovaných testů

Název:	Test 7: 6 neuronů, sigmoidida
Popis:	Neuronová síť o šesti skrytých neuronech a aktivační funkci tvaru sigmoidy.

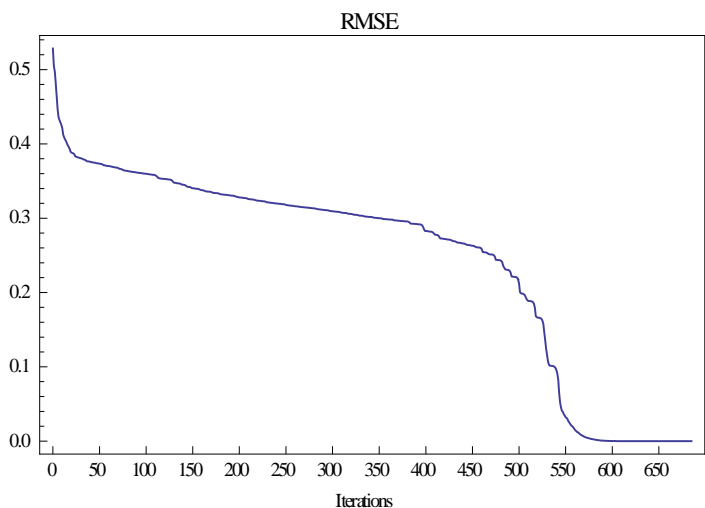
Iterací k naučení: 250	
Chyba:	20%
Chybně určené články:	3, 5

Tab. 8: Test 7: 6 neuronů, sigmoida

Název:	Test 8: 6 neuronů, hyperbolický tangens
Popis:	Neuronová síť o šesti skrytých neuronech a aktivační funkci tvaru hyperbolického tangentu
Iterací k naučení: 114	

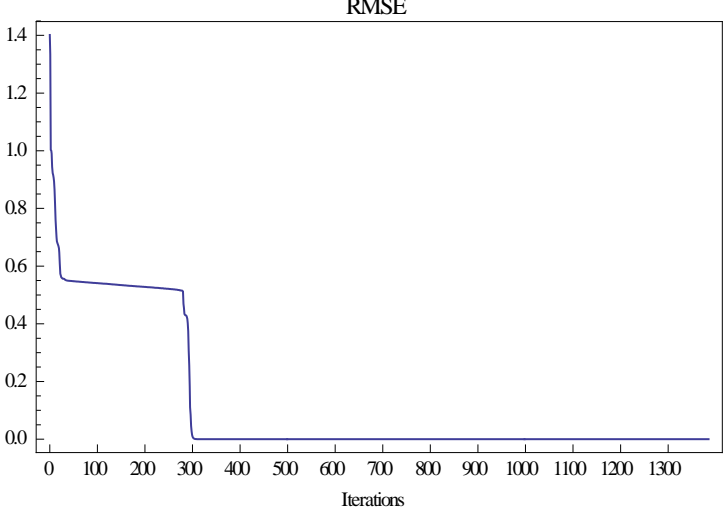
Chyba:	30%
Chybně určené články:	3, 5, 6

Tab. 9: Test 8: 6 neuronů, hyperbolický tangens

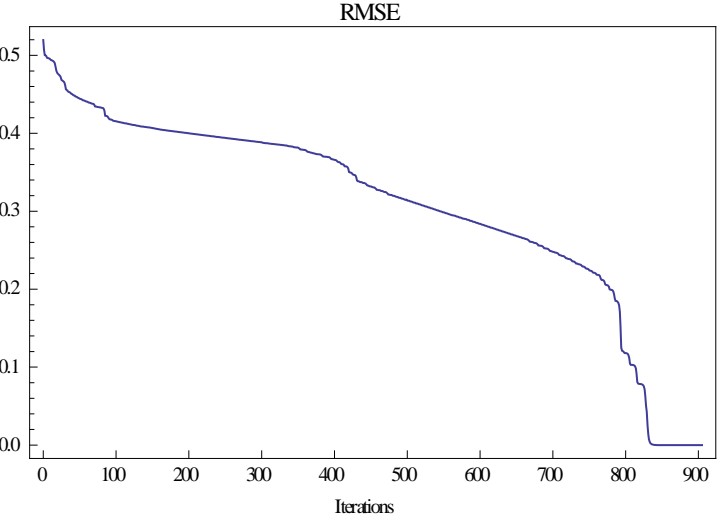
Název:	Test 9: 10 neuronů, sigmoida
Popis:	Neuronová síť o deseti skrytých neuronech a aktivační funkci tvaru sigmoidy.
Iterací k naučení: 685	
Chyba:	30%
Chybně určené články:	1, 2, 3

