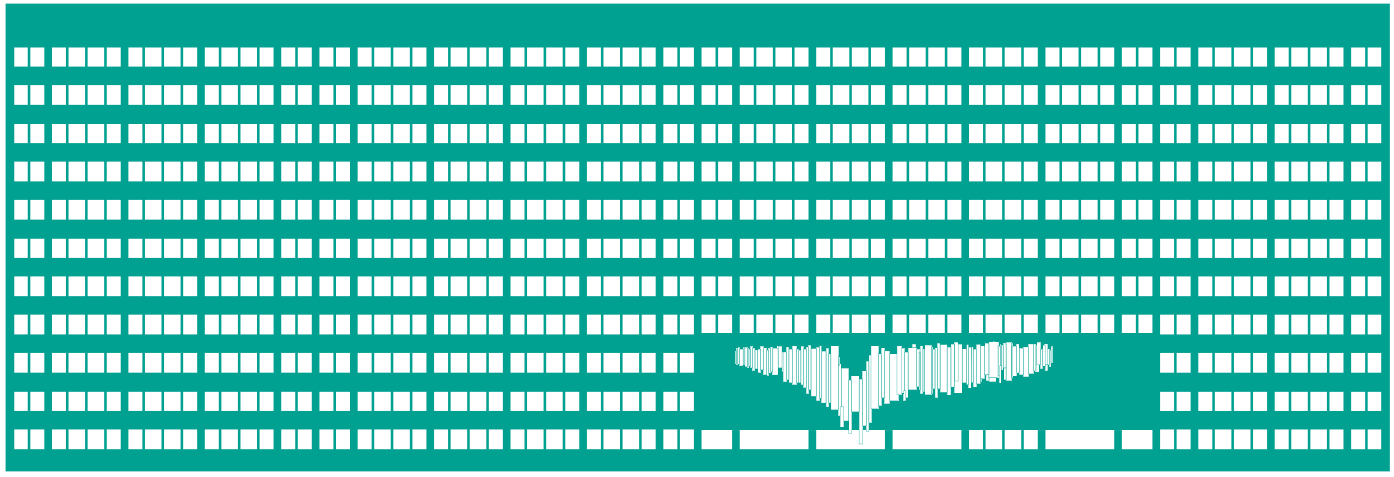
VŠB TECHNICKÁ UNIVERZITA OSTRAVA

VSB TECHNICAL UNIVERSITY OF OSTRAVA



www.vsb.cz

Komprese stromových struktur

Semestrální projekt

Marek Beran

VŠB – Technická univerzita Ostrava

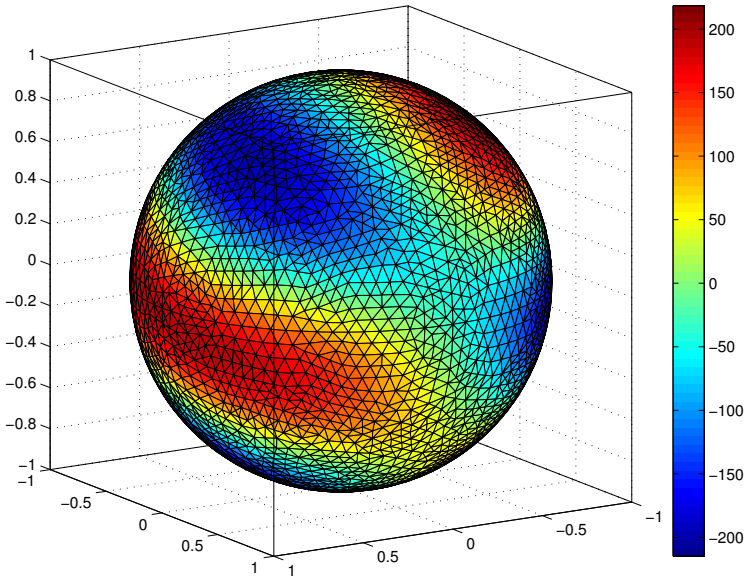
marek.beran.st@vsb.cz

23. května 2019

VŠB TECHNICKÁ UNIVERZITA OSTRAVA | FAKULTA ELEKTROTECHNIKY A INFORMATIKY

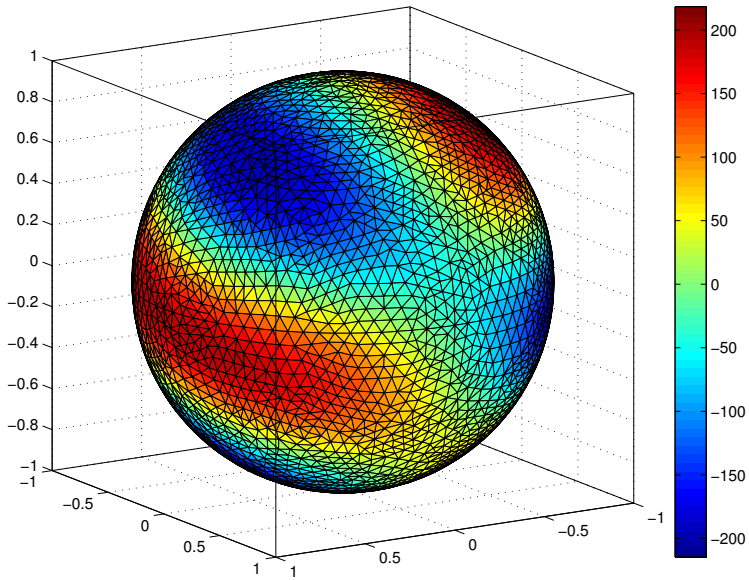
Motivace projektu

- Komprese dat - klíčová oblast informatiky
- Řešení pro sekvenční data běžné (text, multimédia)
- Stromové struktury představují specifickou výzvu
- Přirozený jazyk obsahuje opakující se syntaktické vzorce
- Syntaktické stromy - možný cíl optimalizované komprese?



