UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA
KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY

DIPLOMOVÁ PRÁCE

Fraktální analýza textu

Vedoucí diplomové práce:
prof. RNDr. dr hab. Jan Andres, CSc.
Rok odevzdání: 2010

Vypracoval:
Marcela Martinů
AME, II. ročník

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracovala samostatně pod vedením pana prof. RNDr. dr hab. Jana Andrese, CSc. s použitím uvedené literatury.

V Olomouci dne 8. dubna 2010

Poděkování

Na tomto místě bych chtěla poděkovat především svému vedoucímu diplomové práce panu prof. RNDr. dr hab. Janu Andresovi, CSc. za odbornou pomoc a cenné připomínky. Dále bych chtěla poděkovat paní Mgr. Martině Benešové, která mi pomáhala s lingvistickou částí práce a také paní Mgr. Janě Vrbkové za pomoc se softwarem R.

Obsah

Úvod

V práci, nazvané *Fraktální analýza textu*, se pokusíme o fraktální analýzu novinového článku. Vypočítáme odpovídající fraktální dimenze na různých úrovních jazykových struktur, zkonstruujeme fraktál s vypočtenou dimenzí a na závěr provedeme vizualizace jazykových struktur, jakožto aproximace zmíněného fraktálu.

Práce bude rozdělena do několika částí. Nejprve vytvoříme tabulky pro novinový článek na jazykových úrovních: fonémy / slabiky / slova / klauze / sémantické konstrukty, kde také uvedeme definice těchto lingvistických pojmů. Z tabulek, ze kterých získáme hodnoty potřebné pro výpočet parametrů z Menzerathova - Altmannova zákona, odhadneme požadované parametry metodou nejmenších čtverců. K tomu využijeme softwaru R (algoritmy uvedeme v přílohách). Dále provedeme statistickou analýzu, sestrojíme intervaly spolehlivosti pro odhadnuté parametry. V další části práce se zaměříme na fraktální analýzu. Nadefinujeme fraktální dimenzi, míru sémantičnosti a zkonstruujeme fraktál s vypočtenou dimenzí. Tento fraktál sestrojený pro novinový článek budeme vizualizovat. Aproximacemi matematického fraktálu získáme jazykový fraktál.

Fraktál je nekonečně členitý geometrický objekt s vlastností soběpodobnosti, který je generován opakovaným použitím jednoduchých pravidel. Mezi typické představitele fraktálů patří Sierpinského trojúhelník, Cantorův prach či Von Kochova křivka. V naší práci budeme zkoumat tzv. *jazykové fraktály.* Jedná se o objekty, jejichž vizualizace odpovídají aproximacím. Rozlišujeme jazykový fraktál v silném smyslu a jazykový fraktál v slabém smyslu. Jazykový fraktál v silném smyslu podléhá Menzerathovu - Altmannovu zákonu na všech úrovních stejně, zatímco jazykový fraktál v slabém smyslu nikoli [1].

Jak jsme již uvedli, budeme odhadovat parametry z Menzerthova - Altmannova zákona. P. Menzerath v roce 1928 zjistil, že existují vztahy mezi délkou slov ve slabikách a délkou slabik ve fonémech. Tento vztah formuloval následovně.

Čím delší slovo v počtu slabik, tím kratší průměrná délka jeho slabik [2].

G. Altmann, zakladatel kvantitativní lingvistiky, později toto tvrzení zobecnil.

Čím delší jazykový konstrukt, tím kratší jsou jeho konstituenty [2].

Konstrukt i konstituent jsou relativní pojmy. Konstrukt představuje jazykovou jednotku na vyšší úrovni, zatímco konstituent reprezentuje jazykovou jednotku na nižší úrovni. Například slovo je konstruktem vzhledem ke slabice a konstituentem vzhledem ke klauzi.

G. Altmann formuloval své tvrzení matematicky a pojmenoval ho Menzerathův - Altmannův zákon

$$y = Ax^{-b},$$

kde x je délka konstruktu, y délka konstituentu a A, b jsou reálné parametry. Zpřesňující matematická formulace tohoto zákona je následující

$$y = Ax^{-b}e^{cx},$$

kde x je opět délka konstruktu, y délka konstituentu a A, b, c jsou reálné parametry.

V experimentu se pokusíme vypočítat parametry pomocí obou formulací tohoto zákona na třech jazykových úrovních: slova (počítají se ve slabikách) - slabiky (počítají se ve fonémech), klauze (počítají se ve slovech) - slova (počítají se ve slabikách) a sémantické konstrukty (počítají se v klauzích) - klauze (počítají se ve slovech).

Hlavním cílem práce bude zpracovat novinový článek nazvaný *"Investiční životní pojištění je v Česku stále populárnější - vydělává totiž"* ze Svitavského deníku ze dne 26. října 2009. Využijeme k tomu zmíněný Menzerathův - Altmannův zákon. Hlavní úlohu bude hrát parametr b. Jeho převrácenou hodnotu budeme interpretovat jako fraktální dimenzi na příslušné jazykové úrovni. Ukážeme, že tento zákon platí na všech úrovních.

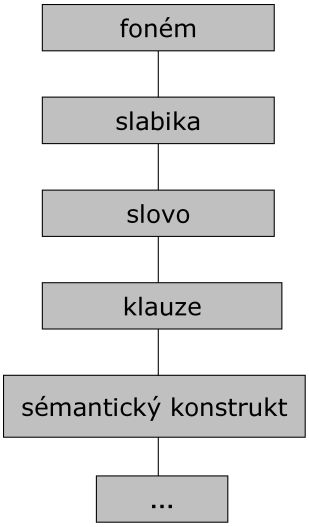
Pokusíme se provést analýzu také pro SMS zprávy (použijeme asi 100 textových zpráv). Postup bude stejný jako v případě zkoumání novinového článku. Sestavíme tabulky na všech uvedených jazykových úrovních a pomocí metody

nejmenších čtverců v programu R odhadneme požadované parametry z Menzerathova - Altmannova zákona.

Na základě zjištěných parametrů, které nemusí vždy vycházet kladně, dojdeme k závěru, že nelze tento zákon aplikovat na výchozí text. Bude nutno zkoumaný text pročíst a zjistit, zda nebude potřeba provést nějaké úpravy nebo určit pravidla, jak s daným textem pracovat.

1. Tabulky a definice lingvistických pojmů

Budeme analyzovat text na různých jazykových úrovních. Zde uvedeme pouze základní členění, které bude pro tuto práci stačit (viz např. v [2]).



Obr.1.: Jazykové úrovně

Foném je nejmenší jednotkou řeči se schopností rozlišit význam. *Jednotlivé fonémy a jejich hláskové konkretizace se mohou stát nositeli významu, ať už ve spojení s jinými fonémy nebo samy o sobě. To však neznamená, že by už fonémy jako jednotky fonologického systému byly samy znaky. V češtině jednofonémová slova jsou nositeli jen formálních významů. Jsou to spojky (a, i) a předložky, jež se skládají jednak ze samohlásek (o, u), jednak ze souhlásek (k, s, z). Souhlásková slova se těsně přimykají k slovu následujícímu (např. k lesu), z hlediska fonetického jde o jedno slovo* (viz [3]).

Slabika je nejmenší jednotkou řeči, v níž dochází k tak těsnému spojení jejích složek, že při rozčlenění proudu řeči nejsme schopni rozdělit je do úseků menších, které by umožnily ještě řeči rozumět. Přestože každý uživatel jazyka je schopen bez obtíží rozčlenit řeč nebo slovo na slabiky, není dosud shody v názoru na podstatu slabiky. Byly definovány pouze různé složky slabiky a jevy, které tvoření slabiky doprovázejí. Z dosavadních teorií jsou nejrozšířenější teorie o podstatě slabiky:

výdechová, motorická, sonoritní a artikulační. Žádná z nich sama o sobě však podstatu slabiky jednoznačně nevysvětluje. Podmínkou pro vznik slabiky je zaznění hlasu. Nejjednodušším příkladem slabiky je těsné spojení souhlásky a samohlásky (viz [3]).

Slovo je základní jednotkou lexikální jazykové roviny jazyka, která je tvořena řadou fonémů (výjimečně fonémem jediným) a nese lexikální a gramatický význam. Slovo má různé významy, pokud bereme v úvahu různé jazykové úrovně. Na slovo můžeme pohlížet ze dvou úhlů. Slovo je konstruktem a konstituenty jsou slabiky, tedy bereme dvojici x slov - y slabik nebo slovo je konstituentem a jeho konstruktem je klauze, tedy x klauzí - y slov a x sémantických konstruktů - y klauzí. V prvním případě se jedná o fonologickou úroveň, ve druhém případě o syntaktickou úroveň. Slovo můžeme považovat za skupinu morfémů (kořen, předpona, přípona, koncovka, atd.) nebo jako lexém, který představuje množinu všech tvarů určitého slova. Např. lexém "*životní*" může v textu nabývat podob *životní, životního, životním, ...*. Toto je důležité pro všechny jazyky, ve kterých skloňujeme, český jazyk není výjimkou (viz např. [4]).

Věta (klauze) představuje uzavřenou jednotku řeči, která se skládá z jednoho nebo více slov. Rozlišujeme stránku obsahovou, mluvnickou (jazykové prostředky použité k vyjádření obsahu), modální (postoj mluvčího k výpovědi) a zvukovou nebo grafickou. To, co dělá ze slov větu, se nazývá predikát [7].

Nyní se podíváme na poslední uvedenou jazykovou úroveň a to na sémantický konstrukt. Tato jazyková úroveň dovršila představu o struktuře textu. Luděk Hřebíček tyto struktury nazval *"agregátem"* a Gabriel Altmann používal pojem *"hreb"*. Pojmy se příliš neujaly, budeme používat *"sémantický konstrukt"*. Sémantické konstrukty jsou konstrukty a věty jsou jako jejich konstituenty na základě rozlišení dvou druhů kontextů nějaké lexikální jednotky (lexému). Existuje užší kontext a širší kontext. Užší kontext, který je tvořen nějakou syntaktickou konstrukcí, budeme uvažovat větu nebo klauzi. A širší kontext, který tvoří všechny věty daného textu, v nichž se vyskytuje daná lexikální jednotka (viz [2]).

1.1. Tabulky pro novinový článek

Ukazuje se, že pro práci s textem je třeba dodat další požadavky, abychom mohli využít Menzerathův - Altmannův zákon. Slova, která mají funkci gramatických modifikátorů, bez ohledu na jejich pravopis, byla počítána jako část slov, ke kterým se tato slova vztahovala. Předložka před jménem byla počítána jako jednotka spolu s následujícím slovem. Nerozlišuje se, jestli se jedná o slovo řídicí či nikoliv. Vybíráme slovo bezprostředně následující. Například uvedeme spojení vyskytující se v našem novinovém článku "v čem", "o životní" nebo "na zdejším". Tato slovní spojení byla počítána vždy jako jedno slovo.

