



TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky
a mezioborových studií ■

Metody vytěžování dat

Úvod

František Kynych
6. 10. 2022 | MVD





TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky
a mezioborových studií ■

Část I.: Organizace předmětu



Organizace předmětu MVD

- Přednášející i cvičící František Kynych (frantisek.kynych@tul.cz)
- Přednášky
 - Každý čtvrtek od 12:30
 - Témata zveřejněna v sylabu (upravený oproti původnímu)
 - PDF přednášky bude zveřejněno na Githubu
- Cvičení
 - Každý čtvrtek od 14:20, navazuje na přednášku
 - 2 dovolené absence
 - Implementace témat z předchozí přednášky

Organizace předmětu MVD

- Cvičení
 - Python
 - Možnost získání bonusových bodů
 - Splnění bonusové úlohy
 - Zajímavé nebo nadstandardně propracované řešení úlohy
- Zápočet
 - Docházka
 - Splnění všech úloh ze cvičení
- Zkouška
 - Písemná
 - Úspěch > 60 %
 - Details budou upřesněny ke konci semestru

Organizace předmětu MVD

Aktualizovaný syllabus

1. Úvod do problematiky
2. Vizualizace dat
3. Vektorizace textu – Word2Vec a GloVe
4. Vyhledávání
5. Vyhledávání 2
6. Vyhledávání na webu
7. Shlukování
8. Shlukování 2
9. BERT model
10. Kategorizace dokumentu podle tématu, detekce sentimentu
11. Doporučovací systémy
12. Detekce anomálií
13. Vyhledávání vzorů
14. Genetické algoritmy

Cíle předmětu

- Seznámení se základními principy NLP, vizualizace a vytěžování dat
 - Jak je implementovat nebo používat
 - Jak je vyhodnotit
 - Jak je vylepšit
 - Kde a jak hledat informace z aktuálního výzkumu

Část II.: Úvod do předmětu

Vytěžování dat

- Velké množství dat vzniká každým dnem
 - Kvůli jejich množství nejsme schopni vše procházet ručně
- Strojová data
 - Logy a výstupy všech možných systémů
 - Senzory a různá zařízení
- Lidská data
 - Chování uživatelů
 - Chaty, příspěvky, komentáře, různé dokumenty



Vizualizace dat

- Grafická reprezentace informací a dat
- Využití pro
 - Získání vhledu do dat
 - Analýzu dat
 - Zobrazení a prezentaci výsledků
 - Zobrazení real time informací o systémech (status, statistiky, ...)
 - ...
- Potřeba téměř u každé aplikace



Vizualizace dat pro ML

- Vizualizace
 - Průběhu učení
 - Stavů systému
 - Výsledků
 - Struktury modelu
- Užitečné nástroje během celého životního cyklu ML modelu



Vyhledávání

- Důležité pro získání informace z velkého množství dat
- Rychlé získání informace
- Vertikální vyhledávače
 - Zaměřené na více specifický problém (nepracují se všemi daty)
 - Poskytují lepší výsledky v dané oblasti
 - Vyhledávání v obchodech, Google Scholar, ...



Vyhledávání na webu

- Rozšíření základní úlohy vyhledávání
- Používá další informace kromě základního textu
 - Analýza odkazů
 - Analýza chování uživatele
 - Počítání prokliků
 - Využití stáří stránky
 - Mnoho dalších vylepšení



Shlukování

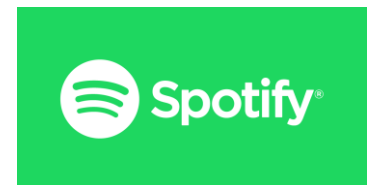
- Unsupervised learning
 - Data neobsahují labely
- Seskupení dat dle podobnosti do shluků
- Chceme, aby podobnost dat uvnitř shluků byla vysoká a mezi různými shluky nízká
- Aplikace v
 - Segmentaci trhu
 - Analýza sociálních sítí
 - Segmentace obrazu
 - ...

Detekce sentimentu

- Extrakce subjektivního pohledu (náklonost / odpor)
- Detekce kladného, záporného, neutrálního nebo bipolárního sentimentu
 - Bipolarita = kladný a záporný zároveň
- Složitá úloha
 - Tón
 - Polarita
 - Sarkasmus
 - Bias

Doporučovací systémy

- Doporučení dalších služeb nebo produktů založené na historii uživatele
 - Navrhování hudby, zboží, filmů, produktů atd.
- Služba se snaží udržet uživatele déle na jeho platformě nebo o zakoupení produktu uživatelem
 - „The Attention Economy“
 - Možnost vytvoření závislosti uživatele
- Content-based
- Collaborative filtering



Detekce anomálií

- Detekuje datové body, které se odlišují od normálního chování
- Může se jednat o kritický incident, technickou chybu nebo změnu chování uživatelů
- Typy anomálií
 - Global
 - Bod vychylující se z celého datasetu
 - Contextual
 - Nevychylují se z datasetu, ale v pouze dané oblasti
 - Collective
 - Vychyluje se podmnožina datasetu

Vyhledávání vzorů

- Automatický proces rozpoznání vzorů nebo zákonitostí v datech
- Co je to vzor?
 - Skupina datových bodů nebo sekvencí, která se společně často nebo pravidelně vyskytuje v datasetu
- Aplikace
 - Jaké produkty se často kupují dohromady
 - Pokud si uživatel koupí nějaký produkt, co si s nejvyšší pravděpodobností koupí poté
 - Hledání inherentních pravidelností v datech

Genetické algoritmy

- Počátek v 60. letech (J. Holland)
- Heuristický postup, který se snaží aplikací principů evoluční biologie nalézt řešení složitých problémů
- Použité techniky napodobují evoluční procesy známé z biologie
 - Dědičnost
 - Mutace
 - Přirozený výběr a křížení
- Cílem je „vyšlechtit“ řešení zadané úlohy
- Conway's Game of Life

Textová data

- Primární zaměření tohoto předmětu
- Obsahují velké množství informací
 - Např. názory a preference lidí

⇒ Vytvoření inteligentních systémů na pomoc lidem
- Generováno a konzumováno člověkem
 - Potřebuje pomoc při procházení obsahu

Co to je NLP?