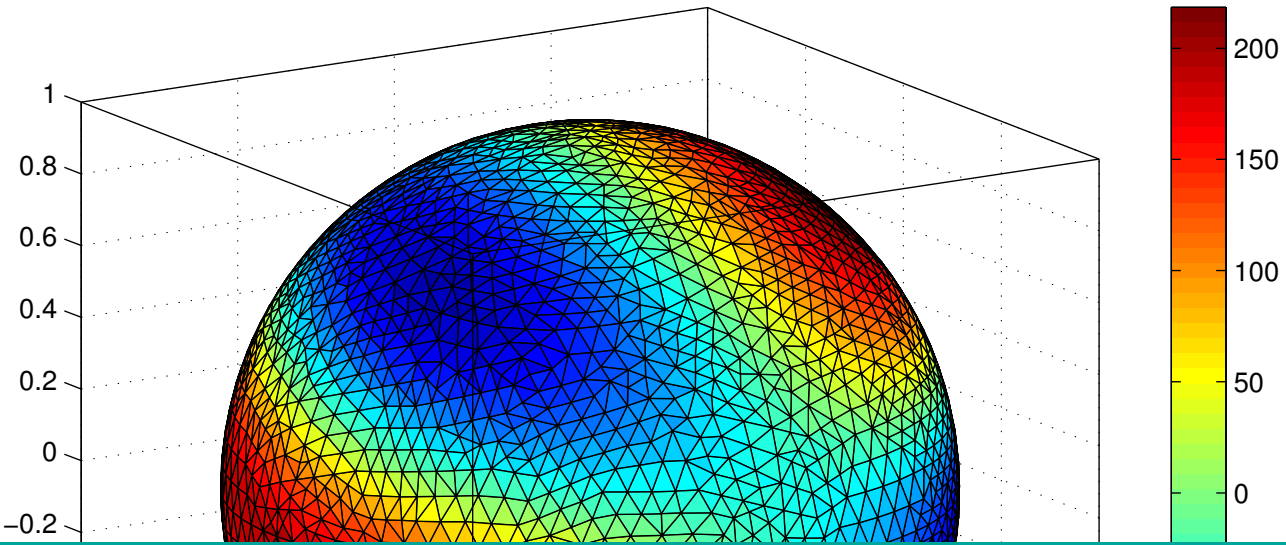
Motivace projektu

- Komprese dat - klíčová oblast informatiky
- Řešení pro sekvenční data běžné (text, multimédia)
- Stromové struktury představují specifickou výzvu
- Přirozený jazyk obsahuje opakující se syntaktické vzorce
- Syntaktické stromy - možný cíl optimalizované komprese?



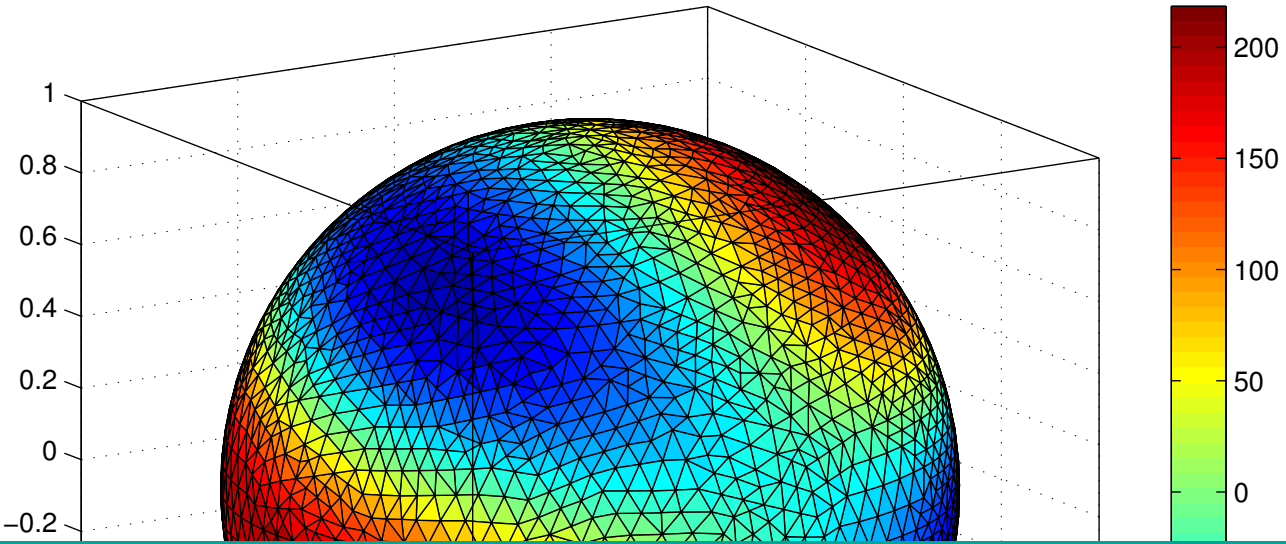
Cíle projektu

- Navrhnout a implementovat knihovnu pro kompresi stromových struktur
- Zaměřit se na syntaktické stromy vytvořené z přirozeného jazyka
- Ověřit hypotézu o možné efektivní kompresi využitím opakujících se vzorců
- Srovnat různé přístupy ke kompresi stromových struktur
- Naivní hypotéza: Velikost komprimovaných stromů roste logaritmicky s délkou textu



Cíle projektu

- Navrhnout a implementovat knihovnu pro kompresi stromových struktur
- Zaměřit se na syntaktické stromy vytvořené z přirozeného jazyka
- Ověřit hypotézu o možné efektivní kompresi využitím opakujících se vzorců
- Srovnat různé přístupy ke kompresi stromových struktur
- Naivní hypotéza: Velikost komprimovaných stromů roste logaritmicky s délkou textu



