

VŠB TECHNICKÁ
UNIVERZITA
OSTRAVA

VSB TECHNICAL
UNIVERSITY
OF OSTRAVA



www.vsb.cz

Kompresa stromových struktur

Semestrální projekt

Marek Beran

VŠB – Technická univerzita Ostrava

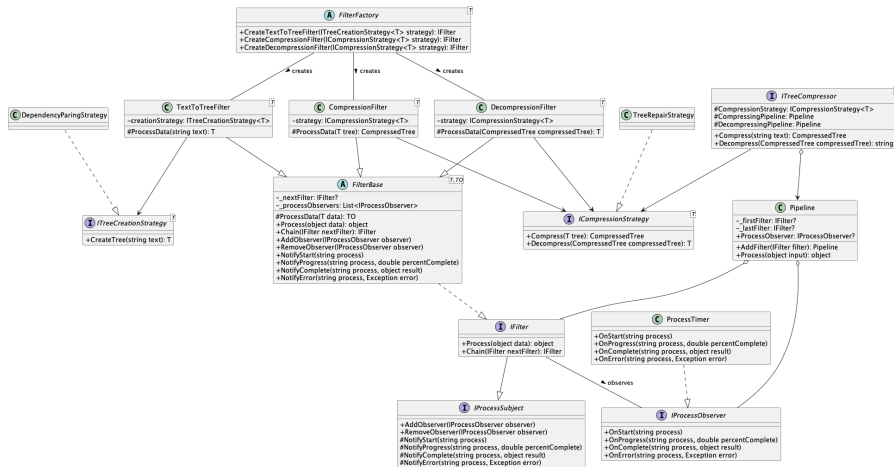
marek.beran.st@vsb.cz

27. května 2025



- Hlavní cíl: Ověřit, zda je možné efektivně komprimovat přirozený jazyk pomocí převodu do stromové struktury a následné komprese stromu

Implementace knihovny



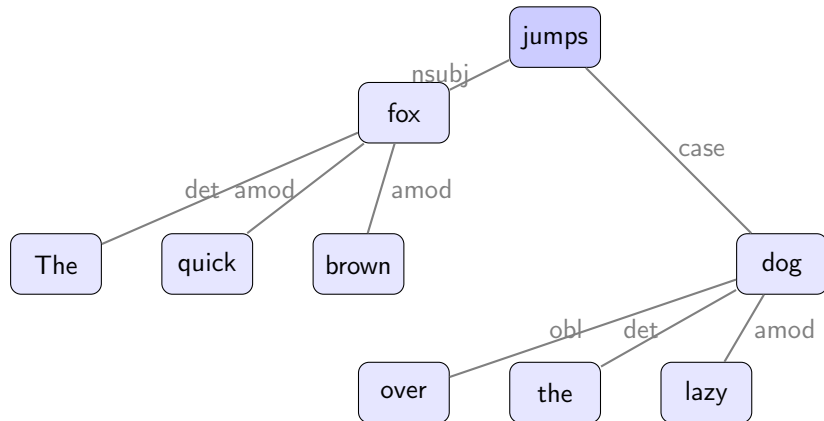
Obrázek: Třídní diagram části implementace zaměřený na řetězení filtrů

Převod textu do stromové struktury



- Dependency parsing - závislosti mezi slovy ve větě
- Využití knihovny UDPipe

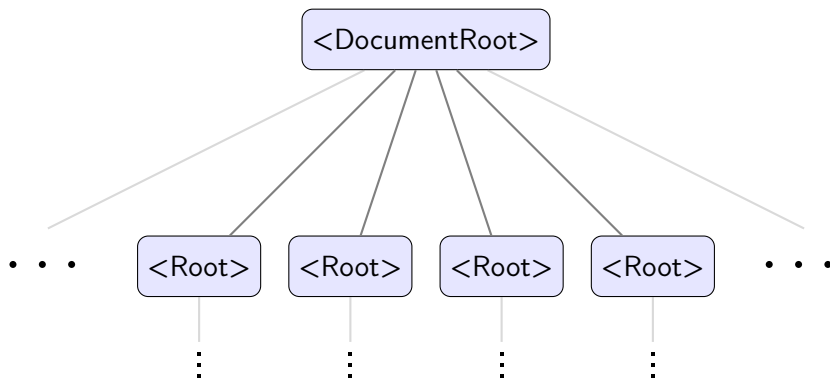
"The quick brown fox jumps over the lazy dog."



