Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Analýza interpunkčných znamienok v rôznojazyčných textoch

Bakalárska práca

2025
Zdenko Németh

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Analýza interpunkčných znamienok v rôznojazyčných textoch

Bakalárska práca

Študijný program:	Aplikovaná Informatika
Študijný odbor:	Informatika
Školiace pracovisko:	Katedra aplikovanej informatiky
Školiteľ:	doc. RNDr. Mária Markošová, PhD.

Bratislava, 2025
Zdenko Németh

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Zdenko Németh
Študijný program: aplikovaná informatika (Jednoodborové štúdium, bakalársky I. st., denná forma)
Študijný odbor: informatika
Typ záverečnej práce: bakalárska
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Analýza interpunkčných znamienok v rôznojazyčných textoch
Analysis of punctuation marks in different languages texts

Anotácia: Študent naprogramuje aplikáciu, ktorá v textovom súbore spracuje distribúciu interpunkčných znamienok. V aplikácii je potrebná možnosť výberu, teda, či pôjde o distribúciu všetkých interpunkčných znamienok, alebo len tých, ktoré budú dopredu špecifikované (napr. len čiarky, len bodkočiarky ....). Aplikácia tiež musí umožniť rôzne typy zobrazení výsledných súborov, ako napr. log log zobrazenie, log lin zobrazenie a podobne. Študent porovná a vyhodnotí distribúcie, ktoré získa pre jednotlivé jazyky.

Cieľ: Cieľom je vytvoriť aplikáciu na analýzu distribúcií interpunkčných znamienok v texte.

Literatúra: Kulig and others. In narrative texts punctuation marks obey the same statistics as words, Information Sciences
Volume 375, 1 January 2017, Pages 98-113

Vedúci: doc. RNDr. Mária Markošová, PhD.
Katedra: FMFI.KAI - Katedra aplikovanej informatiky
Vedúci katedry: doc. RNDr. Tatiana Jajcayová, PhD.

Dátum zadania: 22.02.2024

Dátum schválenia: 23.10.2024 doc. RNDr. Damas Gruska, PhD.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie: You can thank anyone who helped you with the thesis here (e.g. your supervisor).

Abstrakt

Slovenský abstrakt v rozsahu 100–500 slov, jeden odstavec. Abstrakt stručne sumarizuje výsledky práce. Mal by byť pochopiteľný pre bežného informatika. Nemal by teda využívať skratky, termíny alebo označenie zavedené v práci, okrem tých, ktoré sú všeobecne známe.

Kľúčové slová: Slovak, keywords, here

Abstract

Abstract in the English language (translation of the abstract in the Slovak language).

Keywords: English, keywords, here

Obsah

Zoznam obrázkov

Algoritmy

Úvod

Put your introduction here.

Kapitola 1

Prehľad problematiky

Táto kapitola sa zaoberá základnými pojmami a definíciami, ktoré sú nevyhnutné pre pochopenie problematiky tejto práce.

Najskôr sa zameriame na teóriu grafov, ktorá je základom pre analýzu a modelovanie sietí. Definujeme si kľúčové pojmy, ako sú graf, vrchol, hrana, stupeň a ďalšie.

Následne sa pozrieme na rôzne typy modelov sietí a ich vlastnosti, predovšetkým na náhodné a bezškálové siete. Budeme sa sústrediť na ich vznik a charakteristiky, ako je distribúcia stupňov vrcholov. Taktiež sa budeme venovať slovným sieťam, ktoré sú špecifickým typom sietí, využívaným na analýzu jazyka a jeho štruktúry.

Na záver kapitoly sa zameriame na úlohu a význam interpunkcie v jazyku a jej vplyvu na štruktúru textu.

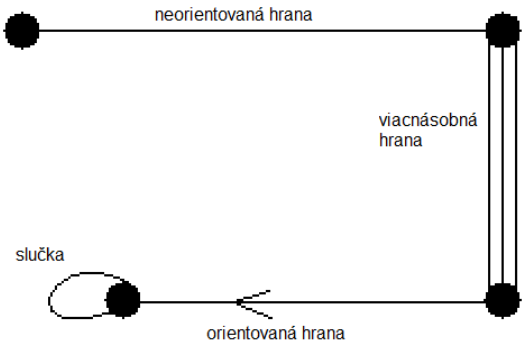
1.1 Teória grafov

Teória grafov je rozsiahla a komplexná oblasť matematiky a informatiky. V tejto kapitole sa zameriame na základné pojmy, definície a koncepty, ktoré sú nevyhnutné na pochopenie práce s grafmi, ich vlastností a rôznych aplikácií. Pre podrobnejšie informácie a hlbšie pochopenie odporúčam odbornú literatúru, najmä publikáciu *Graph Theory Fifth Edition od R. Diestela* [1] .

Graf G je reprezentovaný ako dvojica $G = (V, E)$, pričom V *(z angl. vertices)* je množina všetkých vrcholov a E je množina všetych hrán *(z angl. edges)*, ktoré tvoria spojenia medzi týmito vrcholmi. Vrcholy grafu sa taktiež nazývajú aj uzly.