V novinovém článku bylo přibližně 40 procent jednoslabičných slov předložkou. Tyto předložky byly tvořeny jedním nebo maximálně dvěma fonémy. Novým spojením předložky se slovem následujícím se v mnoha případech nezvýšil počet slabik, protože se mnohdy jednalo o neslabičné předložky. Například "v čem" - i spojením jsme zachovali počet slabik následujícího slova, ale u spojení "o životní" - jsme získali u následujícího slova o slabiku navíc. Toto spojování jsme využili při výpočtech na všech uvedených jazykových úrovních.

Z novinového článku *"Investiční životní pojištění je v Česku stále populárnější - vydělává totiž"* (viz příloha 1.) jsme vytvořili tabulky na daných jazykových hladinách.

Tabulka 1. je pro jazykovou úroveň slova - slabiky. Slova se počítají ve slabikách a slabiky se počítají ve fonémech. Zde x značí počet n-slabičných slov, v našem případě $n = 1, ..., 6$, z je frekvence, kolikrát se vyskytlo n-slabičné slovo a y je průměrná délka slabiky ve fonémech.

x	z	y
1	115	2,486956520
2	181	2,439226519
3	176	2,354166667
4	108	2,296296296
5	30	2,220000000
6	2	2,333333333

Tabulka 1.: slova - slabiky

Následující tabulka 2. je vytvořena na jazykové úrovni klauze - slova. Klauze se počítají ve slovech a slova se počítají ve slabikách. Zde x udává počet n-slovných vět (u nás $n = 3, 4, \ldots$), z je četnost, kolik se v článku vyskytlo x-slovných vět a y je průměrná délka slova v počtu slabik.

x	**z**	**y**
3	4	2,666666667
4	8	2,531250000
5	12	2,450000000
6	11	2,469696970
7	10	2,485714286
8	9	2,541666670
9	8	2,666666667
10	8	2,787500000
11	3	2,757575758
12	4	2,770833333
13	4	2,634615385
15	1	2,866666667

Tabulka 2.: klauze - slova

Tabulka 3. sestavená pro novinový článek je vytvořena na hladině sémantické konstrukty - klauze. Sémantické konstrukty se počítají v klauzích a klauze se počítají v počtu slov. Zde x představuje četnost každého lexému, z je počet slov objevujících se v textu s danou četností a y je průměrná délka klauze pro každý lexém. Např. v novinovém článku se slovo "investiční" vyskytlo devatenáctkrát a je tam 220 slov vyskytujících se jedenkrát.

x	**z**	**y**
1	220	8,395454545
2	64	8,390625000
3	17	8,588235294
4	11	9,090909091
5	3	8,533333333
6	4	7,416666667
7	2	7,571428571
10	1	7,000000000
11	1	8,636363636
12	1	8,750000000
13	1	8,846153846
15	2	9,666666667
18	1	7,666666667
19	1	9,526315789

Tabulka 3.: sémantické konstrukty - klauze

1.2. Tabulky pro SMS zprávy

Tabulky pro SMS zprávy jsme sestavili stejně jako pro novinový článek. Konkrétně, počítali jsme předložku a slovo bezprostředně následující jako jedno slovo. Protože hodnoty parametrů vycházely záporně, zkusili jsme SMS zprávy přepočítat podle původní metody. Předložky a slova následující byly počítány zvlášť.

V tabulce 4. vidíme hodnoty vypočtené na hladině slova (počítána ve slabikách) - slabiky (počítány ve fonémech).

x	z	y
1	900	2.351111111
2	815	2,333128834
3	239	2,292887029
4	46	2,173913043
5	10	2,200000000
6	1	2,833333333

Tabulka 4.: slova - slabiky

Další tabulka 5. je vytvořena na hladině klauze (počítány ve slovech) - slova (počítána ve slabikách).

x	z	y
1	2	1,500000000
2	34	1,808823529
3	71	1,779342723
4	95	1,802631579
5	99	1,694949495
6	48	1,756944444
7	49	1,688046647
8	16	1,726562500
9	7	1,650793651
10	2	1,750000000
11	1	1,363636364

Tabulka 5.: klauze - slova

Poslední tabulka 6. pro SMS zprávy je sestavena na hladině sémantické konstrukty (počítány v klauzích) - klauze (počítány ve větách).

x	**z**	**y**
1	323	5,046439628
2	95	5,105263158
3	32	4,875000000
4	24	5,270833333
5	11	4,945454545
6	5	4,966666666
7	7	5,673469388
8	6	5,562500000
9	7	5,460317460
10	3	5,400000000
11	5	4,654545455
12	2	4,750000000
14	2	5,214285714
15	1	5,600000000
17	4	5,588235294
20	1	5,300000000
22	1	5,636363636
27	1	6,296296296
31	1	5,290322581
35	1	5,142857143
40	1	5,575000000
42	1	5,833333333
45	1	4,888888889
47	1	6,255319149
48	1	4,791666667
50	1	5,420000000
62	1	5,596774194
77	2	5,240259740
79	1	6,037974684
91	1	4,747252747

Tabulka 6.: sémantické konstrukty - klauze

2. Numerická analýza

V této části práce se zaměříme na metodu nejmenších čtverců (MNČ), pomocí níž vypočítáme parametry z Menzerathova - Altmannova zákona, a to ze všech výše uvedených tabulek.

MNČ je jedna z numerických metod. Pomocí této metody aproximujeme funkce, které jsou zadány tabulkou, jsou-li hodnoty funkce zatíženy chybami nebo je-li jich velký počet. Hledaná funkce bude kombinací předem známých funkcí a metodou nejmenších čtverců budeme schopni vypočítat její koeficienty.

Předpokládejme, že máme funkci $f(x)$ a $\{x_i\}, i = 1, \ldots, n$, posloupnost bodů neboli argumentů, ve kterých jsme naměřili hodnoty funkce $f(x)$, které budou obecně zatíženy chybami. Přesnou hodnotu $f(x_i)$ v bodě x_i označíme f_i a naměřenou hodnotu v bodě x_i označíme $\bar{f}_i$. Položme $E_i = f_i - \bar{f}_i$.

Nechť $\{\varphi_j(x)\}, j = 0, 1, \ldots$, je posloupnost funkcí definovaných pro každé x_i. Cílem je aproximovat $\bar{f}_i$ lineární kombinací funkcí $\{\varphi_j(x)\}$ a to následovně

$$\bar{f}_i \approx \sum_{j=0}^{m} a_j^{(m)} \varphi_j(x_i), i = 1, \ldots, n.$$

Koeficienty $a_j^{(m)}$ určíme tak, aby výraz

$$H\left(a_0^{(m)}, \ldots, a_m^{(m)}\right) = \sum_{i=1}^{n} w(x_i) \left[\bar{f}_i - \sum_{j=0}^{m} a_j^{(m)} \varphi_j(x_i)\right]^2 \tag{2.1}$$

$$= \sum_{i=1}^{n} w(x_i) R_i^2$$

byl minimální. Funkci $w(x_i)$ nazýváme váhovou funkcí a platí pro ni, že $w(x_i) \geq 0$ pro $i = 1, \ldots, n$. R_i nazýváme reziduum v bodě x_i a index m u $a_j^{(m)}$ značí, že koeficient u $\varphi_j(x)$ závisí na m (nemusí tomu tak být vždy). Pokud koeficienty $a_j^{(m)}$ jsou určeny tak, že výraz (2.1) nabývá své nejmenší hodnoty, dostáváme aproximaci

$$y_m(x) = \sum_{j=0}^{m} a_j^{(m)} \varphi_j(x),$$

které říkáme aproximace funkce $f(x)$ nad $\{x_i\}$ metodou nejmenších čtverců.

Koeficienty a_j^m určíme jako minimum výrazu (2.1). Vypočítáme parciální derivace podle $a_k^{(m)}$ pro $k = 0, \ldots, m$ a položíme je rovny nule. Tedy

$$\frac{\partial H}{\partial a_k^{(m)}} = -2 \sum_{i=1}^{n} w_i \left[\bar{f}_i - \sum_{j=0}^{m} a_j^{(m)} \varphi_j (x_i) \right] \varphi_k (x_i) = 0,$$

$$k = 0, \ldots, m, \quad w_i = w(x_i).$$

Dostaneme soustavu $m + 1$ lineárních rovnic o $m + 1$ neznámých $a_j^{(m)}$. Tuto soustavu nazýváme soustavu normálních rovnic. Řešením normální soustavy získáme hledané koeficienty (viz např. [5]).

S využitím MNČ a programu R vypočítáme parametry z MAZ jak pro jednoduchou verzi

$$y = Ax^{-b},$$

tak pro úplnou verzi

$$y = Ax^{-b}e^{cx}.$$

2.1. Výpočet parametrů pro novinový článek

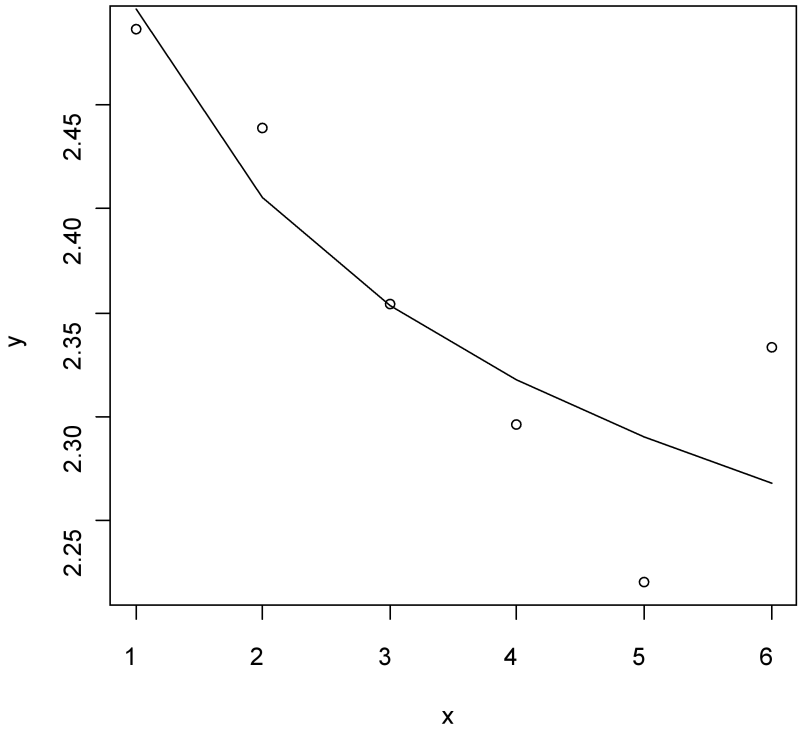
Na hladině slova - slabiky (tabulka 1.) provedeme logaritmizaci, abychom zjistili počáteční hodnoty pro MNČ. Pro jednoduchou verzi MAZ vyjdou parametry

$$A = 2,496176, \quad b = 0,05374.$$

Nyní můžeme použít nelineární MNČ. Pomocí ní nám vyjdou parametry

$$A = 2,49632, \quad b = 0,05363.$$

Grafické znázornění vypadá takto:



Graf 1.: nelineární MNČ, jednoduchá verze MAZ

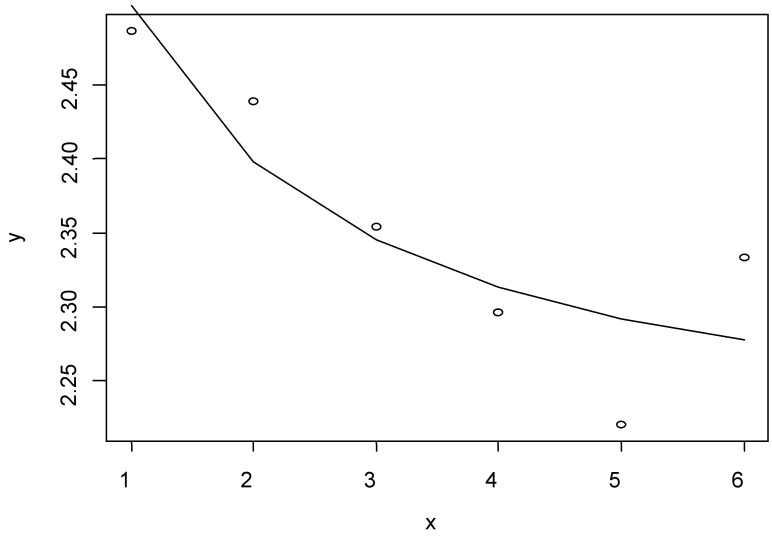
Pro úplnou verzi MAZ nám úvodní logaritmizace dala tyto výsledky.

$$A = 2,485838, \quad b = 0,076224, \quad c = 0,008232.$$

Z toho opět pomocí MNČ vyšlo

$$A = 2,486641, \quad b = 0,072443, \quad c = 0,007005.$$

Graf je následující:



Graf 2.: nelineární MNČ, úplná verze MAZ

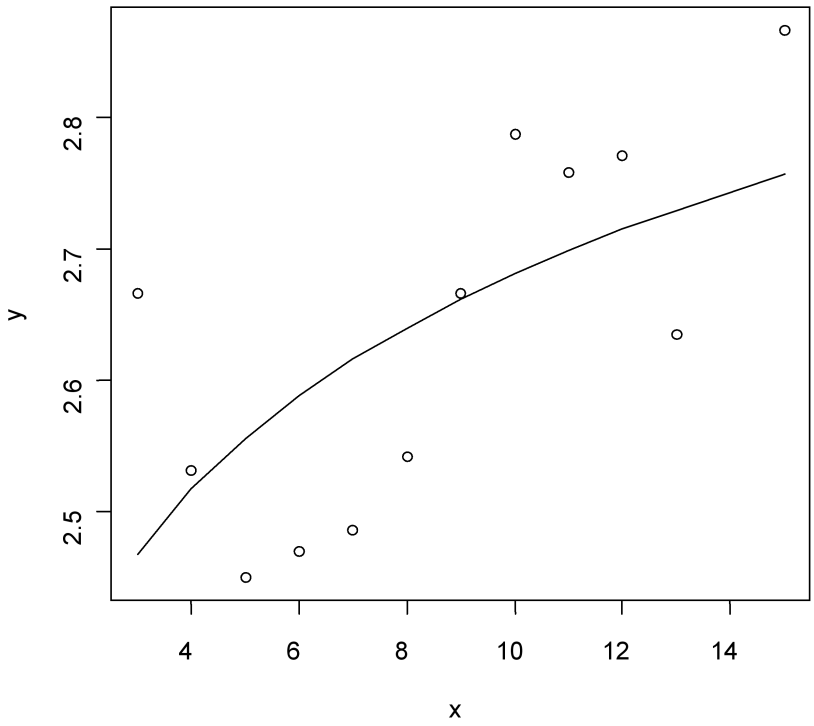
Nyní se zaměříme na tabulku 2. sestavenou pro jazykovou úroveň klauze - slova. Nejprve provedeme logaritmizaci u jednoduché verze MAZ. Získáme hodnoty parametrů

$$A = 2,301175, \quad b = -0,06567.$$

A pomocí MNČ vychází parametry

$$A = 2,28888, \quad b = -0,06866.$$

Graficky:



Graf 3.: nelineární MNČ, jednoduchá verze MAZ

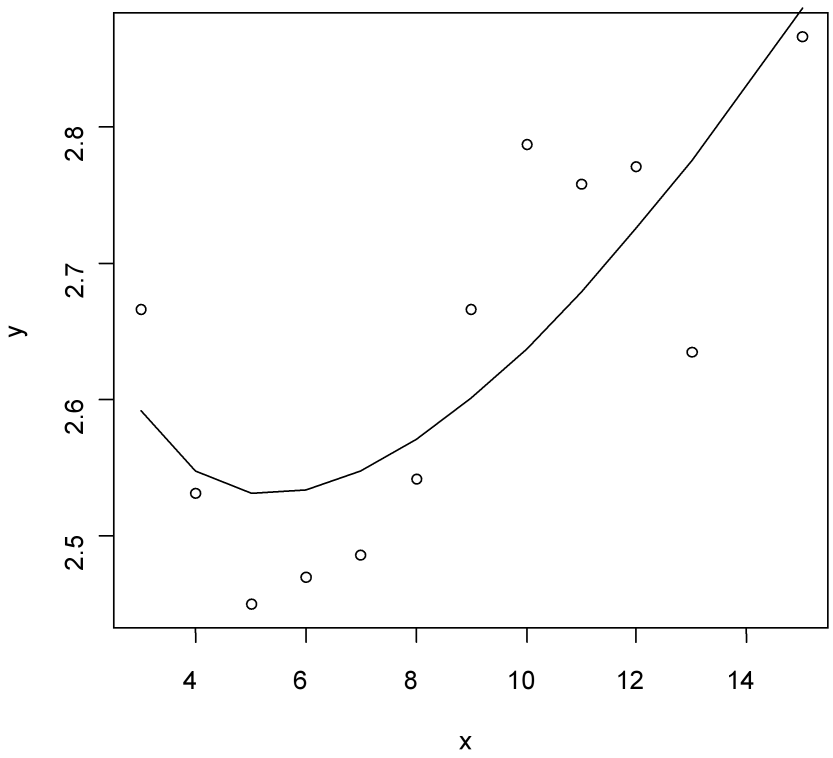
U úplné verze MAZ na jazykové hladině klauze - slova vyjdou parametry z logaritmizace

$$A = 2,859625, \quad b = 0,18043, \quad c = 0,03339$$

a numerickou MNČ vychází

$$A = 2,84268, \quad b = 0,17141, \quad c = 0,03199.$$

A opět uvedeme graf získaný z programu R:



Graf 4.: nelineární MNČ, úplná verze MAZ

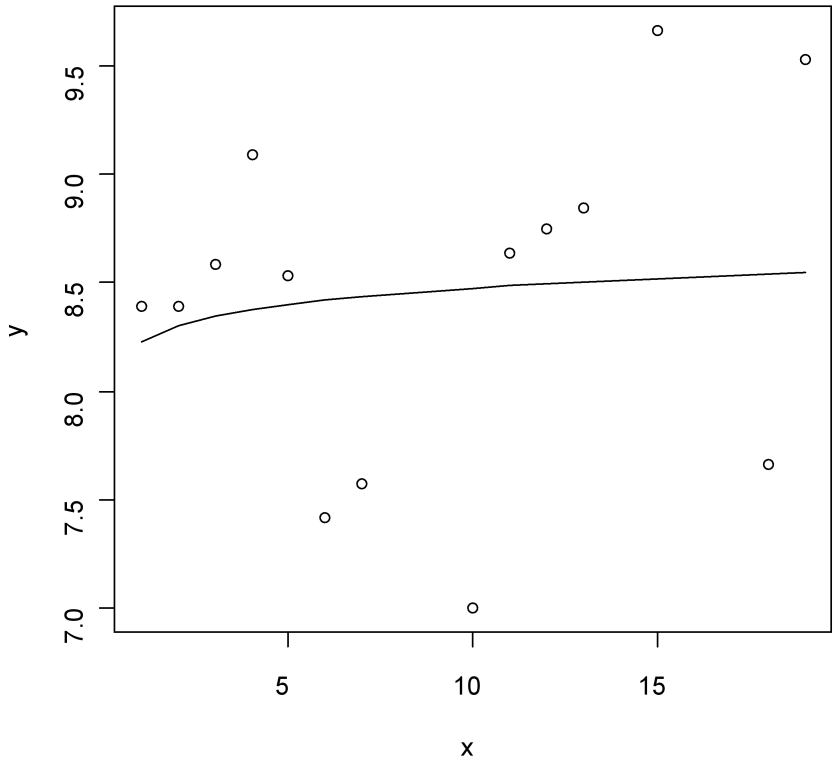
Z poslední tabulky 3. vytvořené na jazykové hladině sémantické konstrukty - klauze jsme získali následující hodnoty koeficientů. Pro jednoduchou verzi MAZ z logaritmizace

$$A = 8,238267, \quad b = -0,01014$$

a z numerické MNČ

$$A = 8,22585, \quad b = -0,01303.$$

Grafické znázornění:



Graf 5.: nelineární MNČ, jednoduchá verze MAZ

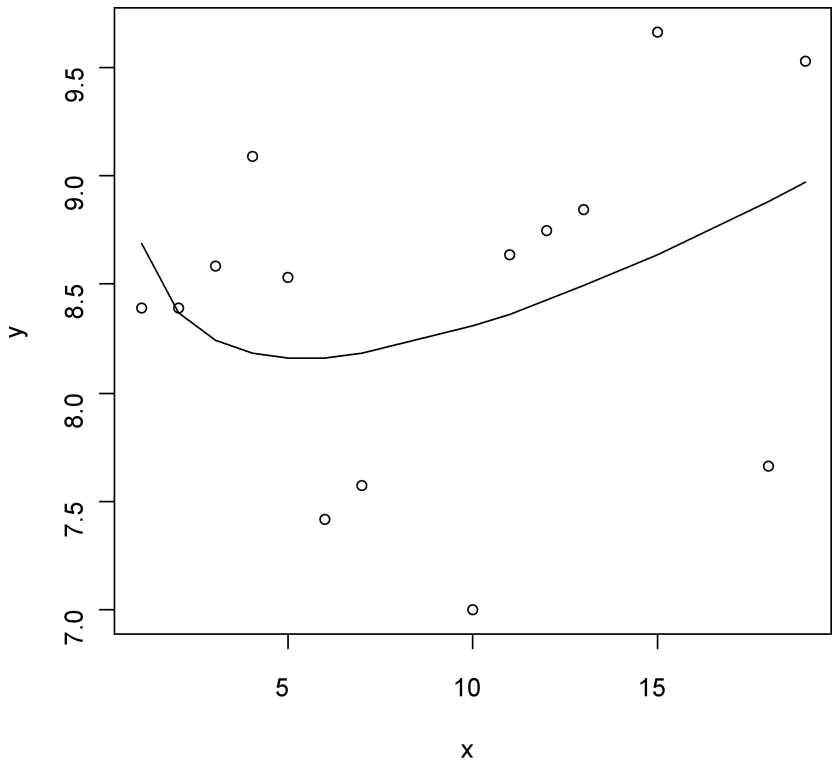
Na stejné jazykové úrovni pro úplnou formuli MAZ získáváme z logaritmizace hodnoty koeficientů

$$A = 8,595727, \quad b = 0,07906, \quad c = 0,014027$$

a z nelineární MNČ

$$A = 8,56823, \quad b = 0,07311, \quad c = 0,01376.$$

Grafické znázornění pro úplnou verzi MAZ:



Graf 6.: nelineární MNČ, úplná verze MAZ

2.2. Výpočet parametrů pro SMS zprávy

Nyní opět vypočítáme parametry z Menzerathova - Altmannova zákona, z jeho jednoduché i úplné formule. Provedeme logaritmizaci, abychom získali startovací hodnoty pro MNČ. Výsledky shrneme zkráceně do tabulky.

V mnoha případech vycházely záporné koeficienty b. Proto v další části práce již nebudeme provádět pro SMS zprávy statistickou ani fraktální analýzu.

sémantické konstrukty - klauze	A	b	c
jednoduchá verze MAZ - logaritmizace	5,031379	-0,019893	-
jednoduchá verze MAZ - MNČ	5,03506	-0,02064	-
úplná verze MAZ - logaritmizace	4,868045	-0,044102	-0,0012641
úplná verze MAZ - MNČ	4,854438	-0,046414	-0,001301
klauze - slova			
jednoduchá verze MAZ - logaritmizace	1,704704	0,0098109	-
jednoduchá verze MAZ - MNČ	1,706874	0,008567	-
úplná verze MAZ - logaritmizace	1,628977	-0,256946	-0,063167
úplná verze MAZ - MNČ	1,63508	-0,24019	-0,05909
slova - slabiky			
jednoduchá verze MAZ - logaritmizace	2,452358	0,10324	-
jednoduchá verze MAZ - MNČ	2,43411	0,09477	-
úplná verze MAZ - logaritmizace	2,586119	-0,184374	-0,1052837
úplná verze MAZ - MNČ	2,57092	-0,16573	-0,09754

Tabulka 7.: Výsledná tabulka vypočtených parametrů z MAZ

3. Statistická analýza

V následující části práce sestrojíme konfidenční intervaly pro spočítané koeficienty A, b, c získané z MAZ. Určíme interval, v němž hledaný parametr leží s danou pravděpodobností.

$100(1-\alpha)$ procentní interval spolehlivosti (konfidenční interval) pro danou náhodnou veličinu (funkci náhodné veličiny, parametr) je takový interval, ve kterém se s pravděpodobností $1-\alpha$ realizace této dané náhodné veličiny nebo parametru nachází. Nejčastěji se využívají intervaly spolehlivosti $99, 95$ nebo 90 procentní.

Pokud známe rozdělení náhodné veličiny X, vytvoříme konfidenční interval pro X jednoduchým způsobem pomocí kvantilů. Můžeme sestrojit buď jednostranné, nebo oboustranné intervaly spolehlivosti.
Levostranný interval spolehlivosti:

$$(x_\alpha, \infty) \quad \text{tj. } P\{x_\alpha < X\} = 1 - \alpha$$

Pravostranný interval spolehlivosti:

$$(-\infty, x_{1-\alpha}) \quad \text{tj. } P\{X < x_{1-\alpha}\} = 1 - \alpha$$

Oboustranný interval spolehlivosti:

$$(x_{\alpha/2}, x_{1-\alpha/2}) \quad \text{tj. } P\left\{x_{\alpha/2} < X < x_{1-\alpha/2}\right\} = 1 - \alpha.$$

Pokud neznáme rozdělení pravděpodobnosti našeho parametru, tak při sestrojování konfidenčních intervalů zvolíme vhodnou výběrovou charakteristiku, jejíž rozdělení známe (viz např. [6]).

3.1. Určení konfidenčních intervalů pro parametry získané z novinového článku

Pomocí softwaru R jsme vypočítali intervaly spolehlivosti pro jednotlivé koeficienty získané z novinového článku. Intervaly spolehlivosti máme spočítány pro koeficienty A, b ze zjednodušené formule MAZ a pro koeficienty A, b, c ze zpřesňující formule MAZ. Konfidenční intervaly určují přesnost odhadu, máme je

tedy vypočítány pro koeficienty získané z logaritmizace, kterou jsme prováděli z důvodu zjištění počátečních neboli startovacích hodnot, abychom mohli použít MNČ.

Pro každý parametr určíme dolní a horní hranici intervalu. Budeme uvažovat $\alpha = 0,05$. Získáme tak oboustranný odhad na hladinách $\alpha/2$ a $1 - \alpha/2$ (tedy $2,5\%$ a $97,5\%$).

Na jazykové hladině slova - slabiky vyšly oboustranné konfidenční intervaly pro jednoduchou verzi MAZ následovně. Pro koeficient A, který jsme pomocí logaritmizace odhadli jako $A \doteq 2,496$, vyšel interval spolehlivosti přibližně

$$(2,362; 2,630)\,.$$

Pro koeficient b, který jsme odhadli jako $b \doteq 0,05374$, vyšel interval spolehlivosti

$$(0,01109; 0,09639)\,.$$

Na stejné hladině, ale pro koeficienty ze zpřesňující formule MAZ vychází konfideční oblasti přibližně takto.
Pro parametr $A \doteq 2,485$

$$(2,30; 2,67)\,.$$

Pro parametr $b \doteq 0,07622$

$$(-0,14327; 0,29571)\,.$$

Pro parametr $c \doteq 0,00823$

$$(-0,06951; 0,08597)\,.$$

Na jazykové úrovni klauze - slova opět spočítáme oboustranné intervaly spolehlivosti, nejprve pro koeficienty A, b získané z logaritmizace.
Pro parametr $A \doteq 2,301$ vychází přibližně

$$(2,030; 2,572)$$

a pro $b \doteq -0,06567$

$$(-0,12504; -0,0063)\,.$$

Na této hladině určíme také intervaly spolehlivosti pro koeficienty A, b, c. Pro koeficient $A \doteq 2,86$ vyjde po zaokrouhlení

$$(2,32; 3,40)\,.$$

Pro koeficient $b \doteq 0,18043$

$$(-0,03508; 0,39594)\,.$$

Pro koeficient $c \doteq 0,03339$

$$(0,00488; 0,06190)\,.$$

Zbývá určit konfidenční intervaly pro všechny parametry na jazykové úrovni sémantické konstrukty - klauze. Pro koeficienty A, b ze zjednodušené formule MAZ vychází intervaly spolehlivosti následovně.
Pro koeficient $A \doteq 8,24$

$$(7,18; 9,30)\,.$$

Pro koeficient $b \doteq -0,01014$

$$(-0,07749; 0,05721)\,.$$

A opět na hladině sémantické konstrukty - klauze vypočítáme konfidenční intervaly pro úplnou verzi MAZ.
Pro koeficient $A \doteq 8,596$

$$(7,3; 9,9)\,.$$

Pro koeficient $b \doteq 0,07906$

$$(-0,10556; 0,26368)\,.$$

Pro koeficient $c \doteq -0,01427$

$$(-0,01324; 0,04178)\,.$$

4. Fraktální analýza

V této části práce se zaměříme na fraktální analýzu, budeme interpretovat převrácené hodnoty

$$D_i = \frac{1}{b_i}$$

parametrů $b_i, i = 1, \ldots, n$ jako *fraktální dimenze matematických fraktálů*. Obecněji můžeme psát

$$D = \frac{n}{b_1 + \ldots + b_n}.$$

Dále nadefinujeme *míru sémantičnosti jazykových objektů* jako

$$D \in [D_{min}, D_{max}],$$

kde $D_{min} := \min D_i, D_{max} := \max D_i, i = 1, \ldots, n$. Budeme vycházet z [1].

Dá se jednoduše ukázat, že úplnou verzi Menzerathova - Altmannova zákona na n-jazykových úrovních

$$y_i = A_i x_i^{-b_i} e^{c_i x_i}, i = 1, \ldots, n$$

lze také ekvivalentně vyjádřit jako

$$\frac{1}{b_i} = \frac{\log x_i}{\log(\frac{A_i}{y_i} e^{c_i x_i})} = \frac{\ln x_i}{\ln(\frac{A_i}{y_i} e^{c_i x_i})}$$

pro $i = 1, \ldots, n$.

Zjednodušená verze MAZ ($c_i = 0$, $i = 1, \ldots, n$) vypadá takto

$$y_i = A_i x_i^{-b_i}, \quad i = 1, \ldots, n$$

a jeho ekvivalentní formule je ve tvaru

$$\frac{1}{b_i} = \frac{\log x_i}{\log \frac{A_i}{y_i}}, \quad i = 1, \ldots, n.$$

Zde se zmíníme o známém *Moran - Hutchinsonově vzorci* pro výpočet fraktální dimenze. Z tohoto vzorce uvidíme, proč můžeme interpretovat převrácené

hodnoty parametrů $b_i, i = 1, \ldots, n$ z Menzerathova - Altmannova zákona jako fraktální dimenze. Moran - Hutchinsonův vzorec, jehož předpokladem je totálně nesouvislý fraktál nebo jen se dotýkající části fraktálu vypadá takto

$$mr^D = 1.$$

Pokud z něho vyjádříme fraktální dimenzi D, tak dostáváme

$$D = \frac{\log m}{\log \frac{1}{r}},$$

kde r je koeficient kontrakce a m značí počet kontrahujících zobrazení v iterovaném funkčním systému. Z výše uvedené ekvivalentní úpravy MAZ vidíme, že $m_i \sim x_i$ a $r_i \sim \frac{y_i}{A_i}$.

