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

Úvod

Put your introduction here.

Kapitola 1

Prehľad problematiky

...

1.1 Teória grafov

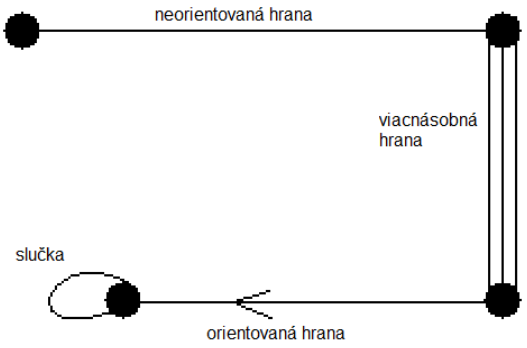
Teória grafov je rozsiahla a komplexná oblasť matematiky a informatiky. V tejto kapitole sa zameriame na základné pojmy, definície a koncepty, ktoré sú nevyhnutné na pochopenie práce s grafmi, ich vlastností a rôznych aplikácií. Pre podrobnejšie informácie a hlbšie pochopenie odporúčam odbornú literatúru, 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.

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



Obr. 1.1: Zobrazenie typov hrán v grafe.

ako napríklad dopravné siete.

Pojem jednoduchý graf nám 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 treba 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

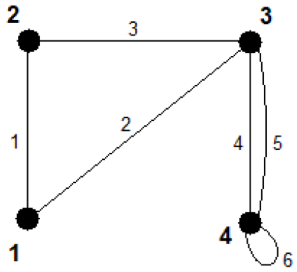
$$v_0, e_0, v_1, e_1, \ldots, e_{k-1}, v_k,$$

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 viacero 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,



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

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

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:

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

kde L predstavuje počet hrán v grafe a n je počet vrcholov[3].

Taktiež často používanou charakteristikou grafu je priemerný stupeň uzla:

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

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

Hustota grafu nám 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:

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

kde D je hustota grafu, L 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:

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

kde C_v je koeficient zhlukovania vrchola v, L_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 nám 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 nám 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_k. Poskytuje pravdepodobnosť, že náhodne vybraný vrchol grafu má práve k susedov[3]. Pre graf s počtom vrcholov n a počtom vrcholov s k susedmi n_k je distribúcia stupňov definovaná ako:

$$p_k = \frac{n_k}{n}, \tag{1.8}$$

kde p_k je distribúcia stupňov, n_k je počet vrcholov s k susedmi, n je počet vrcholov grafu.

1.2 Modely sietí

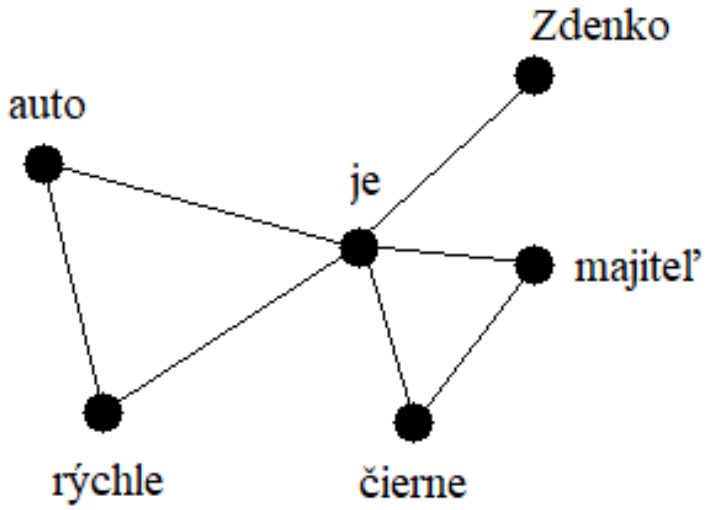
1.2.1 Náhodná sieť

1.2.2 Bezškálová sieť

1.2.3 Dorogovtsev Goltsev Mendes model

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

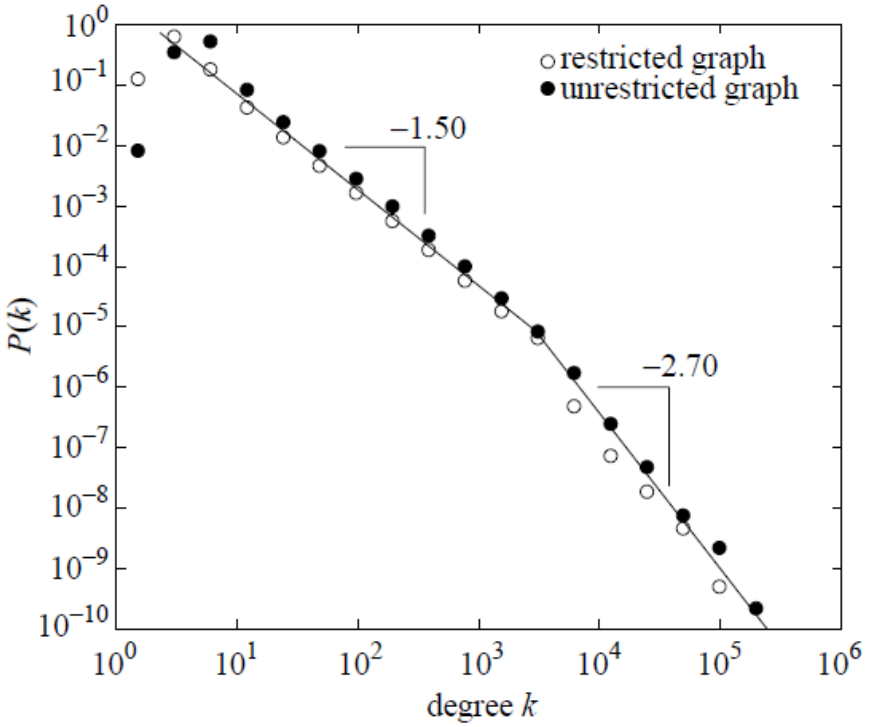


Obr. 1.3: Pozičná slovná sieť z viet: „Auto je rýchle. Auto je čierne. Majiteľ je Zdenko.“.

a vzťahy medzi slovami, ktoré sa navzájom ovplyvňujú alebo súvisia tvoria hrany[4]. 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 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, susedia. Ukážku takejto siete vidíme na obrázku č. 1.3.

Tieto siete majú niekoľko zaujímavých vlastností. Koeficient zhlukovania, ktorý nám určuje, do akej miery slová tvoria skupiny alebo často používané frázy. Distribúcia stupňov uzlov, ktorá sa často podobá na rozdelenie typu mocninného zákona, čo znamená, že väčšina slov má veľmi nízky stupeň, zatiaľ čo malý podiel slov má veľmi vysoký stupeň[5]. Toto rozdelenie je typické pre bezškálové siete. Na obrázku č. 1.4 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 $y_{BA} = -3$[6].



Obr. 1.4: Distribúcia stupňov vrcholov pre dve siete, body zoskupené mocninou dvoch[6].

1.4 NetworkX

Text ohľadom knižnice NetworkX.

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] 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.

[5] 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.

[6] 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.