Medzi dvojicou vrcholov grafu môže, ale nemusí existovať hrana. Ak medzi nimi hrana existuje, hovoríme, že sú navzájom prepojené. Jednotlivé hrany, ktoré obsahuje množina E, sú reprezentované ako dvojice vrcholov (u, v), pričom u a v sú vrcholy z množiny V.

Rozlišujeme niekoľko typov hrán. Ak má hrana grafu daný smer, nazýva sa orientovaná a záleží na poradí vrcholov v usporiadanej dvojici, prvý vrchol dvojice predstavuje



Obr. 1.1: Zobrazenie typov hrán v grafe.

uzol, od ktorého hrana vychádza a druhý vrchol dvojice predstavuje uzol, do ktorého hrana smeruje. Ak hrana nemá určený smer, tak sa jedná o neorientovanú hranu a poradie vrcholov v usporiadanej dvojici nie je dôležité. Hranu, ktorá má začiatok a koniec v rovnakom vrchole, nazývame slučka. Pojem viacnásobná hrana označuje prípad, kedy medzi dvoma vrcholmi existuje viac ako jedna hrana. Rôzne typy hrán je možné vidieť na ilustračnom obrázku č. 1.1 .

Hrany, ktoré majú priradenú číselnú hodnotu, nazývame vážené hrany. Tieto váhy môžu reprezentovať rôzne vlastnosti hrany, ako napríklad vzdialenosť medzi vrcholmi alebo náklady na prechod medzi nimi. Sú využívané hlavne v praktických aplikáciách, ako napríklad dopravné siete.

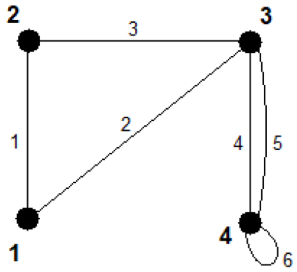
Pojem jednoduchý graf definuje taký graf, ktorý neobsahuje slučky ani viacnásobné hrany. Jednoduchý graf, ktorý neobsahuje orientované hrany a vrcholy sú poprepájané spôsobom každý s každým nazývame kompletný graf [2] .

Pri analýze grafov je potrebné poznať rôzne spôsoby, akými môžeme prechádzať cez vrcholy a hrany grafu. Prechod grafom, pri ktorom sa striedajú vrcholy a hrany, ktoré sa môžu opakovať, sa nazýva sled. Formálny zápis pre sled v grafe $G = (V, E)$ je postupnosť

$$v_0, e_0, v_1, e_1, \ldots, e_{k-1}, v_k,$$

kde $v_i \in V(G)$ pre všetky $i \in \{0, 1, \ldots, k\}$ a $e_j \in E(G)$ pre všetky $j \in \{0, 1, \ldots, k-1\}$ s podmienkou $e_i = (v_i, v_{i+1})$. Ťah je špeciálny typ sledu, pri ktorom sa hrany nemôžu opakovať, teda pre všetky $i \neq j$ platí $e_i \neq e_j$. Vrcholy sa v ťahu môžu opakovať. Sled a ťah majú špeciálny prípad, kedy sa počiatočný a koncový vrchol zhodujú, teda $v_0 = v_k$. Takýto sled sa nazýva uzavretý sled a ťah uzavretý ťah. Ešte existuje pojem cesta, ktorý predstavuje prechod grafom, pri ktorom sa nemôžu opakovať ani vrcholy ani hrany.

Spojitý graf je taký graf, v ktorom existuje cesta medzi každými dvoma vrcholmi. Nie každý graf je spojitý, pretože niektoré grafy sa skladajú z viacerých disjunktných častí, ktoré nie sú navzájom prepojené hranou, teda medzi nimi neexistuje žiadna cesta. Takýmto disjunktným častiam grafu hovoríme komponenty. Graf s viacerými



$$A = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 2 \\ 0 & 0 & 2 & 1 \end{pmatrix} \quad B = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 2 \end{pmatrix}$$

Obr. 1.2: Matica susedností a incidenčná matica pre zobrazený graf.

komponentami je nespojitý graf.

Grafy sa dajú reprezentovať rôznymi spôsobmi. Najčastejšia matematická reprezentácia je pomocou matice susedností a matice incidencie. Nech $G = (V, E)$ je graf s $n = |V|$ vrcholmi a $m = |E|$ hranami. Matica susedností A je $n \times n$ matica, kde každý prvok a_{ij} predstavuje počet hrán medzi vrcholmi v_i a u_j. Incidenčná matica B je $n \times m$ matica, kde pre každý prvok platí $b_{ij} = 1$, ak je hrana e_j incidentná s vrcholom v_i, inak $b_{ij} = 0$. Príklad matice susedností a incidenčnej matice spolu s ilustračným grafom je zobrazený na obrázku č. 1.2 .