Umělé rozšíření stromu



- Rozšíření stromu pro podporu více vět bez nutnosti práce s lesem
- Pro zajištění dostatečné velikosti závislostího stromu pro testování a kompresi





- Inspirace algoritmem RePair pro kompresi řetězců
- Dva hlavní přístupy k aplikaci na stromové struktury:
 - 1 **Linearizace + komprese** – převod stromu na posloupnost, aplikace RePair
 - 2 **Přímá komprese stromu** – adaptace RePair pro práci přímo se stromovou strukturou



Princip:

- Převod stromu na lineární posloupnost uzlů – experimentování s různými metodami linearizace (Preorder, Postorder, Level-order)
- Aplikace standardního RePair algoritmu na tuto posloupnost
- Optimalizace: rozšíření z párů (pairs) na obecné n-tice (n-grams) pro lepší kompresi dlouhých vzorů vytvořených kvůli linearizaci

**Princip:**

- Algoritmus inspirovaný TreeRePair, využívaného pro kompresi XML dokumentů
- Identifikace opakujících se dvojic (digramů) rodič–potomek + pozice potomka v rámci ostatních potomků
- Vytváření gramatických pravidel pro podstromy



Klíčové komponenty:

- **Digram:** pár (rodič, dítě) reprezentující hranu ve stromě
- **Četnost:** počet výskytů daného digramu v celém stromě
- **Pravidlo:** náhrada opakujícího se vzoru neterminálem

Iterativní proces:

- 1 Identifikace nejčastějšího digramu
- 2 Vytvoření nového gramatického pravidla
- 3 Nahrazení všech výskytů digramu neterminálem
- 4 Aktualizace indexu digramů
- 5 Opakování dokud je komprese efektivní



Heuristiky pro ukončení:

- Sledování kompresního zisku po každé iteraci
- Ukončení při poklesu zisku pod stanovený práh

Post-processing optimalizace:

- Odstranění přebytečných pravidel (neterminál \rightarrow neterminál)
- Využití tranzitivního uzávěru pro zjednodušení gramatiky

Složitost:

- Časová: $O(n)$ s omezením iterací, jinak $O(n^2)$



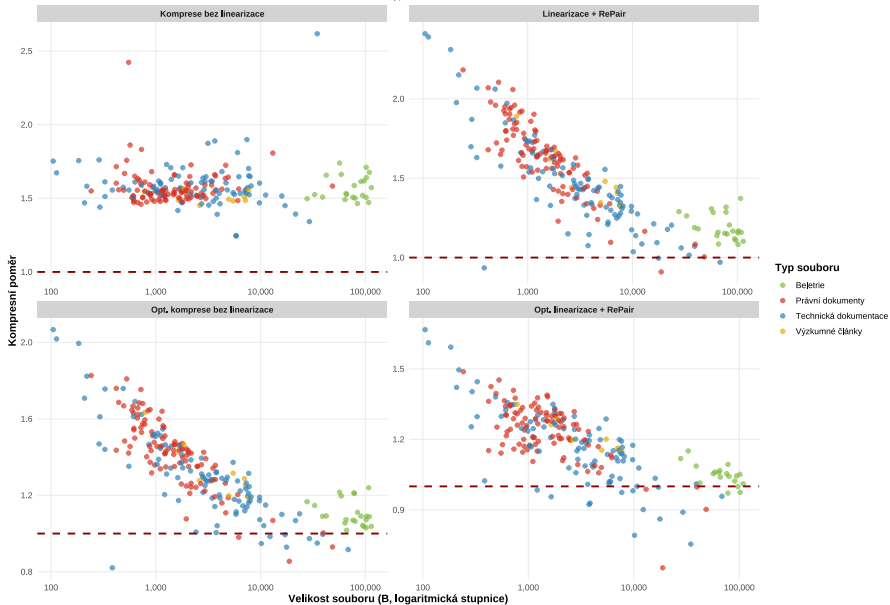
- 4 různé typy textu
- Celkový počet souborů: 242
- Celkový objem 10 MB
- Všechny texty jsou v angličtině