Z Moran-Hutchinsonovy formule pro fraktální dimenzi D můžeme interpretovat převrácený aritmetický průměr $\frac{n}{b_1+\ldots+b_n}$ koeficientů $b_1, \ldots, b_n$ jako dimenzi $D = dim(\mathbf{A})$ vhodného cyklicky soběpodobného fraktálu $\mathbf{A}$,

$$D := \frac{n}{b_1 + \ldots + b_n}. \tag{4.1}$$

Pro $x := x_1 = x_2 = \ldots = x_n$ a $r_i := (\frac{y_i}{A_i e^{c_i x}})^k = \frac{1}{x^{kb_i}}, i = 1, \ldots, n$. (Z toho plyne $r_1 \ldots r_n := (\frac{y_1 \ldots y_n}{A_1 \ldots A_n e^{((c_1+\ldots+c_n)x)}})^k = \frac{1}{x^{k(b_1+\ldots+b_n)}}$), kde $k \geq \max \frac{1}{b_i}, k \in \mathbb{N}, i = 1 \ldots, n$, fraktál $\mathbf{A}$ může být považován za jedinou uzavřenou pozitivně invariantní množinu $\mathbf{A} = F(\mathbf{A})$ kompozice $F = F_n \circ \ldots \circ F_1$ z *Hutchinson - Barnsleyho zobrazení* F_i, kde

$$F_i(\mathbf{x}) := \bigcup_{\mathbf{j}} {}_i f_{\mathbf{j}}(\mathbf{x}),$$

$${}_i f_{\mathbf{j}} : [0,1]^k \to [0,1]^k,$$

$${}_i f_{\mathbf{j}}(\mathbf{x}) := r_i \mathbf{x} + \frac{1}{x}\mathbf{j},$$

$\mathbf{j} = (j_1, \ldots, j_k), j_{\mathbf{l}} \in \{0, 1, \ldots, x-1\}, i = 1, \ldots, n$ a r je koeficient kontrakce.

Dále ještě může být získán pomocí limity (s použitím *Hausdorffovy metriky* d_H) z postupných aproximací $F^0([0,1]), F^m([0,1]), m = 1, 2, \ldots$ fraktálu $\mathbf{A}$. To je

$$\lim_{m \to \infty} d_H(F^m([0,1]), \mathbf{A}) = 0,$$

kde Hausdorffova vzdálenost $d_H(F^m([0,1]), \mathbf{A})$ mezi aproximacemi $F^m([0,1])$ a $\mathbf{A}$ může být odhadnuta následovně

$$d_H(F^m([0,1]), \mathbf{A}) \leq \frac{(r_1 \ldots r_n)^m}{1 - r_1 \ldots r_n} d_H([0,1], F([0,1])) =$$

$$= \frac{((1 - \frac{1}{x}) + (1 - \frac{1}{x}) \sum_{i=2}^{n} r_2 \ldots r_i)\sqrt{k-1}}{x^{mk(b_1 + \ldots + b_n)}(1 - x^{-k(b_1 \ldots + b_n)})}.$$

Poznamenejme, že pro $A := A_1 = A_2 = \ldots = A_n$, $b := b_1 = b_2 \ldots = b_n$ a $c := c_1 = c_2 = \ldots = c_n$ hodnota $\frac{1}{b}$ může být jednoduše interpretována jako *fraktální dimenze* $\mathbf{A} = F_1(\mathbf{A}) = F_2(\mathbf{A}) = \ldots = F_n(\mathbf{A})$. Tedy

$$D = \frac{1}{b}. \tag{4.2}$$

Protože, s ohledem na uvedené výše, máme

$$r := r_1 = r_2 = \ldots = r_n = (\frac{y}{Ae^{cx}})^k = \frac{1}{x^{kb}} \leq \frac{1}{x}.$$

Pro redukovanou formuli MAZ ($c = 0$) platí, že

$$r := (\frac{y}{A})^k = \frac{1}{x^{kb}}.$$

Fraktální dimenzi $D^{(p)}$ p-rozměrné projekce $\mathbf{A}$ lze vypočítat jako

$$D^{(p)} = \frac{p}{k} D.$$

Jednodušší způsob pro konstrukci fraktálu s danou dimenzí je následující. Pro dané číslo $D > 0$ můžeme zkonstruovat soběpodobnostní fraktál, jehož dimenze $D \leq k, k \in \mathbb{N}$, jako jedinou uzavřenou pozitivně invariantní množinu $\mathbf{A}$ iterovaného funkčního systému

$$f_{\mathbf{j}} : [0,1]^k \to [0,1]^k,$$

$$\mathbf{j} = (j_1, \ldots, j_k), j_\mathbf{l} \in \{0, 1, \ldots, x-1\},$$

kde

$$f_\mathbf{j}(\mathbf{x}) := r\mathbf{x} + \frac{1}{x}\mathbf{i}$$

a koeficient kontrakce $r = \frac{1}{x^{k/D}}$, k je nejmenší kladné celé číslo větší nebo rovno než D. Za D vezmeme $D_{\min}$ z $D_i, i = 1, \ldots, n$. Dle definice Hutchinson - Barnsleyho zobrazení

$$F(\mathbf{x}) = \bigcup_\mathbf{j} f_\mathbf{j}(\mathbf{x})$$

uzavřená pozitivně invariantní množina $\mathbf{A}$ splňuje rovnost $\mathbf{A} = F(\mathbf{A})$. Pro odhad Hausdorffovy vzdálenosti mezi postupnými aproximacemi $\mathbf{A}_m$ a $\mathbf{A}$ platí

$$d_H(\mathbf{A}_m, \mathbf{A}) = d_H(F^m([0,1]), \mathbf{A}) \leq \frac{r^m}{1-r} d_H([0,1], F([0,1])) =$$

$$= \frac{\left(1 - \frac{1}{x}\right)\sqrt{k-1}}{\left(1 - \frac{1}{x^{k/D}}\right) x^{mk/D}}$$

Pro $x = 2$ dostáváme

$$d_H(F^m([0,1]), \mathbf{A}) \leq \frac{\frac{\sqrt{k-1}}{2}}{\left(1 - \frac{1}{2^{k/D}}\right) 2^{mk/D}}.$$

Fraktální dimenze $D^{(2)}$ 2 - dimenzionální projekce $\mathbf{A}$ se vypočítá

$$D^{(2)} = \frac{2}{k} D.$$

Nyní uvedeme definice týkající se sémantičnosti. *Sémantika* byla charakterizována mnoha autory jako "čtení mezi řádky". V experimentu zjistíme, jak velký je vliv sémantičnosti u novinového článku. Nejprve musíme rozlišit dva druhy jazykových fraktálů.

Definice 1. Jazykové fraktály v silném smyslu podléhají Menzerathovu - Altmannovu zákonu na všech úrovních stejně. V opačném případě je nazýváme jazykovými fraktály v slabém smyslu.

Definice 2. Pro jazykové fraktály vyšších řádů v silném smyslu, s koeficientem $b = b_1 = \ldots = b_n$, definujeme (vyjmeme-li úrovně slabik a fonémů) jejich míru sémantičnosti

$$D = \frac{1}{b}$$

jako fraktální dimenzi aproximovaného matematického modelu.

Definice 3. Pro jazykové fraktály v slabém smyslu (vyjmeme-li úrovně slabik a fonémů) charakterizované koeficienty $b_1, \ldots, b_n$ definujeme míru sémantičnosti

$$D = \frac{n}{b_1 + \ldots + b_n}$$

jako převrácenou hodnotu aritmetického průměru hodnot koeficientů $b_1, \ldots, b_n$. Míra sémantičnosti D reprezentuje dimenzi jistého aproximovaného matematického modelu.

Podotkněme, že pro $b = b_1 = \ldots = b_n$ se míra sémantičnosti D zjednodušuje na $D = \frac{1}{b}$, odpovídá definici 2.

Obecněji, pro dané lingvistické objekty (s vyloučením úrovní slabik a fonémů) charakterizované koeficienty $b = b_1 = \ldots = b_n$,

$$D_{\min} := \min_{i=1,\ldots,n} \frac{1}{b_i}$$

$$D_{\max} := \max_{i=1,\ldots,n} \frac{1}{b_i}$$

splňuje míra sémantičnosti D nerovnost

$$D_{\min} \leq D \leq D_{\max}.$$

Můžeme říci, že míra sémantičnosti D je přinejmenším rovna $D_{\min}$.

4.1. Výpočet fraktální dimenze pro novinový článek a určení míry sémantičnosti

Zde vypočteme fraktální dimenze ze vztahů (4.1) a (4.2) pro již spočítané parametry z MAZ. Pro větší přehlednost výsledky shrneme do tabulky.

sémantické konstrukty - klauze	A	b	c
jednoduchá verze MAZ - logaritmizace	8,238267	-0,01014	-
jednoduchá verze MAZ - MNČ	8,22585	-0,01303	-
úplná verze MAZ - logaritmizace	8,595727	0,07906	0,01427
úplná verze MAZ - MNČ	8,56823	0,07311	0,01376
klauze - slova			
jednoduchá verze MAZ - logaritmizace	2,301175	-0,06567	-
jednoduchá verze MAZ - MNČ	2,28888	-0,06866	-
úplná verze MAZ - logaritmizace	2,859625	0,18043	0,03339
úplná verze MAZ - MNČ	2,84268	0,17141	0,03199
slova - slabiky			
jednoduchá verze MAZ - logaritmizace	2,496176	0,05374	-
jednoduchá verze MAZ - MNČ	2,49632	0,05363	-
úplná verze MAZ - logaritmizace	2,485838	0,076224	0,008232
úplná verze MAZ - MNČ	2,486641	0,072443	0,007005

Tabulka 8.: Výsledná tabulka vypočtených parametrů z MAZ

Jak jsme zmínili výše, budeme interpretovat převrácené hodnoty parametrů $b_i, i = 1, 2, 3$ jako fraktální dimenze aproximovaných matematických fraktálů. Čili z úplné verze MAZ získáme hodnoty

$$D_1 = \frac{1}{0,07311} \doteq 13,67802$$

$$D_2 = \frac{1}{0,17141} \doteq 5,833965$$

$$D_3 = \frac{1}{0,07244} \doteq 13,80396$$

Dle obecnějšího vzorce získáme fraktální dimenzi D jako

$$D = \frac{n}{b_1 + b_2 + b_3} = \frac{3}{0,07311 + 0,17141 + 0,072443} \doteq 9,464827.$$

U jednoduché verze MAZ na úrovních klauze - slova, sémantické konstrukty - klauze vychází hodnoty D_1, D_2 záporně, na hladině slova - slabiky vychází

$$D_3 = \frac{1}{0,05363} \doteq 18,64628.$$

Dále určíme míru sémantičnosti. Výpočty provedeme pro úplnou verzi MAZ. Pro jazykový fraktál v slabém smyslu se míra sémantičnosti vypočítá jako

$$D = \frac{n}{b_1 + b_2 + b_3} = \frac{3}{0,07311 + 0,17141 + 0,072443} \doteq 9,464827.$$

Obecněji ji můžeme určit tak, že vybereme

$$D_{\min} \doteq 5,833965,$$

$$D_{\max} \doteq 13,80396$$

a pro míru sémantičnosti D platí

$$D_{\min} \leq D \leq D_{\max},$$

tedy

$$5,833965 \leq D \leq 13,80396.$$

Míra sémantičnosti je rovna nejméně $D = 5,833965$.