Tab. 10: Test 9: 10 neuronů, sigmoida

Název:	Test 10: 10 neuronů, hyperbolický tangens
Popis:	Neuronová síť o deseti skrytých neuronech a aktivační funkci tvaru hyperbolického tangentu

Iterací k naučení: 1385	 <p>The plot shows the Root Mean Square Error (RMSE) on the y-axis (ranging from 0.0 to 1.4) against the number of iterations on the x-axis (ranging from 0 to 1300). The error starts at approximately 1.4, drops sharply to about 0.55 by iteration 50, remains relatively stable until iteration 250, and then drops again to near zero by iteration 300, where it remains for the rest of the training process.</p>
Chyba:	30%
Chybně určené články:	3, 5, 6

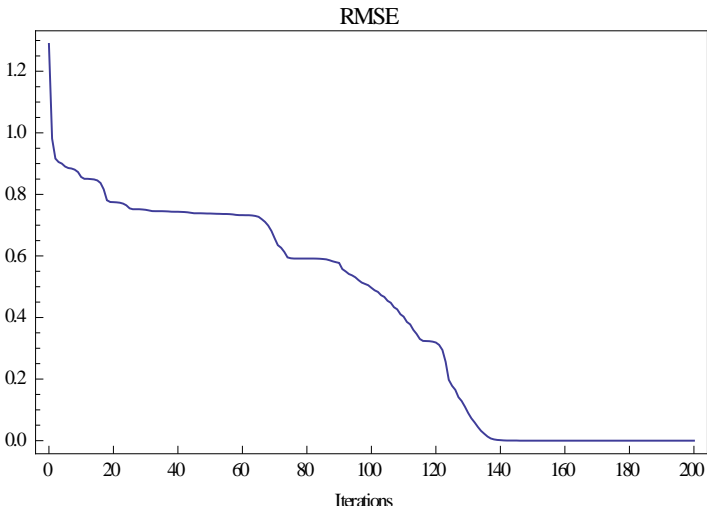
Tab. 11: Test 10: 10 neuronů, hyperbolický tangens

Název:	Test 11: 15 neuronů, sigmoida
Popis:	Neuronová síť o patnácti skrytých neuronech a aktivační funkci tvaru sigmoidy.
Iterací k naučení: 905	 <p>The plot shows the Root Mean Square Error (RMSE) on the y-axis (ranging from 0.0 to 0.5) against the number of iterations on the x-axis (ranging from 0 to 900). The error starts at approximately 0.5, decreases gradually to about 0.35 by iteration 400, and then continues to decrease more slowly until iteration 750. Finally, it drops sharply to near zero by iteration 850, where it remains for the rest of the training process.</p>



Chyba:	20%
Chybně určené články:	3, 5

Tab. 12: Test 11: 15 neuronů, sigmoida

Název:	Test 12: 15 neuronů, hyperbolický tangens
Popis:	Neuronová síť o patnácti skrytých neuronech a aktivační funkci tvaru hyperbolického tangentu
Iterací k naučení: 200	
Chyba:	20%
Chybně určené články:	3, 5

Tab. 13: Test 12: 15 neuronů, hyperbolický tangens

Článek číslo 5 se opět objevoval téměř při každé chybě. Ostatní články (7 a 10) byly při druhém testování vždy správně určeny a článek číslo 6 se vyskytl jako chybný dvakrát. Proto není možné z čísel článků vyvozovat žádné závěry.