Typ textu	Počet souborů	Rozsah velikostí
Beletrie	23	28 – 120 KB
Právní dokumenty	103	1 KB – 800 KB
Technická dokumentace	96	<1 KB – 100 KB
Vědecké články	20	1 KB – 100 KB

Tabulka: Typy textu a jejich velikosti

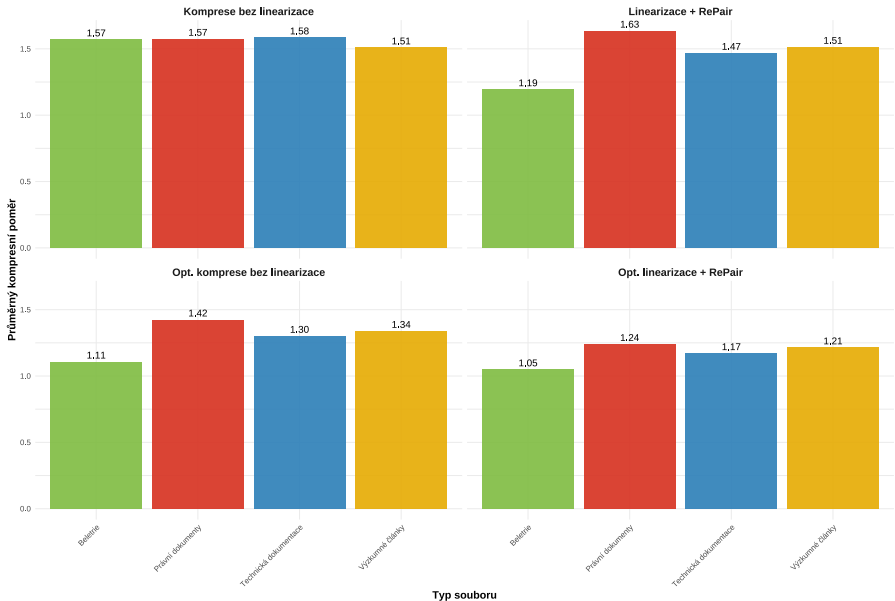
Kompresní poměr podle velikosti souboru

Barva = typ souboru






Srovnání kompresních poměrů stromových algoritmů

Průměrný kompresní poměr podle metody a typu souboru





-  Katja Filippova a Michael Strube. “Dependency tree based sentence compression”. In: *Proceedings of the Fifth International Natural Language Generation Conference*. 2008, s. 25–32.
-  Daniel Jurafsky a James H Martin. “Speech and Language Processing: An introduction to Natural Language Processing”. In: *Computational Linguistics, and Speech Recognition with Language Models. Third Edition draft* (2024).
-  Sandra Kübler, Ryan McDonald a Joakim Nivre. “Dependency parsing”. In: *Dependency parsing*. Springer, 2009, s. 11–20.
-  Markus Lohrey, Sebastian Maneth a Roy Mennicke. “XML tree structure compression using RePair”. In: *Information Systems* 38.8 (2013), s. 1150–1167.



-  Colt McAnlis a Aleks Haecky. *Understanding compression. data compression for modern developers*. 1st Edition. Sebastopol, CA: O'Reilly, 2016. ISBN: 978-1-491-96153-7.
-  F. Oquendo, J. Leite a T. Batista. "Pipe-filter architectural style". In: *Undergraduate Topics in Computer Science* (2016), s. 171–177. DOI: 10.1007/978-3-319-44339-3_13.
-  Milan Straka a Jana Straková. "Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe". In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver, Canada: Association for Computational Linguistics, srp. 2017, s. 88–99. URL: <http://www.aclweb.org/anthology/K/K17/K17-3009.pdf>.



Jana Straková, Milan Straka a Jan Hajič. “Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition”. In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Baltimore, Maryland: Association for Computational Linguistics, čvn. 2014, s. 13–18. URL: <http://www.aclweb.org/anthology/P/P14/P14-5003.pdf>.

Děkuji za pozornost

Marek Beran

VŠB – Technická univerzita Ostrava

marek.beran.st@vsb.cz

27. května 2025