- Natural Language Processing
- Soubor technik pro zpracování, analýzu a generování textu
- Počátek v 50. letech minulého století (Alan Turing)
 - Turingův test
- Snaha vytvoření programů k porozumění přirozeného jazyka
- Kombinace lingvistiky, statistiky, machine learningu (ML) a deep learningu (DL)

Proč je NLP složitý?

- Jedna věta lze vyjádřit více způsoby
- Potřeba porozumění textu i jeho smyslu
 - Různé jazyky mohou mít různá pravidla
- Použití sarkasmu, abstrakce, dvojznačnost

Příklady aplikací #1

- Vyhledávání
- Filtrování obsahu
 - Detekce spamu, reklamy
- Doporučování obsahu
- Kategorizace
- Extrakce informací
- Detekce sentimentu

Příklady aplikací #2

- Rozpoznání řeči
- Part of Speech tagging (PoS)
 - Rozpoznání slovních druhů
- Rozlišování smyslu slova
- Named Entity Recognition (NER)
 - Identifikace názvu produktu
- Rozlišování odkazování
 - Na koho se v textu nepřímě odkazujeme (on, ona, ...)
- Generování textu
- Strojový překlad
- Sumarizace textu
- ...

Současný stav

- Vyřešeno
 - PoS
 - Klasifikace textu
 - Detekce spamu
 - NER
 - Detekce jmen, lokalit, organizací, ...
- Velké pokroky
 - Analýza sentimentu
 - Rozlišování odkazování
 - Určení významu slova z jeho kontextu
 - Strojový překlad

Současný stav

- Velké výzvy
 - Obecné dialogové systémy a chat boti
 - Odpovídání na otázky ([Odkaz na současné state-of-the-art](#))
 - NLP pro jazyky s nízkými zdroji
 - Nedostatek dat pro některé jazyky
 - Univerzální jazykový model
 - Vícejazyčné modely

Část III.: Základní metody

Tokenizace

- Rozdělení textu na menší prvky (tokeny)
+ Odstranění interpunkce

Metody vytěžování dat -> Metody vytěžování dat

- Závislá na jazyku
- Je potřeba jasná definice pravidel

Aren't -> ? Aren't

Arent

Are n't

Aren t

Stop words

- Nežádoucí slova v textu
 - Příliš častý výskyt
 - Ovlivnění výsledku algoritmu

Metody a vytěžování dat ->

Metody

vytěžování

dat

- Např. odstranění spojek a předložek
- Vytvoření stop listu
 - Malé 7-12 slov
 - Velké 200-300 slov
- Moderní systémy je nepotřebují používat (viz Inverse Document Frequency)

Normalizace textu

- Cílem je nalezení slov i v případě odlišností v posloupnosti znaků tokenu
⇒ Redukce slova na jeho základní tvar
- 1. Rozdělení textu na tokeny + lowercase
- 2. Odstranění stop words
- 3. Stematizace (Stemming) / Lematizace
- (4.) PoS tagging

Stemming vs. Lematizace

- Stemming může redukovat token na slovo, které neodpovídá gramatice
 - people -> peopl
 - V některých aplikacích nemusí vadit
- Lematizace převádí token do základního gramatického tvaru
 - am, are, is -> be

N-gramy

- Sekvence N slov
- Využívané k základnímu modelování jazyka

Metody vytěžování dat

Unigramy -> (Metody), (vytěžování), (dat)

Bigramy -> (Metody, vytěžování), (vytěžování, dat)

Trigramy -> (Metody, vytěžování, dat)

- Aplikace např. na slovech, písmenech nebo fonémech

N-gramy

- Pravděpodobnost celé věty – řetězkové pravidlo

$$p(w) = p(w_1)p(w_2|w_1) \dots p(w_k|w_1 \dots w_{k-1})$$

- Markovský předpoklad
 - Současné slovo závisí pouze na pevně daném počtu předchozích slov

$$p(w_i|w_1 \dots w_{i-1}) = p(w_i|w_{i-n+1} \dots w_{i-1})$$

- Příklad bigramového modelu

$$p(w) = \prod_{i=1}^{k+1} p(w_i|w_{i-1})$$

- $k+1$ pro využití end tokenu

N-gramy

- Výpočet pravděpodobnosti

$$p(w_i | w_{i-1}) = \frac{c(w_{i-1}w_i)}{\sum_{w_i} c(w_{i-1}w_i)} = \frac{c(w_{i-1}w_i)}{c(w_{i-1})}$$

- Příklad bigramu:

<s> **Metody** vytěžování dat <end>

<s> Aplikace metody Monte-Carlo <end>

<s> **Metody** psychologie ... <end>

Pravděpodobnost výskytu slova „metody“ na začátku věty:

$$p(\textit{metody} | < s >) = \frac{c(< s > \textit{metody})}{c(< s >)} = \frac{2}{3}$$

Užitečná literatura, kurzy nebo odkazy

- [Natural Language Toolkit \(NLTK\)](#)
- [Fast AI](#)
- [Hugging Face](#)
- [Weights & Biases](#)
- [Papers with Code](#)
- [Coursera NLP kurz od DeepLearning.ai](#)