Teraz si predstavíme niekoľko základných vlastností, s ktorými budeme pracovať, ich definície z teórií grafov a ich aplikáciu. Budeme sa zaoberať iba jednoduchými, neorientovanými grafmi, ak nebude uvedené inak.

Jedna zo základných vlastností vrchola grafu je jeho stupeň. Stupeň vrchola v označovaný aj ako $deg(v)$ nám udáva počet hrán, ktoré sú incidentné s vrcholom v. Minimálna hodnota stupňa vrchola je 0, vtedy tvorí vrchol samostatný komponent, taktiež sa nazýva izolovaný vrchol. Maximálna hodnota stupňa vrchola je $n-1$, nastáva keď vrchol susedí s každým vrcholom grafu. Celkový počet hrán v grafe potom získame ako polovicu súčtu všetkých stupňov vrcholov grafu, pretože každá hrana je incidentná s dvoma vrcholmi. Tento vzťah môžeme zapísať ako:

$$|E| = \frac{1}{2} \sum_{i=1}^{n} deg(v_i), \tag{1.1}$$

kde $|E|$ predstavuje počet hrán v grafe a n je počet vrcholov [3] .

Taktiež často používanou charakteristikou grafu je priemerný stupeň uzla:

$$\bar{d} = \frac{1}{n} \sum_{i=1}^{n} deg(v_i) = \frac{2|E|}{n}, \tag{1.2}$$

kde $\bar{d}$ je priemerný stupeň uzla a n je počet vrcholov grafu [3] .

Hustota grafu určuje, ako blízko je graf ku kompletnému, teda akú časť z maximálneho počtu hrán obsahuje. Nadobúda hodnoty v intervale $[0, 1]$, pričom 0 znamená, že

graf je prázdny a 1 znamená, že graf je kompletný. Formálny zápis pre hustotu grafu je:

$$D = \frac{2|E|}{n(n-1)}, \tag{1.3}$$

kde D je hustota grafu, $|E|$ je počet hrán a n je počet vrcholov.

Ďalšou vlastnosťou grafu je koeficient zhlukovania, inak známy aj ako klasterizačný koeficient. Tento koeficient udáva ako veľmi sú prepojené susedné vrcholy uzla. Hodnota koeficientu sa pohybuje v intervale $[0, 1]$, kde 0 značí, že medzi susedmi vrcholu neexistujú žiadne hrany a 1 znamená, že všetci susedia sú medzi sebou navzájom prepojení. Nech v je vrchol grafu a $deg(v)$ je jeho stupeň, potom:

$$C_v = \frac{2|E_v|}{deg(v)(deg(v)-1)}, \tag{1.4}$$

kde C_v je koeficient zhlukovania vrchola v, $|E_v|$ predstavuje počet hrán medzi $deg(v)$ susedmi vrchola v. Týmto spôsobom vieme získať lokálny koeficient zhlukovania pre každý vrchol grafu, z ktorých potom vieme vypočítať priemerný koeficient zhlukovania pre celý graf:

$$C = \frac{1}{n}\sum_{i=1}^{n} C_{v_i}, \tag{1.5}$$

kde C je priemerný koeficient zhlukovania grafu a n je počet vrcholov grafu [3] .

Pojem najkratšia cesta v grafe definuje takú cestu, ktorá spája dva vrcholy grafu a obsahuje najmenší možný počet hrán. Priemerná najkratšia cesta medzi dvoma vrcholmi je definovaná ako priemer všetkých najkratších ciest medzi všetkými kombináciami dvojíc vrcholov grafu.

$$\ell = \frac{1}{n(n-1)}\sum_{\substack{i,j=1 \\ i\neq j}}^{n} d(v_i, v_j), \tag{1.6}$$

kde ℓ je priemerná najkratšia cesta, $d(v_i, v_j)$ je dĺžka najkratšej cesty medzi vrcholmi v_i a v_j a n je počet vrcholov grafu [3] .

Priemer grafu udáva najväčšiu dĺžku najkratšej cesty medzi dvoma rôznymi vrcholmi grafu.

$$diam = \max_{i\neq j} d(v_i, v_j), \tag{1.7}$$

Veľmi dôležitým parametrom pri analýze grafu je distribúcia stupňov vrcholov, značené ako p_{deg}. Poskytuje pravdepodobnosť, že náhodne vybraný vrchol grafu má práve deg susedov [3] . Pre graf s počtom vrcholov n a počtom vrcholov so stupňom deg označeným ako n_{deg} je distribúcia stupňov definovaná ako:

$$p_{\text{deg}} \cong \frac{n_{\text{deg}}}{n}, \tag{1.8}$$

kde p_{deg} predstavuje pravdepodobnosť výskytu daného stupňa, n_{deg} je počet vrcholov s daným stupňom a n je počet vrcholov grafu.