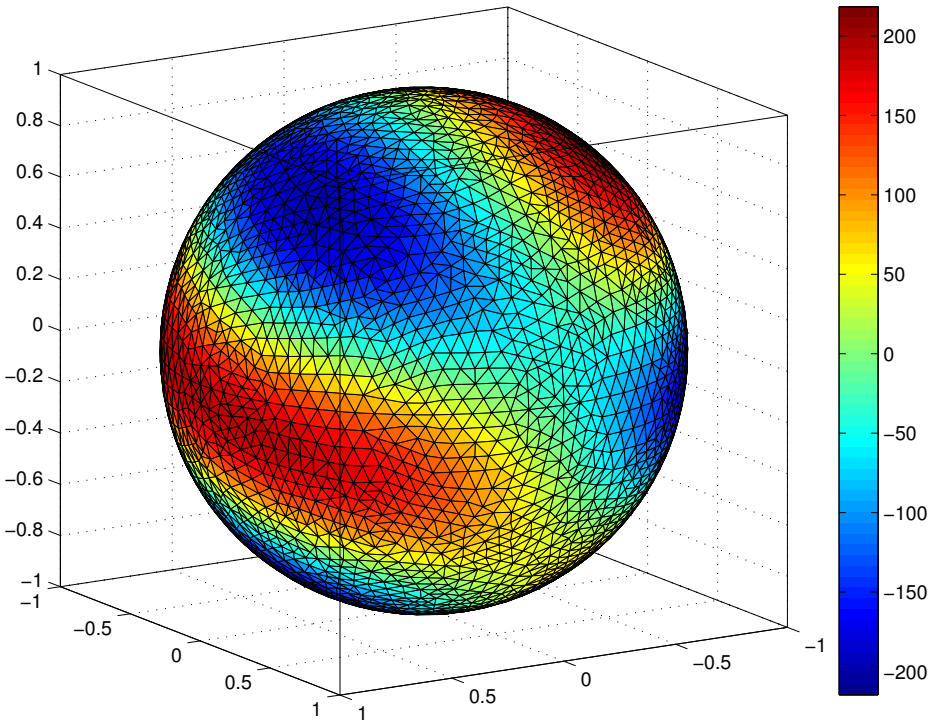
Použité technologie a nástroje

Implementační prostředí:

- Programovací jazyk C#
- Objektově orientovaný přístup
- Vzor Pipes and Filters

NLP nástroje:

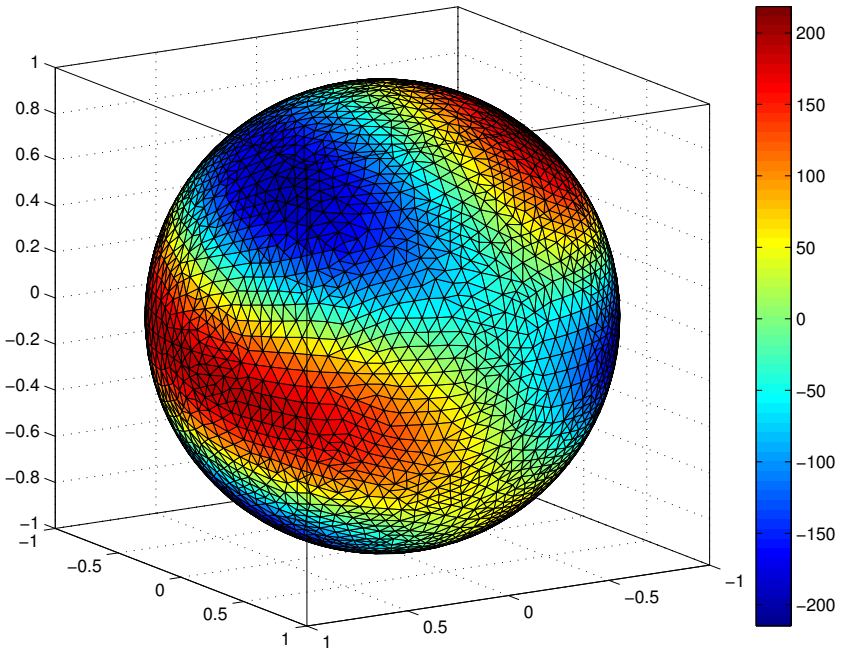
- MorphoDiTa - morfologická analýza
- UDPipe - dependency parsing



Syntaktická analýza

Dependency parsing:

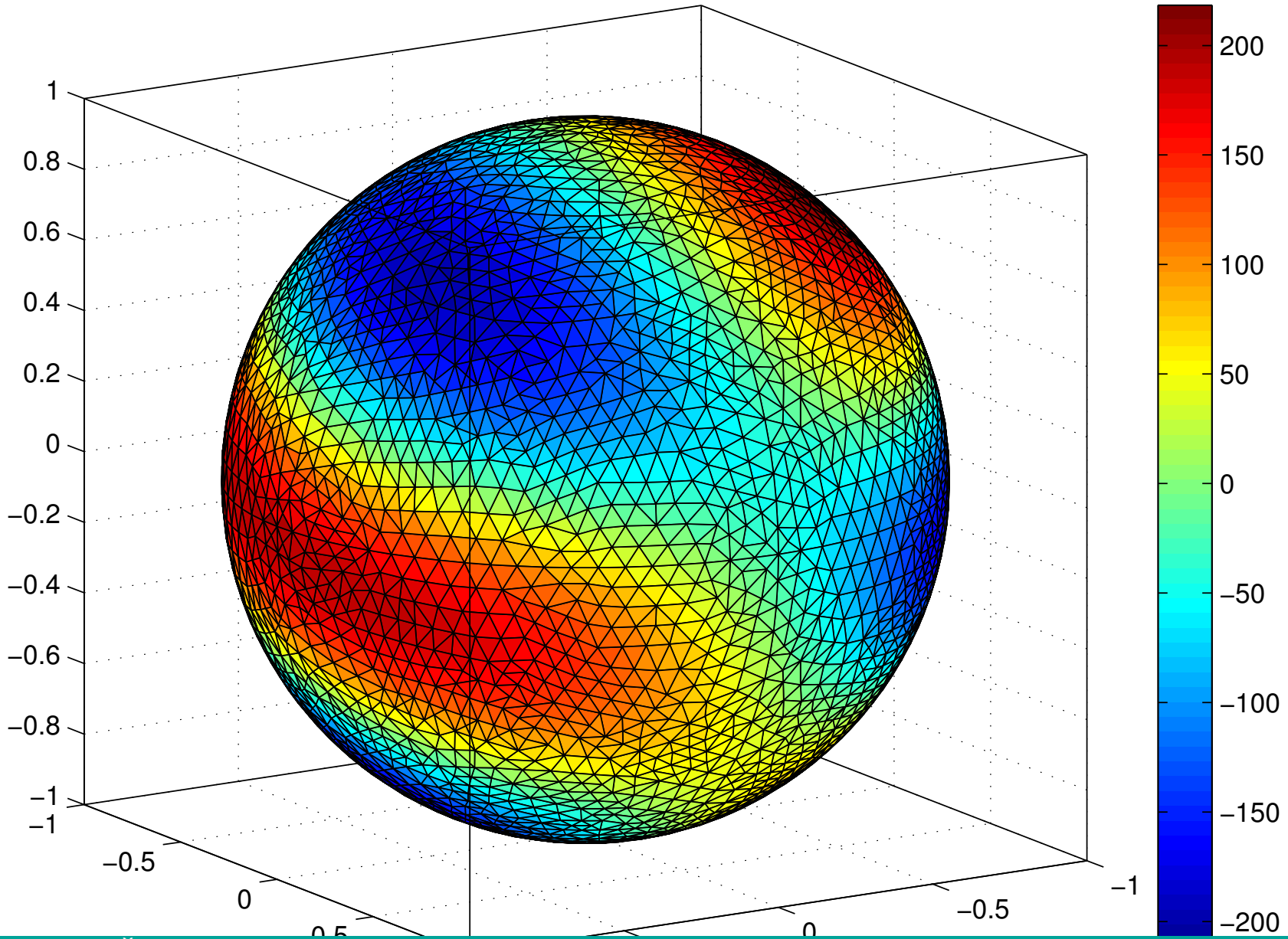
- Metoda pro vytvoření syntaktického stromu
- Uzly = slova, hrany = vztahy mezi slovy
- Zachycuje gramatickou strukturu věty
- Klíčový pro vytvoření vstupních dat