4.2. Konstrukce fraktálu s vypočtenou dimenzí pro novinový článek - 1. způsob

V této části práce sestrojíme fraktál s předem spočítanou dimenzí. Nejprve ukážeme jednodušší postup.

Máme již spočítané hodnoty

$$D_1 \doteq 13,67802,$$

$$D_2 \doteq 5,833965,$$

$$D_3 \doteq 13,80396.$$

$D_2 \doteq 5,833965$ je dolní odhad D, zkonstruujeme fraktál s touto dimenzí a to následovně.

Vezmeme $x = 2$ a $k = 6$, jako nejmenší kladné celé číslo větší než D. Pro koeficient kontrakce r platí, že

$$r = \frac{1}{x^{k/D}} \doteq \frac{1}{2^{6/5,833965}} \doteq 0,490233.$$

Čili iterovaný funkční systém se skládá z $x^k = 2^6 = 64$ kontrakcí se stejným koeficientem $r \doteq 0,490233$. Dále

$$f_{\mathbf{j}}(\mathbf{x}) := r\mathbf{x} + \frac{1}{x}\mathbf{j} = 0,490233\mathbf{x} + \frac{1}{2}\mathbf{j},$$

$$\mathbf{x} = (x_1, \ldots, x_6) \in [0,1]^6, \quad \mathbf{j} = (j_1, \ldots, j_6), j_l \in \{0,1\}.$$

Dle definice Hutchinson - Barnsleyho zobrazení, tedy

$$F(\mathbf{x}) := \bigcup_{\mathbf{j}} f_{\mathbf{j}}(\mathbf{x}) = \bigcup_{\mathbf{j}} 0,490233\mathbf{x} + \frac{1}{2}\mathbf{j}, \quad \mathbf{x} \in [0,1]^6,$$

uzavřená pozitivně invariantní množina $\mathbf{A}$ splňuje rovnost $\mathbf{A} = F(\mathbf{A})$. Postupné aproximace $\mathbf{A}_m = F^m([0,1]), m = 1, 2, \ldots$ vyhovují odhadům

$$d_H(\mathbf{A}_m, \mathbf{A}) = d_H(F^m([0,1]), \mathbf{A}) \leq \frac{r^m}{1-r} d_H([0,1], F([0,1]))$$

$$\doteq\frac{(0,490233)^m}{0,509767}\sqrt{2}\doteq 2,774235\,(0,490233)^m\,.$$

Rovinná projekce $\mathbf{A}$ má dimenzi

$$D^{(2)}=\frac{2}{6}5,833965\doteq 1,944655.$$

4.3. Konstrukce fraktálu s vypočtenou dimenzí pro novinový článek - 2. způsob

Podle teorie uvedené na začátku této kapitoly nyní opět sestrojíme fraktál pro novinový článek.

Na třech jazykových úrovních jsme získali následující koeficienty,

pro sémantické konstrukty: $b_1\doteq 0,07311$,

pro klauze: $b_2\doteq 0,17141$,

pro slova: $b_3\doteq 0,07244$.

Z toho nám vyšly hodnoty

$$D_1=\frac{1}{0,07311}\doteq 13,67802,$$

$$D_2=\frac{1}{0,17141}\doteq 5,833965,$$

$$D_3=\frac{1}{0,07244}\doteq 13,80396.$$

Zvolíme $x=2$, $k=14$, protože

$$k\geq\max_{i=1,2,3}\frac{1}{b_i}=\max\{13,67802;5,833965;13,80396\}=13,80396$$

a k je přirozené číslo. Míra sémantičnosti $D=\frac{3}{b_1+b_2+b_3}\doteq 9,464827$ je fraktální dimenze uzavřené množiny $\mathbf{A}=F(\mathbf{A})$, kde

$$F=F_3\circ F_2\circ F_1,\quad F_i(\mathbf{x})=\bigcup_{\mathbf{j}}{}_i f_{\mathbf{j}}(\mathbf{x}),\quad i=1,2,3.$$

Nyní vypočítáme koeficienty kontrakce

$$r_1 = \frac{1}{x^{kb_1}} = \frac{1}{2^{14\cdot 0,07311}} \doteq 0,491918,$$

$$r_2 = \frac{1}{x^{kb_2}} = \frac{1}{2^{14\cdot 0,17141}} \doteq 0,189499,$$

$$r_3 = \frac{1}{x^{kb_3}} = \frac{1}{2^{14\cdot 0,07244}} \doteq 0,495117,$$

dostáváme tedy

$${}_1f_{\mathbf{j}}(\mathbf{x}) = 0,491918\mathbf{x} + \frac{1}{2}\mathbf{j}, \mathbf{j} = (j_1, \ldots, j_{14}), j_{\mathbf{l}} \in \{0,1\},$$

$${}_2f_{\mathbf{j}}(\mathbf{x}) = 0,189499\mathbf{x} + \frac{1}{2}\mathbf{j}, \mathbf{j} = (j_1, \ldots, j_{14}), j_{\mathbf{l}} \in \{0,1\},$$

$${}_3f_{\mathbf{j}}(\mathbf{x}) = 0,495117\mathbf{x} + \frac{1}{2}\mathbf{j}, \mathbf{j} = (j_1, \ldots, j_{14}), j_{\mathbf{l}} \in \{0,1\}.$$

A z definice Hutchinson - Barnsleyho zobrazení získáme

$$F_1(\mathbf{x}) = \bigcup_{\mathbf{j}} {}_1f_{\mathbf{j}}(\mathbf{x}) = \bigcup_{\mathbf{j}} 0,491918\mathbf{x} + \frac{1}{2}\mathbf{j},$$

$$F_2(\mathbf{x}) = \bigcup_{\mathbf{j}} {}_2f_{\mathbf{j}}(\mathbf{x}) = \bigcup_{\mathbf{j}} 0,189499\mathbf{x} + \frac{1}{2}\mathbf{j},$$

$$F_3(\mathbf{x}) = \bigcup_{\mathbf{j}} {}_3f_{\mathbf{j}}(\mathbf{x}) = \bigcup_{\mathbf{j}} 0,495117\mathbf{x} + \frac{1}{2}\mathbf{j},$$

kde $\mathbf{x} = (x_1, \ldots, x_{14}) \in [0,1]^{14}$.

Rovinná projekce $\mathbf{A}$ má dimenzi

$$D^{(2)} = \frac{2}{14}9,464827 \doteq 1,35212.$$

Postupné aproximace $\mathbf{A}_m = F^m([0,1]), m = 1,2,\ldots$ vyhovují odhadům

$$d_H(\mathbf{A}_m, \mathbf{A}) = d_H(F^m([0,1]), \mathbf{A}) \leq \frac{(r_1r_2r_3)^m}{1 - r_1r_2r_3} d_H([0,1], F([0,1])) =$$

$$\frac{(0,0461538)^m}{0,9538462}\sqrt{2}\doteq 1,482643(0,0461538)^m.$$

Na závěr této kapitoly uvedeme srovnání při konstrukci fraktálů podle obou způsobů. Podle druhého způsobu je koeficient kontrakce z $F = F_3 \circ F_2 \circ F_1$ roven $r_1 r_2 r_3 = 0,0461538$, zatímco pro $F^3 = F \circ F \circ F$ se rovná $r^3 = \frac{1}{2^{18/5,833965}} = 0,117817$.

Na druhé straně $F = F_3 \circ F_2 \circ F_1$ se skládá z $2^{3 \cdot 14} = 2^{42} \doteq 4,398 \cdot 10^{12}$ zobrazení, kdežto $F^3 = F \circ F \circ F$ se skládá z $2^{3 \cdot 6} = 2^{18} = 262144$ zobrazení.

5. Vizualizace jazykových struktur

V této kapitole se budeme zabývat vizualizacemi jazykových struktur pomocí postupných aproximací matematických fraktálů s vypočtenou dimenzí. Budeme vycházet převážně z [1].

Vzhledem k fraktální analýze (viz výše) můžeme říci, že

$\mathbf{A}_1 := F_1([0,1])$,

$\mathbf{A}_2 := F_2 \circ F_1([0,1])$,

$\vdots$

$\mathbf{A}_n := F_n \circ F_{n-1} \circ \ldots \circ F_1([0,1]) = F([0,1])$,

kde $\mathbf{A} := F([0,1])$ je n-tá postupná aproximace fraktálu $\mathbf{A}$. Aproximace $\mathbf{A}_1, \ldots, \mathbf{A}_n$ můžou být považovány za *vizualizovanou strukturu lingvistických objektů na n-linvistických úrovních charakterizovaných koeficienty* $A_i, b_i, c_i, i = 1, \ldots, n$ z Menzerathova - Altmannova zákona.

Uveďme, že pro mn-té aproximace $\mathbf{A}_{mn} := F^m([0,1])$ máme

$$\lim_{m\to\infty} d_H(\mathbf{A}_{mn}, \mathbf{A}) = 0$$

a odhad pro Hausdorffovu metriku $d_H(\mathbf{A}_{mn}, \mathbf{A})$ mezi $\mathbf{A}_{mn}$ a $\mathbf{A}$ platí.

5.1. Vizualizace pro novinový článek

Mějme slabý jazykový fraktál. Dle definice uvedené výše jsme vypočítali míru sémantičnosti D pro novinový článek jako

$$D = \frac{3}{b_1 + b_2 + b_3} \doteq 9,464827,$$

která je fraktální dimenzí uzavřené množiny $\mathbf{A}$, $\mathbf{A} = F(\mathbf{A})$, kde

$$F = F_3 \circ F_2 \circ F_1,$$

$$F_i(\mathbf{x}) = \bigcup_{\mathbf{j}} {}_i f_{\mathbf{j}}(\mathbf{x}), i = 1, 2, 3$$

a máme již spočítané ($x = 2$)

$$_1f_{\mathbf{j}}(\mathbf{x}) \doteq 0,491918\mathbf{x} + \frac{1}{2}\mathbf{j},$$

$$_2f_{\mathbf{j}}(\mathbf{x}) \doteq 0,189499\mathbf{x} + \frac{1}{2}\mathbf{j},$$

$$_3f_{\mathbf{j}}(\mathbf{x}) \doteq 0,495117\mathbf{x} + \frac{1}{2}\mathbf{j},$$

$\mathbf{j} = (j_1 \dots j_{14}), j_\mathbf{l} \in \{0, 1\}$.