1.2 Modely sietí

Siete sú štruktúry, ktoré možno reprezentovať ako grafy skladajúce sa z uzlov (vrcholov) a prepojení (hrán), pričom uzly predstavujú jednotlivé entity a prepojenia reprezentujú vzťahy medzi nimi. Pojem graf sa používa matematickej reprezentácií, zatiaľ čo termín sieť často odkazuje na reálne systémy. Umožňujú modelovanie a analýzu komplexných systémov v rôznych odvetviach, ako informatika, biológia, sociológia a iné [3] . Medzi základné topológie sietí patria náhodné siete a bezškálové siete, ktoré predstavujú odlišné prístupy ku vzniku a distribúcii prepojení.

1.2.1 Náhodná sieť

Náhodná sieť predstavuje jeden zo základných konceptov v teórií sietí. Definujú sa dva základné modely náhodných sietí.

Prvý z nich je Erdős-Rényiho model, $G_{V,H}$, kde V je množina vrcholov a H je množina náhodne vybraných hrán [4][3] . Tento model generuje náhodnú sieť tak, že začne s $n = |V|$ izolovanými vrcholmi a následne náhodne pridáva $m = |H|$ rôznych hrán medzi nimi, ktoré sa vyberajú z $\frac{n*(n-1)}{2}$ všetkých možných hrán.

Druhý model je Gilbertov model, $G_{V,p}$, takisto obsahuje $n = |V|$ vrcholov, však hrany nie sú pridelené pevným počtom, ale nezávisle s pravdepodobnosťou p [5][3] . Model začína s n izolovanými vrcholmi a následne pre každú dvojicu vrcholov u a v pridá hranu medzi nimi s pravdepodobnosťou p, pričom výsledný počet hrán je náhodný a závisí od hodnoty p.

Priemerný stupeň vrchola v náhodnej sieti pre model $G_{V,H}$ vieme vypočítať jednoduchým vzorcom:

$$\bar{d} = \frac{2|H|}{n}, \tag{1.9}$$

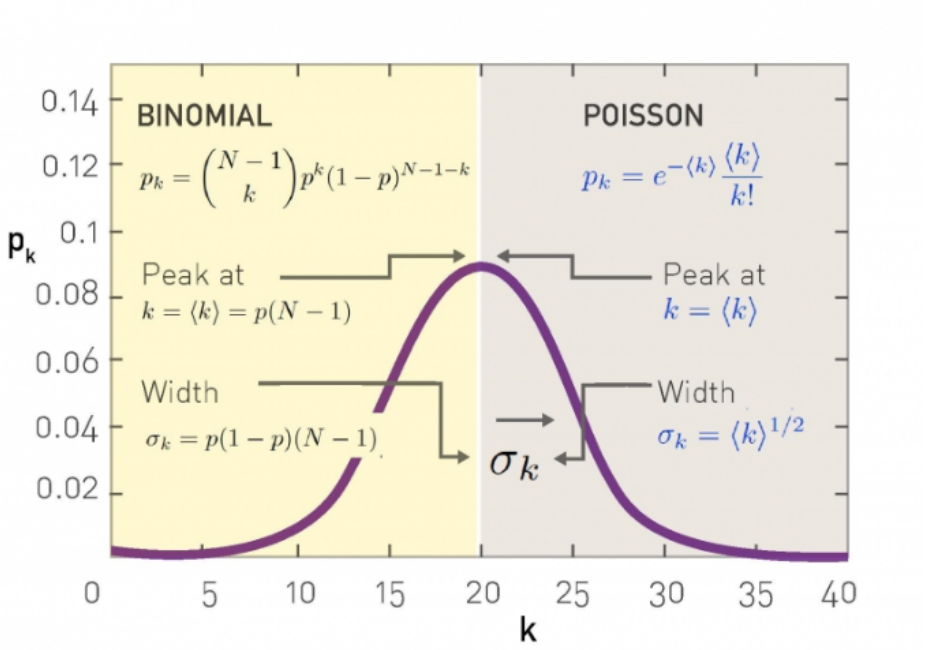
keďže poznáme počet všetkých hrán $|H|$, aj vrcholov n [3] . Pre výpočet priemerného stupňa vrchola v Gilbertovom modeli $G_{V,p}$ môžeme použiť vzorec [3] :

$$\bar{d} = \frac{2 < L >}{n} = p(n-1), \tag{1.10}$$

kde $< L >$ definuje súčin pravdepodobnosti p a počtu párov, ktoré sa snažíme spojiť [3] :

$$< L >= p\frac{n(n-1)}{2}. \tag{1.11}$$

Na tvorbu náhodných sietí nemáme veľký vplyv, preto nastáva, že niektoré vrcholy majú veľmi vysoký stupeň, zatiaľ čo iné majú veľmi nízky stupeň, dokonca aj nulový. Tieto rozdiely je možné pozorovať na distribúcii stupňov vrcholov p_k, ako je vidieť na obrázku č. 1.3 [3] .



Obr. 1.3: Binomická vs. Poissonová distribúcia stupňov. Presná forma distribúcie náhodnej siete je binomická distribúcia (ľavá strana), pri veľkom počte vrcholov sa dá priblížiť Poissonovou distribúciou (pravá strana) [3] .