Obecně lze z testování vyvodit, že neuronová síť s aktivační funkcí hyperbolického tangentu je schopna učení v menším počtu iterací, nežli síť se sigmoidou.

## ZÁVĚR

Cílem této diplomové práce bylo seznámit se s metodami určování autorství, tyto implementovat a otestovat jejich úspěšnost. Byly popsány existující postupy, jejich vývoj a hlavní výhody i nevýhody v určování autorství. Ve své teoretické části podala práce také přehled o neuronových sítích.

V praktické části práce byly algoritmy metod i samotná neuronová síť úspěšně naimplementovány. Pro jejich základní otestování bylo sesbíráno 26 článků od dvou autorů. U každého autora bylo osm článků použito pro učení sítě (jako trénovací množina) a pět článků pro test (testovací množina). Všechny algoritmy byly zkombinovány, aby rys autora byl co nejzřetelnější. Učení neuronové sítě proběhlo vždy v pořádku a v případě aktivační funkce tvaru sigmoidy se síť byla schopna naučit trénovací sadu dat v průměru za cca 690 iterací, v případě hyperbolického tangentu postačilo průměrně 452 iterací. Výsledky testů však zobrazovaly chybu určení autora o hodnotě 40-50%. Proto došlo k modifikaci algoritmů a pro každý byl zaveden parametr délky textu. Při opakování testů na stejné sadě dat se zavedení tohoto parametru ukázalo jako přínosné, avšak chybovost určení autora se pohybovala stále vysoko a její hodnota byla převážně 30-40%.

Nabízí se několik možných vysvětlení pro tak vysokou chybu. Jedním z nich může být malá sada dat. Pokud by se od každého autora podařilo sesbírat 40-50 článků, došlo by pravděpodobně ke snížení chyby. Druhým důvodem může být typ dat – jednalo se převážně o technické články, které autorovi nenabízí tolik otevřeného prostoru, aby jasně vynikly rysy autorova stylu. V neposlední řadě může být příčinou vysoké chyby určení autora relativně malý počet zkoumaných rysů. Pokud by došlo k implementaci dalších metod pro určení autorství, mělo by tím dojít i ke snížení chyby.

Práce je závislá na anglických textech převážně jen kvůli použitému anglickému slovníku pro kontrolu gramatiky a překlepů. Pokud by tento slovník byl nahrazen např. českým, bylo by možné testovat autorství česky psaných textů.

## ZÁVĚR V ANGLIČTINĚ

The aim of the thesis is to describe and implement methods solving the problem of determining authorship. Together with the survey of neural networks, the history and basic characteristics of existing techniques were presented in the theoretical part.

The practical part offers the implemented algorithms and neural networks. 26 articles written by two authors were collected for the experiments. Eight articles from each of the authors were used as a neural network training set, the rest represented the testing set. The algorithms were combined to make the simplest author recognition. Neural network learning was always right and in case of the sigmoid, the network was able to learn the training data set for 690 iterations on an average in the case of hyperbolic tangent sufficed approximately 452 iterations. The results of the tests showed the average error between 40-50%. Better results were observed in the second round of the tests. The algorithms were modified and a new text length parameter was added, however, the error of determining authorship was still high, 30-40%.

The probable cause of the high error rate might be the small set of texts collected for the experiments. A larger number of articles (40-50 from the each author) would probably reduce the error rate. Last but not least, may cause high error rate relatively small number of studied method of authorship attribution. The implementation of other methods for determining authorship may reduce the error, as well.

The algorithms presented in the thesis are language independent, the replacement of the English dictionary, e. g. with the Czech dictionary, would enable determining authorship even for texts in Czech.