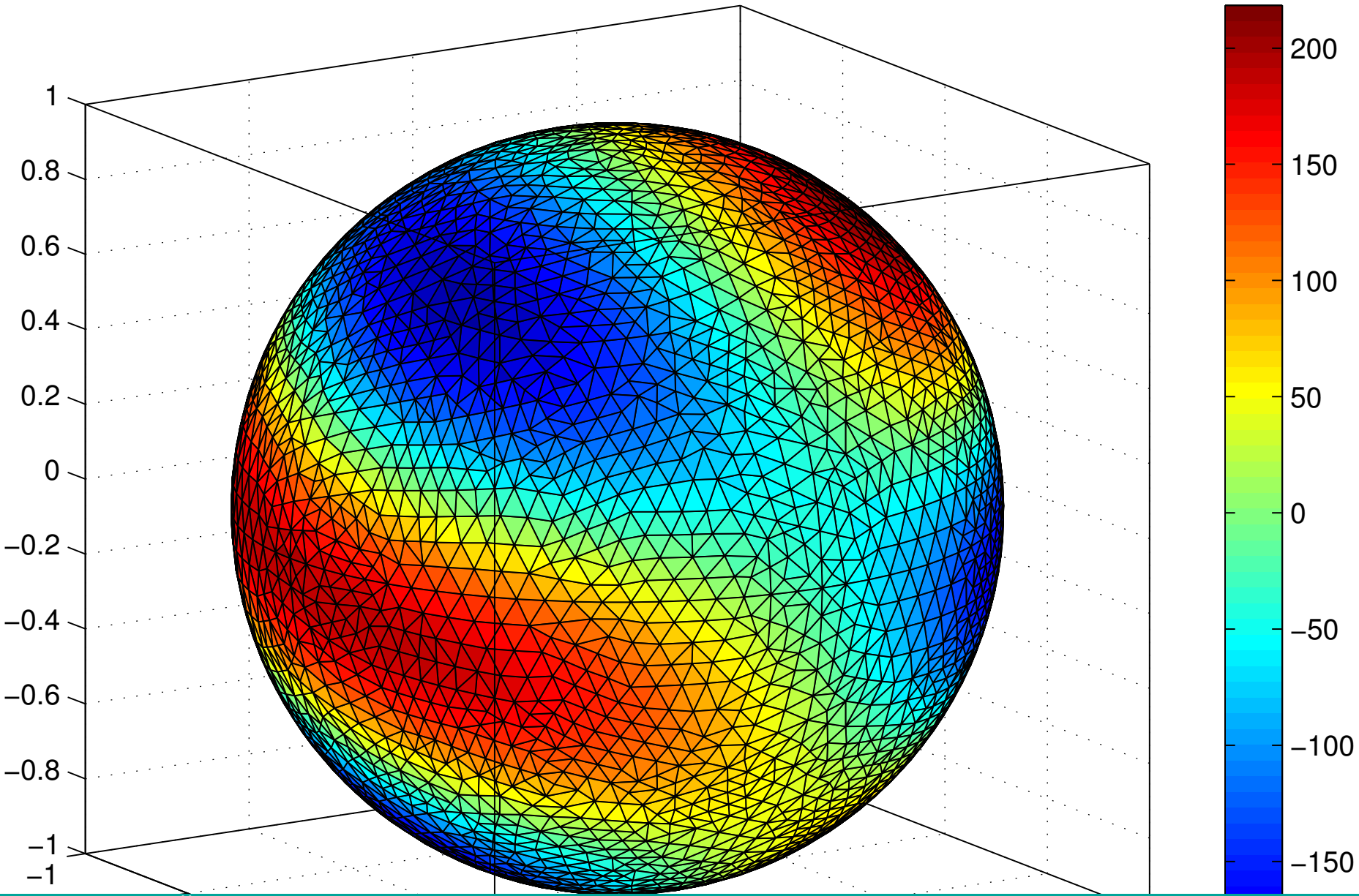
Z věty na strom:

1. Tokenizace a lemmatizace
2. POS tagging
3. Dependency parsing

Návrh knihovny



Pipeline zpracování



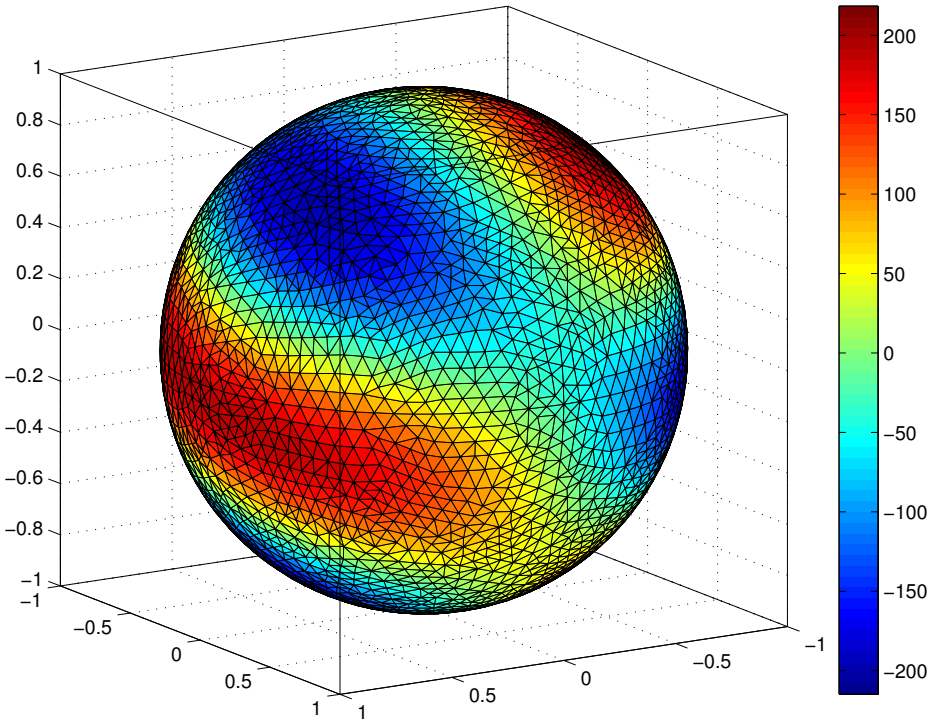
Implementované algoritmy

Algoritmus RePair:

- Recursive Pair - nahrazování opakujících se párů
- Vytváření gramatických pravidel

Dva hlavní přístupy:

1. Přímá komprese stromu bez linearizace
2. Linearizace + aplikace RePair



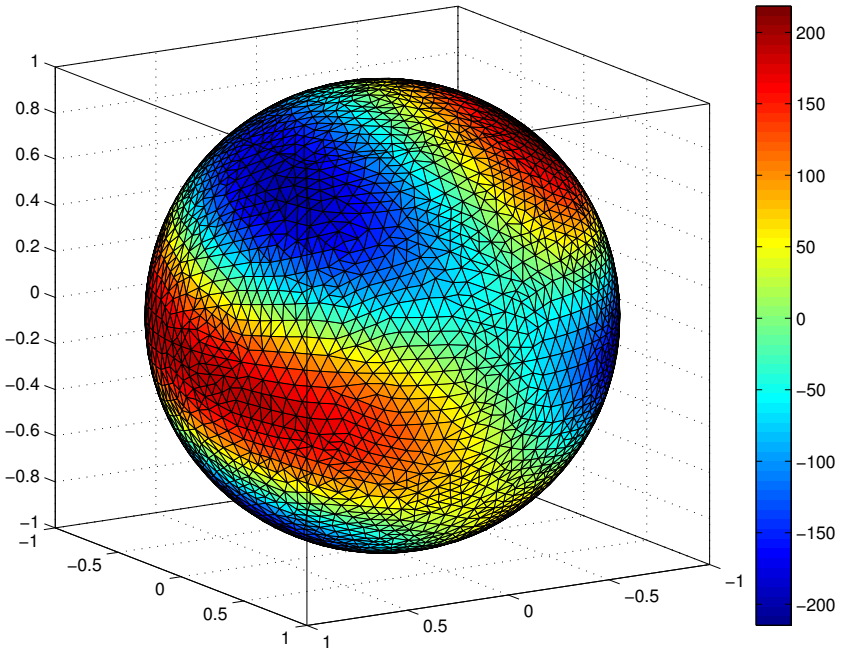
TreeRePair - komprese bez linearizace

Základní varianta:

1. Identifikace opakujících se podstromů
2. Výběr nejčastějších
3. Nahrazení neterminály
4. Vytvoření gramatických pravidel

Optimalizovaná varianta:

- Komplexní metrika pro výběr podstromů
- Prioritizace podle velikosti, četnosti a hloubky
- Paměťové optimalizace



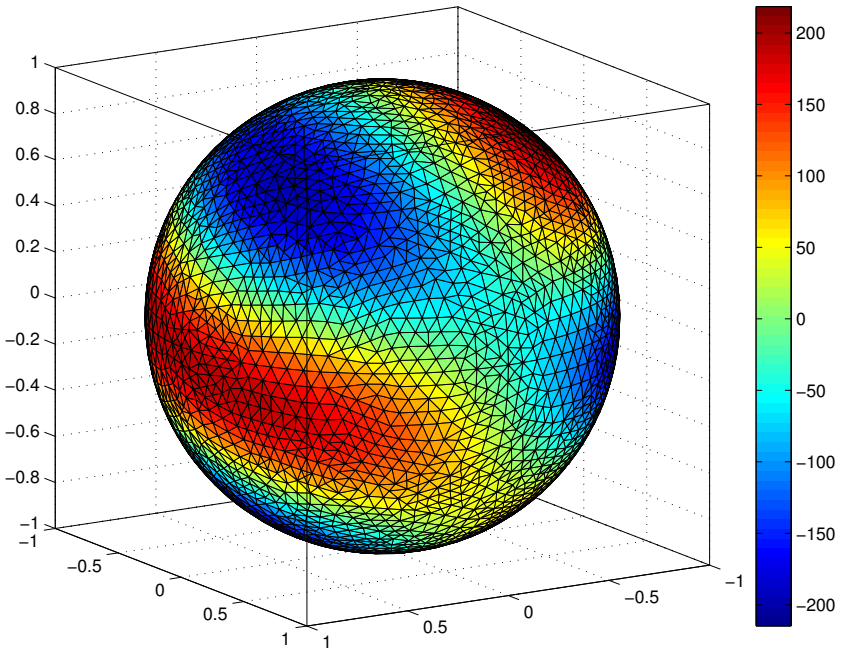
Linearizace a RePair

Základní varianta:

1. Preorder průchod stromem
2. Vytvoření lineární sekvence
3. Aplikace RePair na sekvenci

Optimalizovaná varianta:

- Vylepšený způsob linearizace
- Rozšíření na n-gramy místo digramů
- Kontextově citlivé nahrazování
- Optimalizace gramatiky