Distribúcia stupňov vrcholov v náhodnej sieti sa riadi binomickou distribúciou, ktorá je definovaná ako [3] :

$$p_{\text{deg}} = \binom{n-1}{\text{deg}} p^{\text{deg}} (1-p)^{n-1-\text{deg}}, \tag{1.12}$$

pričom tvar binomickej distribúcie je daný počtom vrcholov n, počtom hrán m a pravdepodobnosťou p. Distribúcia stupňov vrcholov v náhodnej sieti sa dá priblížiť Poissonovou distribúciou, ak má veľký počet vrcholov, ktorá je definovaná ako [3] :

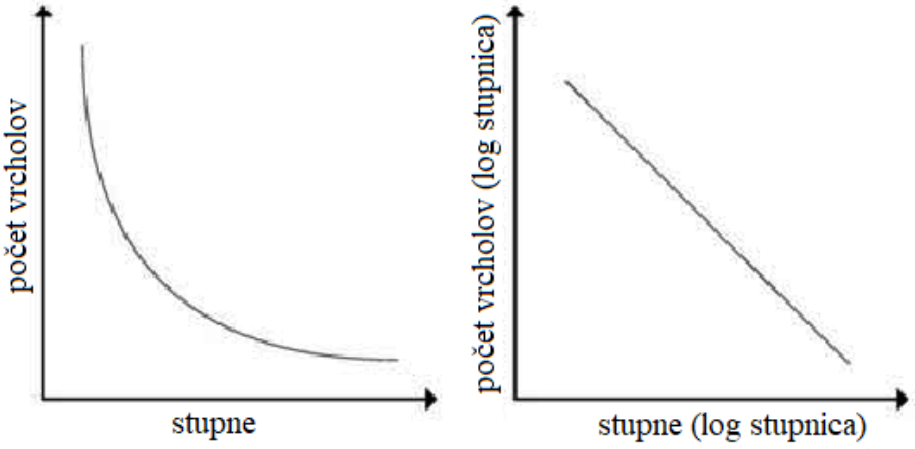
$$p_{\text{deg}} = \frac{\bar{d}^{\text{deg}} e^{-\bar{d}}}{\text{deg}!}, \tag{1.13}$$

$\bar{d}$ je priemerný stupeň vrchola, ktorý je daný vzorcom 1.10 .

Koeficient zhlukovania v náhodnej sieti je veľmi nízky, pretože väčšina vrcholov je izolovaných a nie sú prepojené. Vieme ho jednoducho vypočítať ako:

$$C = p = \frac{\bar{d}}{n}, \tag{1.14}$$

predstavuje pravdepodobnosť, že dva náhodne vybrané vrcholy sú prepojené [3] .



Obr. 1.4: Distribúcia stupňov vrcholov v bezškálovej sieti [7] .

1.2.2 Bezškálová sieť

Pri bezškálových sieťach sa jedná o špecifický typ grafu, ktorý sa vyznačuje tým, že existuje niekoľko uzlov s extrémne vysokým počtom prepojení a väčšina má malý počet prepojení. Tento typ siete sa v reálnom živote vyskytuje oveľa častejšie ako náhodné siete, najmä v kontextoch ako internet, biologické siete alebo sociálne siete [6] [3] .

Na rozdiel od náhodných sietí, kde sú stupne vrcholov rozdelené podľa binomickej alebo Poissonovej distribúcie, ako je vidno na obrázku č. 1.3 , bezškálové siete dodržujú tzv. mocninové rozdelenie (power-law distribution):

$$p_{\deg} \sim \deg^{-\gamma}, \tag{1.15}$$

kde γ je škálovací exponent v rozmedzí $2 < \gamma < 3$, ktorý určuje tvar [3] . Pri tomto rozdelení nás najviac zaujíma rozdelenie stupňov vrcholov, zobrazené v dvojitej logaritimickej mierke, kde produkuje priamku, ako je vidieť na obrázku č. 1.4 .

Najznámejší model na generovanie bezškálových sietí je Barabási–Albertov model, označovaný ako $G_{BA}(n, m)$, ktorý má škálovací exponent $\gamma = 3$. Pri generovaní využíva princíp preferenčného pripájania. Začína s malým počtom počiatočných uzlov a v každom kroku sa pridá nový uzol, ktorý sa pripojí k m už existujúcim uzlom. Pravdepodobnosť prepojenia nového uzla so starým závisí od stupňa starého uzla [3] :

$$p(\deg(\mathrm{i})) = \frac{\deg(\mathrm{i})}{\sum_j \deg(\mathrm{j})}, \tag{1.16}$$

Ďalším modelom na generovanie bezškálových sietí je Dorogovtsev–Mendes model, často označovaný ako $G_{DM}(n)$. Tento model taktiež generuje siete s mocninovým rozdelením stupňov. Od Barabási-Albertovho modelu sa odlišuje mechanizmom rastu siete. Model začína s trojuholníkom (kompletným grafom s tromi uzlami) a v každom kroku sa pridá nový uzol, ktorý sa pripojí na oba koncové body náhodne vybranej

existujúcej hrany [8]. Takýmto spôsobom sa v každom kroku vytvorí nový trojuholník, čo vedie k vysokému koeficientu zhlukovania. Model teda zabezpečuje vznik sietí, ktoré sú nielen bezškálové, ale majú aj malé priemerné vzdialenosti medzi uzlami a vysokú mieru zhlukovania.