Provedeme celkem šest "aproximací"(de facto se jedná pouze o dvě aproximace složeného zobrazení $F = F_3 \circ F_2 \circ F_1$), které budou dostačující (grafy 7. až 12.).
$\mathbf{A}_1 := F_1([0, 1])$,
$\mathbf{A}_2 := F_2 \circ F_1([0, 1])$,
$\mathbf{A}_3 := F_3 \circ F_2 \circ F_1([0, 1])$,
$\vdots$
$\mathbf{A}_6 := F^2([0, 1])$.

První aproximace složeného zobrazení $\mathbf{A}_3 = F_3 \circ F_2 \circ F_1([0, 1]) = F([0, 1])$, jejíž projekce je znázorněna v grafu 9., představuje vizualizaci našeho slabého jazykového fraktálu. Druhá aproximace složeného zobrazení $\mathbf{A}_6$, jejíž projekce je zobrazena v grafu 12., simuluje matematický fraktál $\mathbf{A}$.

Postupné aproximace $\mathbf{A}_{3m} = F^m([0, 1]), m = 1, 2, \dots$ splňují

$$d_H(\mathbf{A}_{3m}, \mathbf{A}) = d_H(F^m([0, 1]), \mathbf{A}) \leq \frac{(r_1r_2r_3)^m}{1 - r_1r_2r_3} d_H([0, 1], F([0, 1]))$$

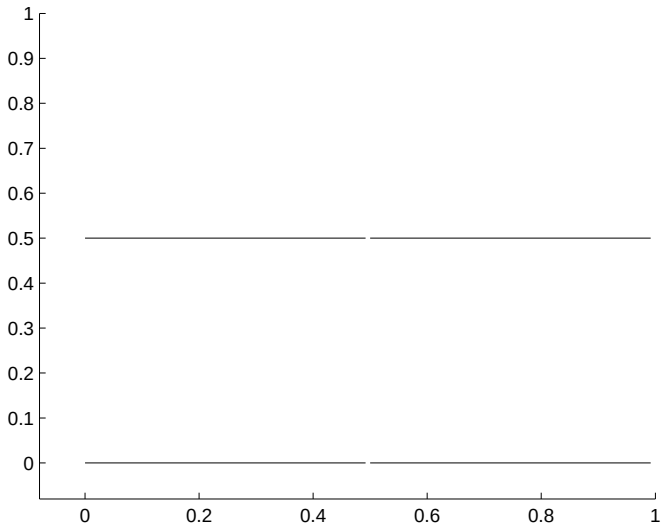
$$= \frac{((1 - \frac{1}{x}) + (1 - \frac{1}{x}) \sum_{i=2}^{n} r_2 \dots r_i)\sqrt{k - 1}}{x^{mk(b_1+\dots+b_n)}(1 - x^{-k(b_1\dots+b_n)})}$$

$$= \frac{(\frac{1}{2} + \frac{1}{2}(r_1 + r_2 + r_3))\sqrt{k - 1}}{2^{mk(b_1+b_2+b_3)}(1 - 2^{-k(b_1+b_2+b_3)})}$$

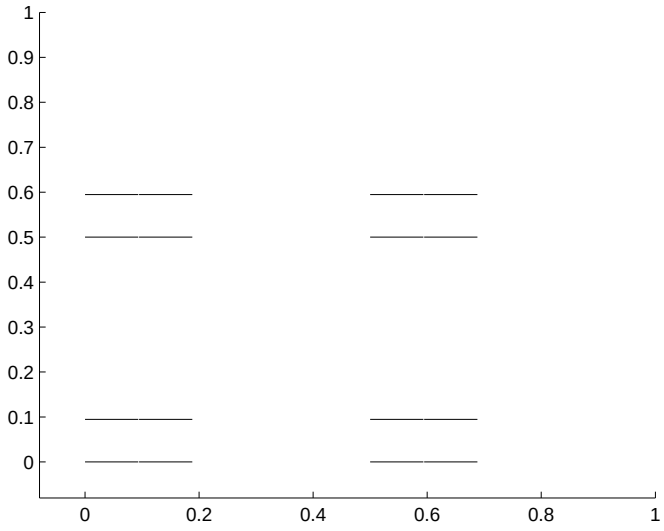
$$= \frac{(\frac{1}{2} + \frac{1}{2}(0,491918 + 0,189499 + 0,495117))\sqrt{13}}{2^{14 \cdot m(0,07311+0,17141+0,07244)}(1 - 2^{-14(0,07311+0,17141+0,07244)})} \doteq \frac{4,113659}{2^{4,43744 \cdot m}}$$

Pokud zvolíme za $m = 2$, potom dostáváme (graf 12.)

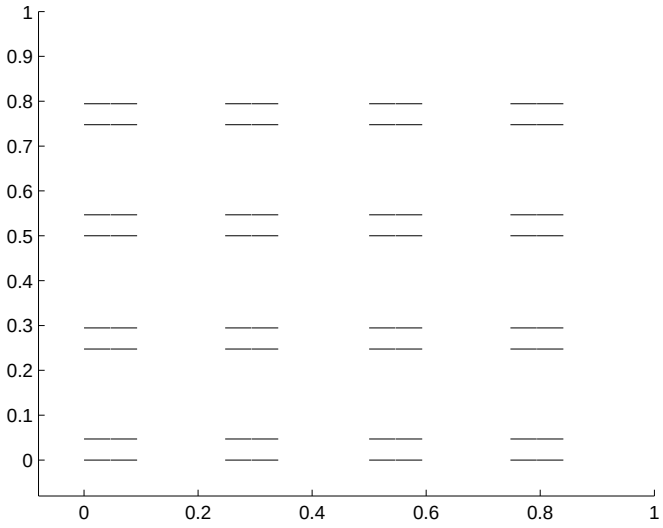
$$d_H(\mathbf{A}_6, \mathbf{A}) \leq 8,76240 \cdot 10^{-3}.$$



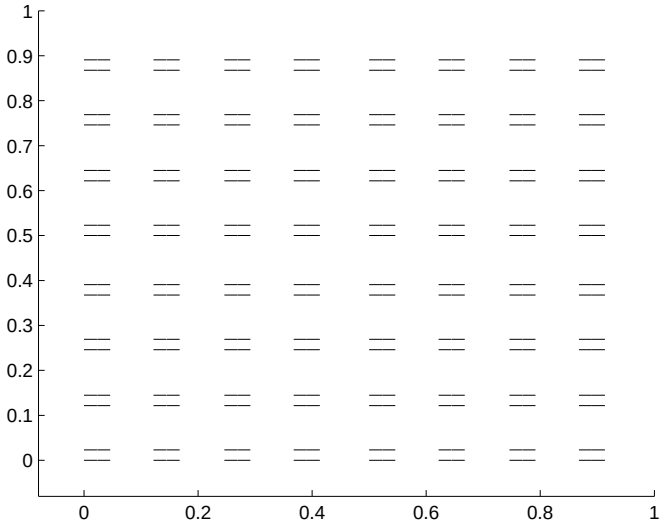
Graf 7.: První "aproximace" $A_1 = F_1([0, 1])$



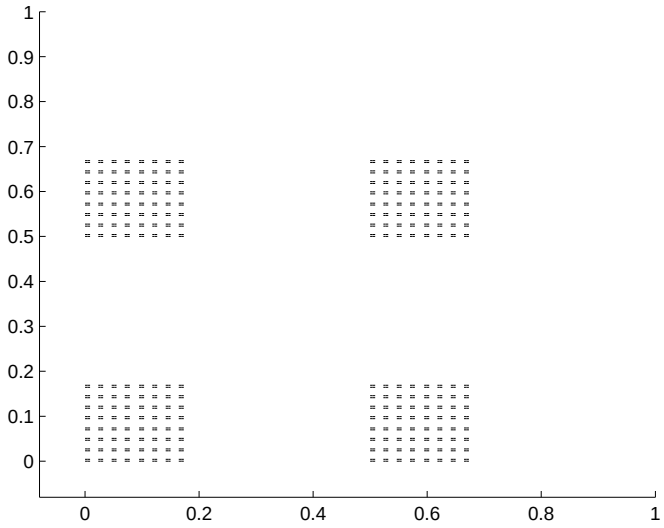
Graf 8.: Druhá "aproximace" $A_2 = F_2 \circ F_1([0, 1])$



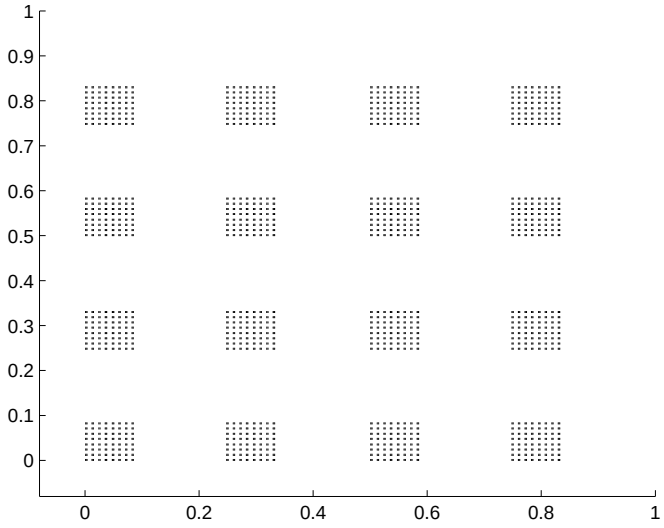
Graf 9.: Třetí "aproximace" $A_3 = F([0,1])$



Graf 10.: Čtvrtá "aproximace"



Graf 11.: Pátá "aproximace"



Graf 12.: Šestá "aproximace" $A_6 = F^2([0,1])$

Závěr

Cílem práce bylo provést dva experimenty. První experiment se týkal analýzy novinového článku a druhý analýzy SMS zpráv. Nejprve k novinovému článku.

Novinový článek názorně demonstroval, že pro nižší jazykové úrovně je potřeba použít úplnou formuli pro Menzerathův - Altmannův zákon. Při použití zkrácené formule zákona vycházel ve většině případů koeficient b nepřípustně záporně. Dále jsme zjistili, že při vytváření tabulek na jazykové úrovni slova je potřeba provést úpravu textu. Na této úrovni jsme museli započítávat předložky spolu s bezprostředně následujícím slovem jako jedno slovo. Při počítání předložky a následujících slov zvlášť nám vycházel parametr b záporně.

Při zvážení těchto dvou bodů experiment potvrdil, že náš text představuje jazykový fraktál (koeficienty b_1, b_2, b_3 vycházely kladně) a to ve slabém smyslu, protože koeficienty b_1, b_2, b_3 byly vzájemně různé.