Metodika testování

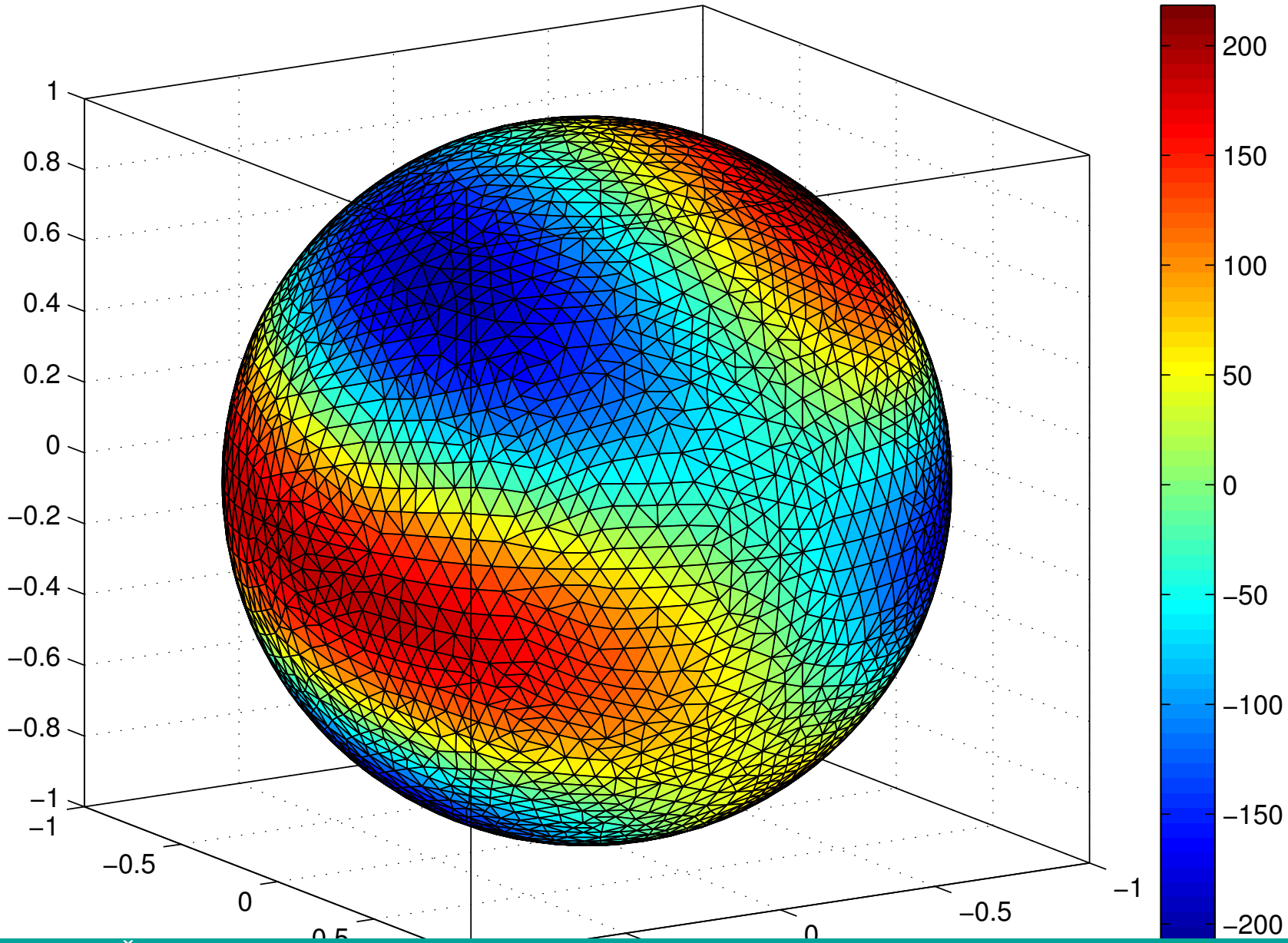
Testovací data:

- Technická dokumentace - specifický jazyk, časté opakování výrazů
- Próza - proměnlivý styl, složitější větné konstrukce
- Právní dokumenty - formální jazyk, přísná pravidla struktury

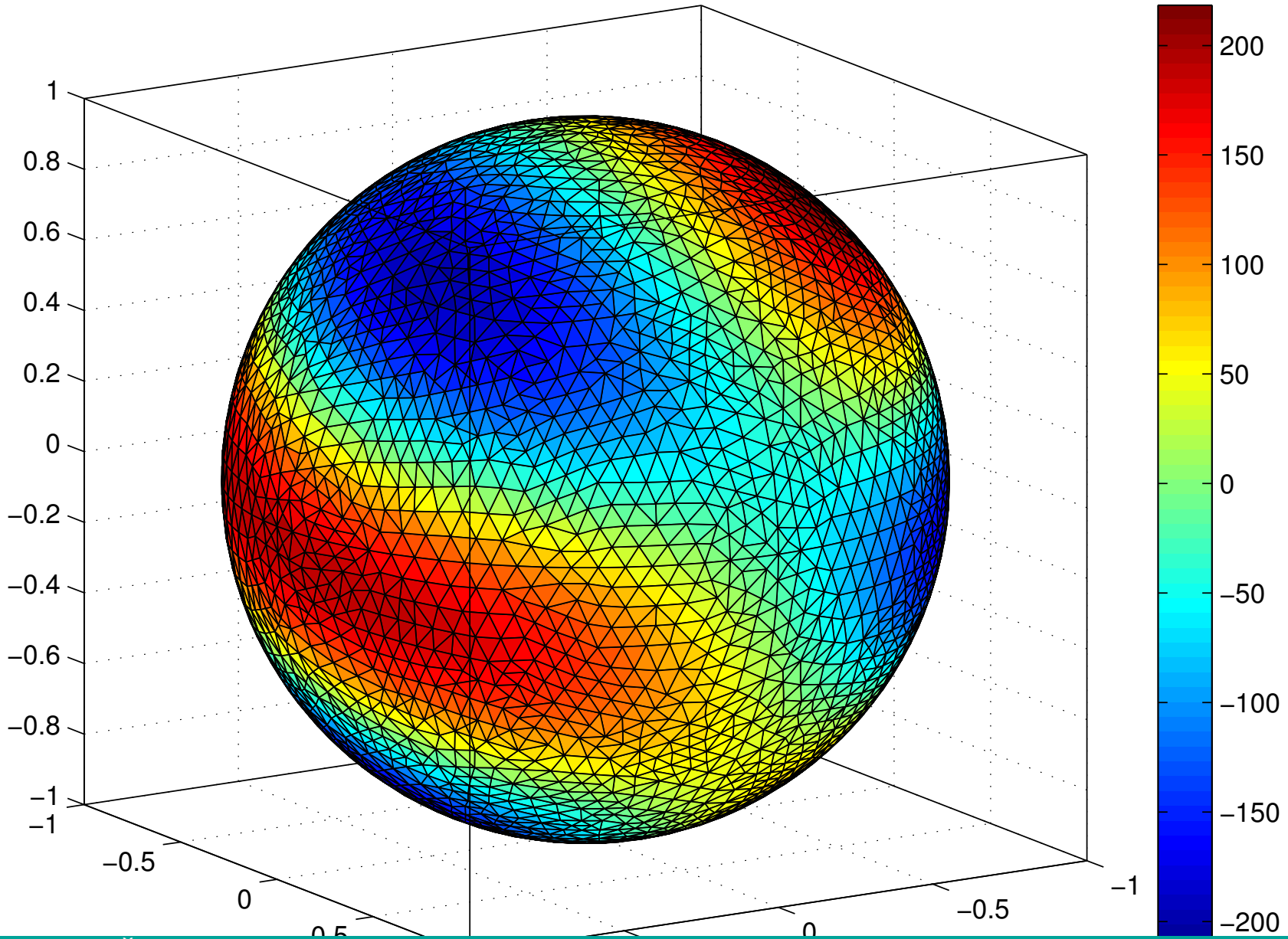
Měřené metriky:

- Kompresní poměr - poměr velikosti komprimovaných dat ku původním
- Kompresní zisk - procentuální úspora místa
- Doba komprese a dekomprese
- Paměťová náročnost

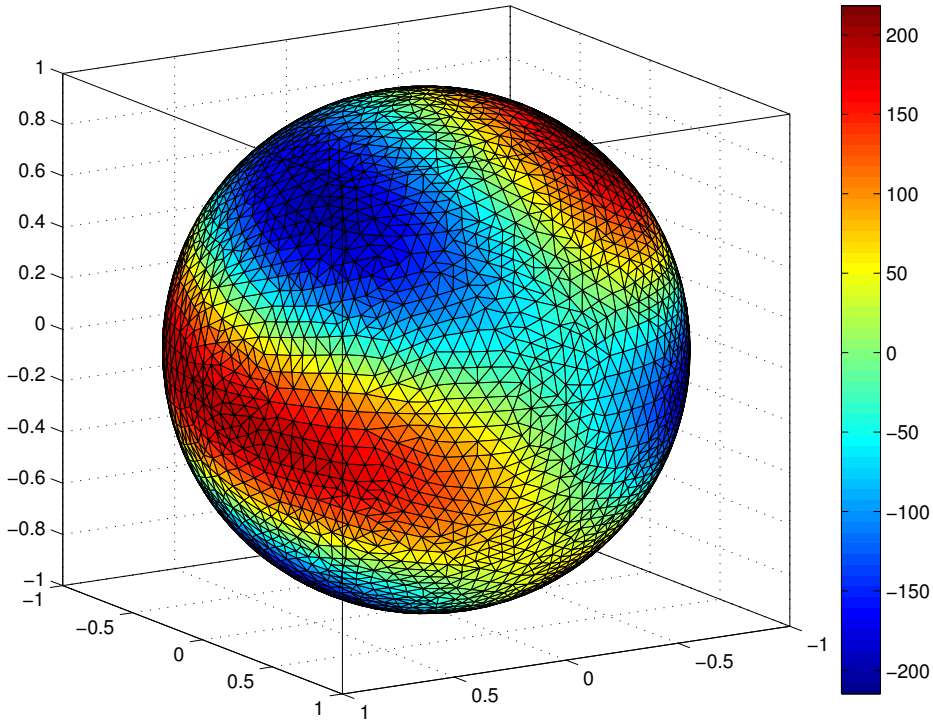
Porovnání kompresních poměrů

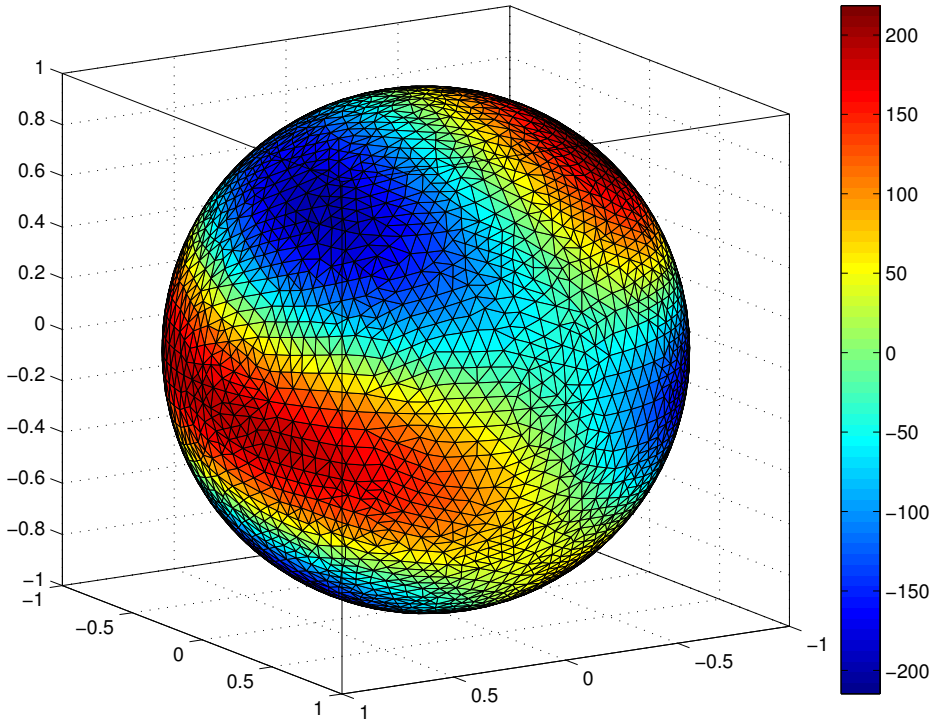


Vliv velikosti vstupních dat



Vliv optimalizací a linearizace





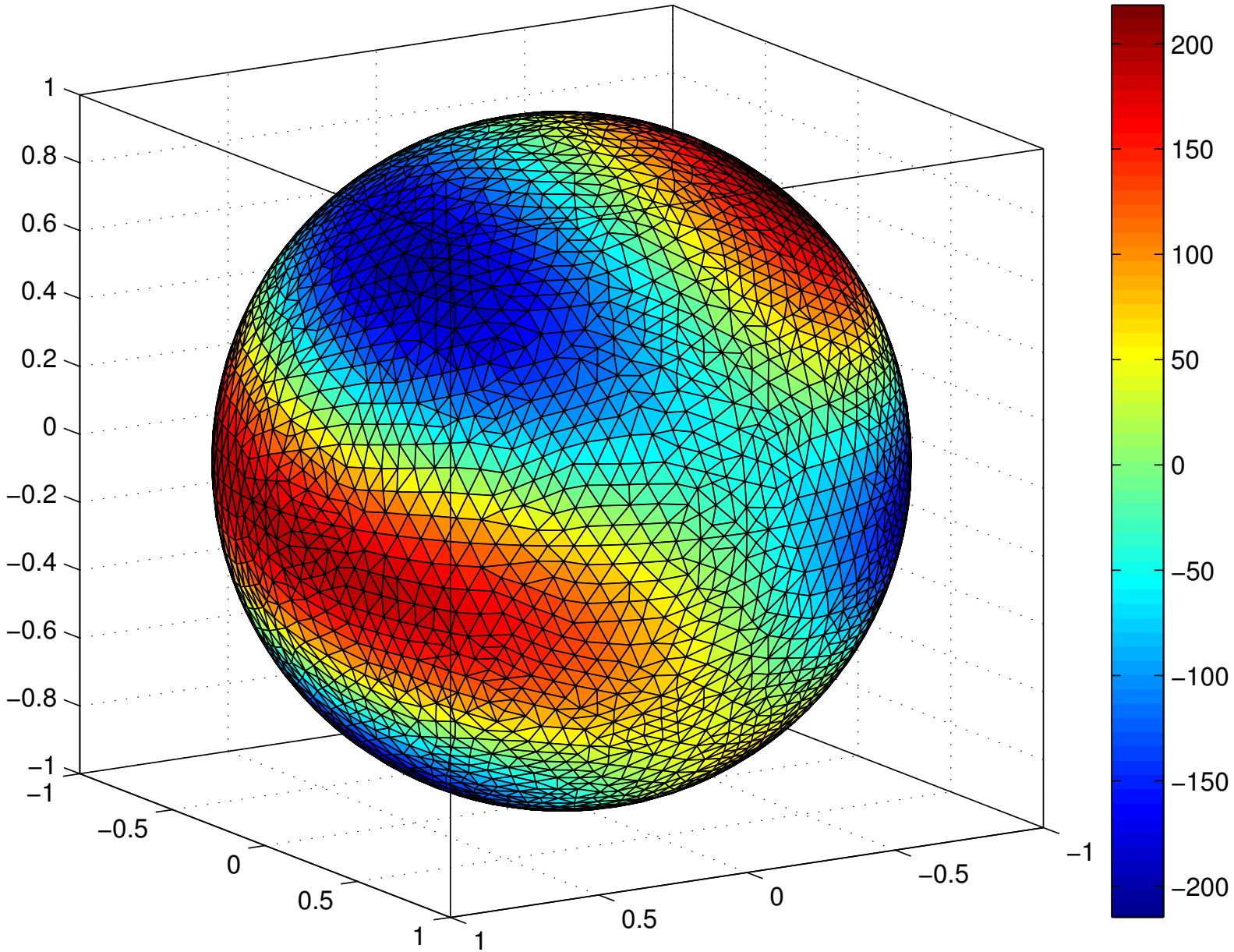
Vliv optimalizací:

- Zlepšení průměrného kompresního poměru o 14-24%
- Největší přínos u právních dokumentů

Vliv linearizace:

- Algoritmy s linearizací obecně efektivnější
- Výraznější rozdíl u větších souborů

Časová náročnost



Shrnutí výsledků

Hlavní zjištění:

- Komprese stromových struktur představuje značnou výzvu
- Nejlepších výsledků dosahuje optimalizovaná linearizace s RePair