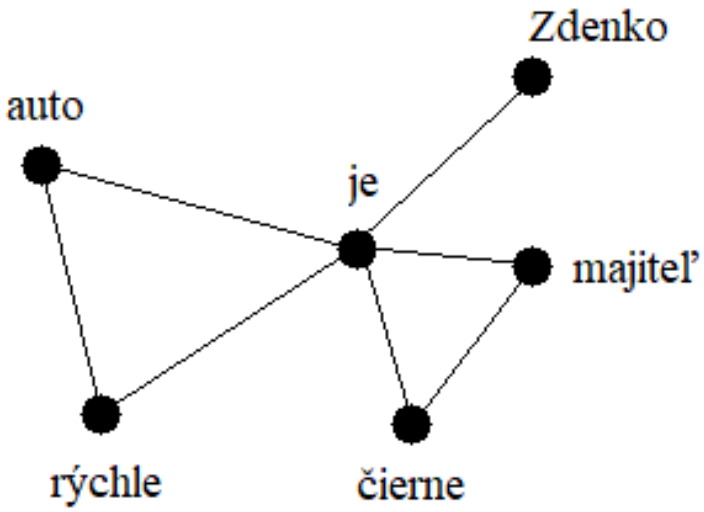
Zvýšený koeficient zhlukovania sa vyskytuje najmä pri analýze slovných sietí, kde sa rovnaké slová často vyskytujú v blízkosti seba a vytvárajú frázy, prirodzene sa zhlukujú do skupín. Práve preto je Dorogovtsev-Mendes model vhodný na analýzu jazykových dát, keďže dokáže simulovať podobné vlastnosti, štruktúru ako majú slovné siete vytvorené na reálnych jazykových dátach.

1.3 Slovné siete

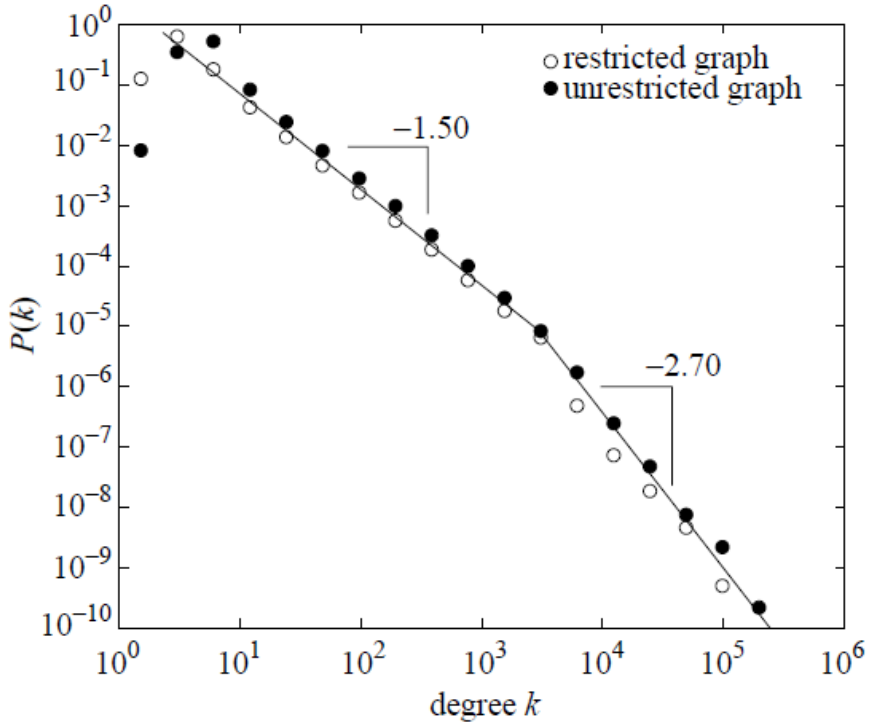
Slovné siete predstavujú efektívny spôsob reprezentácie a analýzy jazyka a jeho štruktúry za pomoci teórie grafov. V takýchto sieťach sú jednotlivé slová reprezentované ako uzly a vzťahy medzi slovami, ktoré sa navzájom ovplyvňujú alebo súvisia tvoria hrany [9] . Podľa metód reprezentácie vzťahov medzi slovami dokážeme slovné siete rozlišovať. V sémantických slovných sieťach vznikajú prepojenia medzi slovami na základe ich významovej podobnosti. Narozdiel od sémantických slovných sietí, v syntaktických slovných sieťach sú interakcie medzi slovami založené na slovnej štruktúre a gramatických pravidlách.