**SEZNAM POUŽITÉ LITERATURY**

- [1] MALONE, Edmond. A Dissertation on the Three Parts of King Henry VI. Tending to Shew That Those Plays Were Not Written Originally by Shakspeare. Gale Ecco, Print Editions, 2010, ISBN 978-1140682981
- [2] STAMATATOS, Efstathios. A Survey of Modern Authorship Attribution Methods. Journal of the American Society for Information Science and Technology [online]. 2009, roč. 60(č. 3), 538-556. Přepis dostupný z: <http://www.icsd.aegean.gr/lecturers/stamatatos/papers/survey.pdf>
- [3] KOPPEL Moshe, SCHLER, Jonathan. Exploiting Stylistic Idiosyncrasies for Authorship Attribution. Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis. 2003, 69-72. Přepis dostupný z: <http://u.cs.biu.ac.il/~koppel/papers/ijcai-idiosyncrasy-final.pdf>
- [4] KOPPEL, Moshe, SCHLER, Jonathan, ARGAMON, Shlomo. Computational Methods in Authorship Attribution. Journal of the American Society for Information Science and Technology [online]. 2009, roč. 60(č. 1), 9-36. Přepis dostupný z: <http://u.cs.biu.ac.il/~koppel/papers/authorship-JASIST-final.pdf>
- [5] WILLIAMS, C. B.. Mendenhall's studies of word-length distribution in the works of Shakespeare and Bacon. Biometrika. 1975. 207-212
- [6] COYOTL-MORALES, Rosa María, VILLASEÑOR-PINEDA, Luis, MONTES-Y-GÓMEZ, Manuel, ROSSO, Paolo. Authorship Attribution using Word Sequences. [online], 10s, [cit. 2013-04-05]. Dostupné z WWW: [http://users.dsic.upv.es/~proso/resources/CoyotlEtAl\\_CIARP06.pdf](http://users.dsic.upv.es/~proso/resources/CoyotlEtAl_CIARP06.pdf)
- [7] STAMATATOS, E., FAKOTAKIS, N., KOKKINAKIS, G.. Computer-Based Authorship Attribution Without Lexical Measures. Computers and the Humanities [online], 2001, 193-214. Dostupné z WWW: <http://www.iula.upf.edu/materials/050401sanchez.pdf>
- [8] GRIEVE, Jack. Quantitative Authorship Attribution: An Evaluation of Techniques [online]. 2007, č. 3, 251-270. Dostupné z WWW: <https://lirias.kuleuven.be/bitstream/123456789/331335/1/Grieve++authorship+attribution.pdf>
- [9] VONDRÁK, Ivo, Neuronové sítě. [online]. 1994, 56s. [cit. 2013-04-29]. Dostupné z WWW: [http://vondrak.cs.vsb.cz/download/Neuronove\\_site.pdf](http://vondrak.cs.vsb.cz/download/Neuronove_site.pdf)

- [10] ŠÍMA, Jiří, NERUDA, Roman. Teoretické otázky neuronových sítí. 1.vyd.Praha: MATFYZPRESS,1996. 390s. Dostupné z WWW: <http://www2.cs.cas.cz/~sima/kniha.pdf>
- [11] DUŠEK, Jakub. Neuronové sítě. [online], 5s. [cit. 2013-05-09]. Dostupné z WWW: [http://lzd.spsejecna.net/web/beranek/I3B/Du%C5%A1ekJakub\\_Neuronov%C3%A9s%C3%ADt%C4%9B.pdf](http://lzd.spsejecna.net/web/beranek/I3B/Du%C5%A1ekJakub_Neuronov%C3%A9s%C3%ADt%C4%9B.pdf)
- [12] VAŠÁK, Pavel. Metody určování autorství. 1. Vyd. Praha: Academia, 1980.
- [13] ZELINKA, I., Umělá inteligence I, VUT Brno, 1998, ISBN 80-214-1163-5

**SEZNAM POUŽITÝCH SYMBOLŮ A ZKRATEK**

URL	Unique Resource Locator
XML	Extensible Markup Language
RMSE	Root Mean Squared Error

**SEZNAM OBRÁZKŮ**

Obr. 1: Biologický vzor [9].....	13
Obr. 2: Model umělého neuronu [11] .....	14
Obr. 3: Rekurentní vícevrstvá neuronová síť [9] .....	15
Obr. 4: Aktivační funkce tvaru sigmoidy .....	23
Obr. 5: Aktivační funkce hyperbolický tangens .....	23
Obr. 6: Graf trénovací množiny autora1 .....	24
Obr. 7: Graf trénovací množiny autora2 .....	24
Obr. 8: Graf testovací množiny autora1 .....	25
Obr. 9: Graf testovací množiny autora2.....	25