- Skutečná komprese (poměr $< 1{,}0$) dosažena jen v 9% případů
- Větší soubory umožňují efektivnější kompresi
- Optimalizace přinesly výrazné zlepšení oproti základním variantám

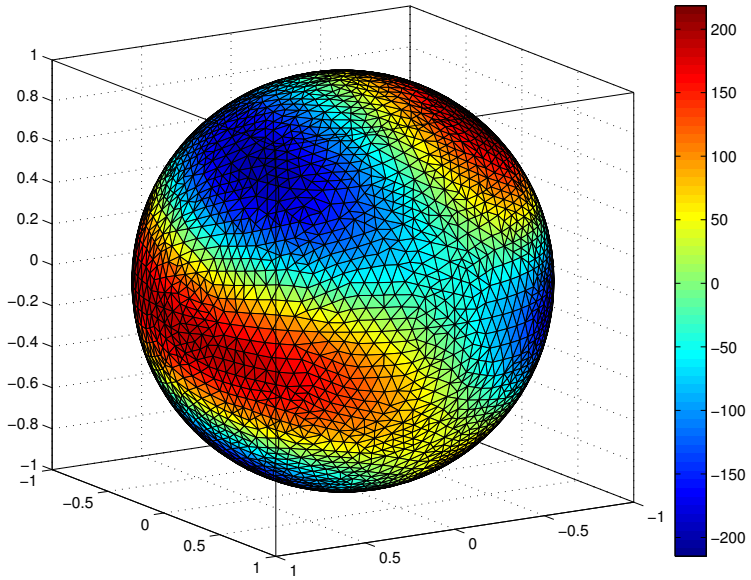
Ověření hypotézy:

- Logaritmický trend potvrzen, ale ne tak výrazný
- Opakující se vzorce existují, ale jejich využití není tak efektivní

Možnosti dalšího vývoje

Směry budoucího výzkumu:

- Adaptivní přepínání mezi variantami algoritmů
- Další optimalizace linearizace
- Paralelizace zpracování
- Hybridní přístupy kombinující různé metody
- Specializace na konkrétní typy textů
- Ztrátová komprese stromových struktur



Děkuji za pozornost

Marek Beran

VŠB – Technická univerzita Ostrava

marek.beran.st@vsb.cz

23. května 2019

VŠB TECHNICKÁ UNIVERZITA OSTRAVA | FAKULTA ELEKTROTECHNIKY A INFORMATIKY