Špecifický typ syntaktickej slovnej siete, ktorej sa budeme venovať v tejto práci sa nazýva pozičná slovná sieť. Táto sieť je zameraná na štruktúru a pozície slov v texte. V pozičnej slovnej sieti uzly tvoria jednotlivé slová a hrany sú vytvorené na základe toho, či sa slová nachádzajú vedľa seba v texte. Ukážku takejto siete vidíme na obrázku č. 1.5 .

Tieto siete majú niekoľko zaujímavých vlastností. Koeficient zhlukovania, ktorý určuje, do akej miery slová tvoria skupiny alebo často používané frázy. Distribúcia stupňov uzlov je typu mocninného zákona, čo znamená, že väčšina slov má veľmi nízky stupeň, zatiaľ čo malý podiel slov má veľmi vysoký stupeň [10] . Toto rozdelenie je typické pre bezškálové siete. Na obrázku č. 1.6 je zobrazená distribúcia stupňov uzlov s dvomi rôznymi priemernými exponentami $y = -1.5$ a $y = -2.7$. Exponent $y = -2.7$ sa blíži k exponentu pre Barabási-Albertov model, ktorý je $y_{BA} = -3$ [11] .



Obr. 1.5: Pozičná slovná sieť z viet: „Auto je rýchle. Auto je čierne. Majiteľ je Zdenko.“.



Obr. 1.6: Dorogovtsev Mendes model, distribúcia stupňov vrcholov pre dve siete, body zoskupené mocninou dvoch [11] .

1.4 Interpunkcia

Interpunkcia je základný prvok jazyka, používa sa na štruktúrovanie a organizovanie textu, vyjadrenie významových vzťahov. Pri správnom použití interpunkcia zlepšuje čitateľnosť a porozumenie textu, zatiaľ čo nesprávne použitie môže viesť k nejednoznačnosti a nedorozumeniu [12].

Tvorí ju súbor znakov, kde každý znak má svoj osobitý význam a funkciu, ako napríklad pauza, dôraz, ukončenie vety, vzťahy medzi vetnými členmi a podobne. Pravidlá písania interpunkcie sú kodifikované v gramatických a jazykových príručkách. Taktiež sa pravidlá písania interpunkcie líšia v závislosti od jazyka, v ktorom je text napísaný.

Najbežnejšie používané interpunkčné znamienka možno rozdeliť do niekoľkých skupín podľa ich funkcie: ukončovacie (bodka, otáznik, výkričník), oddeľovacie (čiarka, bodkočiarka, dvojbodka), typografické (úvodzovky, zátvorky, pomlčky), prípadne aj iné znaky (trojbodka, lomka, atď.).

Z gramatického hľadiska interpunkcia rozdeľuje text, vyznačuje vetnú štruktúru, oddeľuje vetné členy, hlavné a vedľajšie vety. Napríklad umiestnenie čiarky pred spojkami často naznačuje, že ide o vedľajšiu vetu, zatiaľ čo bodka, otáznik, výkričník označujú koniec vetného celku.

Z fonetického hľadiska interpunkcia určuje intonáciu, pauzy pri čítaní textu, dôraz, rytmus a tón. Napríklad bodka naznačuje koniec vety a vyžaduje prestávku, kým čiarka naznačuje, že čitateľ by mal pokračovať v čítaní s malou pauzou. Takýmto spôsobom interpunkcia dokáže preniesť aspekty reči do písaného textu, čo môže ovplyvniť jeho význam a interpretáciu.

Zo štylistického hľadiska interpunkcia ovplyvňuje štýl a tón textu. Napríklad použitie výkričníka môže naznačovať dôraz, vážnosť alebo nadšenie, zatiaľ čo otáznik môže naznačovať pochybnosť. Taktiež môže ovplyvniť rytmus a plynulosť textu. Napríklad dlhé vety s množstvom čiarkami môžu spôsobiť, že text bude pôsobiť ťažkopádne a zložito, zatiaľ čo krátke vety s minimálnym použitím interpunkcie môžu pôsobiť rýchlo a dynamicky.

Kapitola 2

Použité technológie

2.1 NetworkX

NetworkX je open-source knižnica pre Python, ktorá sa zaoberá tvorbou, manipuláciou a analýzou komplexných sietí a grafov. Táto knižnica podporuje všetky druhy grafov, vrátane orientovaných a neorientovaných grafov, vážených a nevážených grafov, ako aj multigrafov [13].

V tejto práci je knižnica NetworkX využitá na základnú grafovú analýzu. Na analýzu sietí sú použité rôzne zabudované funkcie, ktoré umožňujú efektívne vypočítať rôzne metriky, ako napríklad stupeň uzlov, koeficient zhlukovania, priemernú dĺžku najkratšej cesty, hustotu a priemer siete.