Teorie uvedená v [1] a [2] umožňuje interpretovat převrácenou hodnotu aritmetického průměru koeficientů b_1, b_2, b_3 jako míru sémantičnosti daného textu. U novinového článku vyšla $D \doteq 9,464827$. Toto číslo lze rovněž interpretovat jako dimenzi matematického fraktálu, jehož konstrukce byla popsána v kapitole 4. Jeho první aproximace složeného zobrazení modeluje strukturu daného jazykového fraktálu. Rovinná projekce této aproximace je znázorněna v grafu 9. Vzhledem k tomu, že druhá aproximace složeného zobrazení, jejíž Hausdorffova vzdálenost činí $d_H(\mathbf{A}_6, \mathbf{A}) \leq 8,76240 \cdot 10^{-3}$ (je opticky téměř nerozlišitelná), může být považována za simulaci matematického fraktálu. Rovinná projekce matematického fraktálu, jejíž dimenze je $D^{(2)} \doteq 1,35212$, je simulována pomocí druhé aproximace v grafu 12.

Co se týkalo experimentu s SMS zprávami, bohužel koeficienty b vyšly nepřípustně záporně na všech hladinách při použití úplné verze MAZ a u jednoduché verze MAZ vyšly záporně na hladině sémantické konstrukty - klauze a to i po úpravách provedených stejně jako u novinového článku. Nejedná se tudíž o jazykový fraktál ani ve slabém smyslu. Tuto skutečnost si vysvětlujeme výraznou absencí kontextu. Dalo by se očekávat, že pokud by byly SMS zprávy doplněny

o zmíněný kontext (do běžné hovorové mluvy), zřejmě by se opět jednalo o jazykový fraktál. Touto možností jsme se již ale nezabývali z důvodu poměrně velké časové náročnosti projektu.

Podotkněme, že míra sémantičnosti, která u novinového článku vyšla přibližně $9,464827$, je víceméně srovnatelná s nemnoha experimenty tohoto typu. Míra sémantičnosti u básně "Havran" od E. A. Poea provedená mým školitelem vyšla přibližně $12,121$, tedy řádově stejně.

Mluvená podoba je jednosekvenciální. Analýza ukazuje, že nárůst dimenzionality z mluvené do myšlené podoby je způsoben právě vlivem sémantičnosti.

Příloha 1

Životní pojištění trvá obvykle desítky let. Aby vložené peníze nezahálely, pojišťovny nabízejí kombinaci pojištění a investování. V čem jsou pozitiva tohoto řešení?

PAVEL NEBESKÝ

Zájem o životní pojištění na zdejším trhu stále stoupá. Za vyspělou Evropou sice Česko stále zaostává, ale situace se pozvolna zlepšuje. Na druhou stranu jsou to právě Češi, kteří v nejvyšší míře v Evropě uzavírají životní pojistky také pro děti nebo vnuky. Rodiče či prarodiče přitom často volí investiční typ pojistky.

Zhruba čtyři pětiny všech nově uzavíraných životních pojištění ve světě připadá na investiční pojistky. Podobně je tomu v posledních letech i v Česku, kde má rozvoj životního pojištění stále velký potenciál z důvodu menší nasycenosti trhu.

Investiční životní pojištění jako takové existuje v Česku asi deset let. „Na český trh jsme před deseti lety přinesli investiční životní pojištění, které zde bylo tehdy novinkou. Dnes už mezi novými smlouvami jednoznačně dominuje," říká Milan Starý, marketingový manažer životní pojišťovny Aviva.

Investiční životní pojištění se začaly rychle objevovat i v nabídkách dalších pojišťoven. A co je zajímavé, rychle si je začaly osvojovat i pojišťovny, které se do té doby profilovaly spíše v jiných oblastech, jako například Česká podnikatelská pojišťovna.

„Z hlediska naší produkce jsme dříve byli výraznější na poli neživotního pojištění, zejména v povinném ručení. V současné době však v životním pojištění vidíme daleko větší potenciál," říká Renata Svobodová, mluvčí České podnikatelské pojišťovny.

Postup určuje klient

Investiční životní pojištění není nic jiného než kapitálová životní pojistka, kde si klient spoří. U investiční varianty má však pojištěný daleko více možností sám si zvolit řadu parametrů a ty podle potřeby dále měnit. Především si sám vybírá z řady investičních programů, které jsou navíc nemusejí omezovat jen na jeden investiční program, ale mohou jich využívat najednou. Když nějaký fond přestane být přitažlivý, peníze lze přesunout jinam. Klienti tedy mohou s penězi aktivně nakládat, nebo naopak jen je ukládat v rámci vybraného investičního programu a o nic dalšího se už nestarají. Formou výpisu z účtu mají k dispozici neustále informace o tom, zda prostředky vydělávají.

Varovný prst krize

O výhodách investičních životních pojistek lidé nepochybují, ovšem je potřeba znát i úskalí těchto produktů. Ty se naplno ukázaly v době, kdy ve světě řádila finanční krize. A nejvíce to pozorovali lidé, jež při výběru investičních strategií vsadili na dynamičtější modely s větším podílem akcií. Hodnota jejich portfolia v průběhu několika měsíců značně kolísala, a to až o desítky procent. Pokud to lidé nejsou schopni unést, měli by uvažovat buď o standardní kapitálové pojistce s garantovaným výnosem, nebo u investiční pojistky volit méně agresivní strategii.

Velká rozkolísanost hodnoty investičních pojistek se projevuje jen v kratším čase. Pokud jsou však lidé schopni akceptovat tyto zkušenosti a hlavně „udržet nervy na uzdě" – tedy nevybírat peníze pokaždé, když dojde ke krátkodobému propadu –, neměly by nastat žádné větší potíže.

V delším čase jejich portfolio poroste. A kombinace se životní pojistkou, která se obvykle sjednává i na desítky let, je pak přímo ideální.

Dlouhodobý charakter životní pojistky se dá dobře využít v případě dětských kapitálových pojistek. Proto i zde rodiče dávají často přednost investiční formě pojištění. Smlouvu totiž lze uzavřít hned po narození dítěte.

Češi myslí na děti

Velkou výhodou těchto smluv je krytí rizika, že spoření bude přerušeno nebo předčasně ukončeno z důvodů smrti nebo vážného zranění rodičů.

A právě to pojišťovna ošetřuje – v případě úmrtí či zavážného zranění rodiče většina pojišťoven přebírá placení pojistného až do konce trvání smlouvy. Dítě tak získá kapitál v původní výši navzdory okolnostem.

Právě z těchto důvodů jsou dětské kapitálové pojistky v Česku velmi populární. Podle zástupců pojišťoven se tyto pojistky neprodávají tak dobře v žádné jiné zemi Evropy.

„Pojištění Talisman nenabízíme nikde jinde. Důvod je jednoduchý, lidé v cizině o něj nemají zájem," tvrdí Milan Starý z pojišťovny Aviva.

PŘI INVESTIČNÍM ŽIVOTNÍM pojištění si zákazník určuje sám, kam vloží peníze. Foto: Shutterstock

Obr.2.: Novinový článek

Příloha 2

```
nazev=winDialogString("Zadej nazev souboru s daty: ",".txt")
words=read.table(nazev,header = TRUE)

# ------------------------------------------------------------
# funkce A*x^(-b)
# linearni model - logaritmizace: ln(Y) = ln(A) - b*ln(X)
lnY = log(words$y)
lnX = log(words$x)
plot(lnX,lnY)
vysledek = lm (lnY ~ lnX)
summary(vysledek)
koef=vysledek$coefficients
exp(koef[1]); -koef[2]
# konfidencni intervaly
koefStd=confint(vysledek)
exp(koefStd[1,]); -koefStd[2,]

f = function(x,A,b) A*x^(-b)
plot(words$x,words$y,main="Linearni model I (logaritmizace)",xlab="x",ylab="y")
lines(words$x,f(words$x,A=exp(koef[1]),b=-koef[2]))

# ------------------------------------------------------------
# funkce A*x^(-b)
# nelinearni model - nelinearni MNC
# hodnoty z logaritmizace
b0=
A0=

f = function(x,A,b) A*x^(-b)
nls(y ~ f(x,A,b), start=list(A=A0,b=b0),data=words)

plot(words$x,words$y,main="Neinearni model I (nelinearni MNC)",xlab="x",ylab="y")
lines(words$x,f(words$x,A= ,b= ))
```

Příloha 3

```
nazev=winDialogString("Zadej nazev souboru s daty: ",".txt")
words=read.table(nazev,header = TRUE)

# ------------------------------------------------------------
# funkce A*x^(-b)*e^(-c*x)
# linearni model - logaritmizace: ln(Y) = ln(A) - b*ln(X)-c*X

X = words$x
lnY =log(words$y)
lnX =log(words$x)
vysledek = lm (lnY ~ lnX + X)
summary(vysledek)
koef=vysledek$coefficients
exp(koef[1]);-koef[2];-koef[3]
# konfidencni intervaly
koefStd=confint(vysledek)
exp(koefStd[1,]);-koefStd[2,];-koefStd[3,]


f = function(x,A,b,c) A*x^(-b)*exp(-c*x)
plot(words$x,words$y,main="Linearni model II (logaritmizace)",xlab="x",ylab="y")
lines(words$x,f(words$x,A=exp(koef[1]),b=-koef[2],c=-koef[3]))

# ------------------------------------------------------------
# funkce A*x^(-b)*e^(-cx)
# nelinearni model - nelinearni MNC
# hodnoty z logaritmizace
b0=
A0=
c0=

f = function(x,A,b,c) A*x^(-b)*exp(-c*x)
nls(y ~ f(x,A,b,c), start=list(A=A0,b=b0,c=c0),data=words)

plot(words$x,words$y,,main="Nelinearni model II (nelinearni MNC)",xlab="x",ylab="y")
lines(words$x,f(words$x,A=  ,b=  ,c=  ))
```

Literatura

[1] Andres, J.: On de Saussure's principle of linearity and visualization of language structures, Glottotheory 2,2 (2010), 1-14.

[2] Hřebíček, L.: Vyprávění o lingvistických experimentech s textem, Academia, Praha, 2002.

[3] Petr, J.: Mluvnice češtiny 1: Fonetika, Fonologie, Morfonologie a morfemika, Tvoření slov, Academia, Praha, 1986.

[4] Petr, J.: Mluvnice češtiny 2: Tvarosloví, Academia, Praha, 1986.

[5] Ralston, A.: Základy numerické matematiky, Academia, Praha, 2002.

[6] Http://www.math.muni.cz/∼pokora/lsm1-1.pdf.

[7] Http://cs.wikipedia.org/wiki/Věta_(lingvistika).