**SEZNAM TABULEK**

Tab. 1: Historie studií pro určování autorství [4] .....	11
Tab. 2: Test 1: 6 neuronů, sigmoida .....	26
Tab. 3: Test 2: 6 neuronů, hyperbolický tangens.....	26
Tab. 4: Test 3: 10 neuronů, sigmoida .....	27
Tab. 5: Test 4: 10 neuronů, hyperbolický tangens.....	27
Tab. 6: Test 5: 15 neuronů, sigmoida .....	28
Tab. 7: Test 6: 15 neuronů, hyperbolický tangens.....	29
Tab. 8: Test 7: 6 neuronů, sigmoida .....	30
Tab. 9: Test 8: 6 neuronů, hyperbolický tangens.....	31
Tab. 10: Test 9: 10 neuronů, sigmoida .....	31
Tab. 11: Test 10: 10 neuronů, hyperbolický tangens.....	32
Tab. 12: Test 11: 15 neuronů, sigmoida .....	33
Tab. 13: Test 12: 15 neuronů, hyperbolický tangens.....	33



## SEZNAM PŘÍLOH

Příloha 1: Ukázkový článek

Příloha 2 : Ukázka souboru s autorovými články

Příloha 3: Disk CD s diplomovou prací

## **PŘÍLOHA 1: UKÁZKOVÝ ČLÁNEK**

It's not just Bill Gates that has a benevolent eye turned towards Africa, as Microsoft has launched the second stage of its 4Afrika initiative in Tanzania. The company has teamed up with local provider UhuruOne to roll out white space broadband to the University of Dar es salaam, and Redmond is working with banks to help students get loans to buy Windows 8 hardware. Microsoft will also employ some students as on-campus network support, offering training and qualifications to help them in the future. While there's no mention of the custom Huawei W1 the company is offering in Kenya, we assume it'll also be available as part of this project, too.

## **PŘÍLOHA 2 : UKÁZKA SOUBORU S ČLÁNKY AUTORA**

CEO departures normally come in one of two forms -- a dignified slope to the exit, or an explosive, controversial parting of ways. Departing Scarlet Motors CEO Julien Fourgeaud has taken a different approach with a stream-of-consciousness blog post, describing the company's origins and cryptically hinting at divisions within the EV maker. Naming no names, the former Rovio executive mentions that he wishes people "kept their commitments, their word" and were "working together towards building an amazing product." But despite the dissatisfaction, Fourgeaud says that he'll continue to support the business as a shareholder -- in between making sure his LinkedIn profile is up to date.

#####

Spring is here, which means it's high time that Sony refreshed its laptop line for the cool kids in Japan. The company is talking up a quartet of new VAIOs including the previously-reviewed Duo 11 as well as the unfamiliar trio of the VAIO Fit 15E, Fit 14 and Fit 15. Delving deep into that last model, the 22.5mm-thick unit comes with a 1,920 x 1,080 touchscreen display and runs the 64-bit version of Windows 8 on a 2GHz Intel Core i7-3537U CPU. Look deeper and you'll find 8GB RAM, Intel HD Graphics 4000, 1TB of hybrid storage as well as 802.11 b/g/n WiFi, NFC and an Exmor R webcam. The various models will filter into Japanese stores between May 18th and June 1st, with the stock Fit 15 setting you back 180,000 yen (\$1,818).

#####

Ever since Acer's Linxian Lang said that Microsoft would eat "hard rice" for building its own Windows RT hardware, the company has treated the operating system with something bordering on contempt. When asked about Acer's long-gestating RT device, Acer president Jim Wong said "to be honest, there's no value doing [hardware for] the current version of RT." Given the underwhelming interest in RT gear that other companies have reported, we're not sure if Wong's comments qualify as a sick burn or merely kicking an adolescent piece of software when it's down.