Okrem toho je knižnica využitá aj na generovanie rôznych typov bezškálových sietí, ako sú Barabási-Albertov model a Dorogovtsev-Mendesov model. Pre generovanie daných modelov boli použité zabudované funkcie, $nx.barabasi\_albert\_graph()$ a $nx.dorogovtsev\_goltsev\_mendes\_graph()$, ktoré umožňujú rýchle, efektívne vytvorenie týchto sietí.

Knižnica NetworkX bola vybraná pre jednoduchosť implementácie, rozsiahlu dokumentáciu a podporu integrácie s inými knižnicami v Pythone, ako sú NumPy a Matplotlib.

Na ukážke 2.1 je zobrazený kód, ktorý generuje Dorogovtsev-Mendesov model za pomoci funkcie $nx.dorogovtsev\_goltsev\_mendes\_graph()$ a následne využije funkcie $G.number\_of\_nodes()$ pre výpočet počtu vrcholov, $G.number\_of\_edges()$ pre výpočet počtu hrán, $nx.average\_clustering(G)$ pre výpočet koeficientu zhlukovania a $nx.average\_shortest\_path\_length(G)$ pre výpočet priemernej dĺžky najkratšej cesty v grafe.

```
import networkx as nx

G = nx.dorogovtsev_goltsev_mendes_graph(5)

numNodes = G.number_of_nodes()
numEdges = G.number_of_edges()
avgClustering = nx.average_clustering(G)
avgShortestPath = nx.average_shortest_path_length(G)

print(f"Number of nodes: {numNodes}")
print(f"Number of edges: {numEdges}")
print(f"Average clustering coefficient: {avgClustering}")
print(f"Average shortest path length: {avgShortestPath}")
```

Algoritmus 2.1: Ukážka použitia NetworkX pre generovanie DGM modelu a grafovú analýzu.

Kapitola 3

Tvorba aplikácie

uvod do kapitoly

3.1 Prístup k problematike

pristup k problematike

- app v pythone
- pozicna slovna siet iba slova alebo aj s interpunkciou, lowercase
- zobrazovanie charakteristik pomocou matplotlib
- vypocty pomocou numpy

3.2 Používateľské rozhranie

obsah gui

- implementacia gui
- obrazok gui

3.3 Tvorba slovnej siete

pomocou regex, vyseknutie slov z textu alebo aj s interpunkciou, lowercase

- tvorba dictionary
- implementacia

3.4 Zobrazovanie charakteristík

growth rate, degree distribution raw to log-log with log binning

- implementacia

3.5 Analýza siete

grafova analyza pomocou networkx, jazykova analyza custom implementacia

3.6 Ukladanie siete

ukazat implementaciu

moznost ulozenia siete do suboru pre dalsie pouzitie

3.7 Testovanie

tvorba pozicnej slovnej siete ma unittesty

zobrazovanie charakteristik rucne testovane

3.8 Možnosti rozšírenia

nacitanie siete zo suboru

manualne nastavenie parametrov pri charakteristikach, stupnice aky velky vysek..

longest decreasing slice

log-log binning viac interaktivne pre uzivatela

vypocet analyzy siete interaktivne pre uzivatela ktore veliciny sa maju vypocitat

Kapitola 4

Analýza textov

4.1 Vývoj slovnej siete

4.2 Distribúcia stupňov vrcholov

4.3 Grafová analýza

4.4 Jazyková analýza

Literatúra

[1] Reinhard Diestel. *Graph Theory*. Springer, Berlin, Germany, 5th edition, 2017.

[2] Mária Markošová. Dynamika sietí. *Umelá inteligencia a kognitívna veda II*, pages 321–379, 2010.

[3] Albert-László Barabási. *Network Science*. Cambridge University Press, Cambridge, UK, 2016.

[4] A. Rényi P. Erdős. On random graphs. *I. Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959.

[5] Edgar N Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

[6] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[7] Akhlaq Ahmad, Mohamed Ridza Wahiddin, Abdur Rahman, Imad Afyouni, Bilal Sadiq, Faizan Rehman, and Sohaib Ghani. Scale free network analysis of a large crowd through their spatio-temporal activities. 12 2015.

[8] Sergey N Dorogovtsev and José FF Mendes. Evolution of networks. *Advances in Physics*, 51(4):1079–1187, 2002.

[9] Adilson E Motter, Alessandro PS De Moura, Ying-Cheng Lai, and Partha Dasgupta. Topology of the conceptual network of language. *Physical Review E*, 65(6):065102, 2002.

[10] Sergey N Dorogovtsev and José Fernando F Mendes. Language as an evolving word web. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1485):2603–2606, 2001.

[11] Ramon Ferrer I Cancho and Richard V Solé. The small world of human language. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1482):2261–2265, 2001.

[12] Yani Lubis, Fardhan Syahri, Nur Annisa Ameliya, Zaskia Amanda Putri, and Ari Rajasa Tampubolon. Mastering of punctuation marks. *Jurnal Penelitian Ilmiah Multidisiplin*, 9(1), 2025.

[13] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008.