Vyšší odborná škola
a Střední průmyslová škola elektrotechnická,
Plzeň, Koterovská 85

Ročníková práce s obhajobou

Téma: SMDM - Social Media Data Model

Autor práce: Marek Ruttner
Třída: 3.L
Vedoucí práce: Jiří Švihla
Dne: 30. 4. 2024

Hodnocení:

Vyšší odborná škola a
Střední průmyslová škola elektrotechnická Plzeň,
Koterovská 85

ZADÁNÍ ROČNÍKOVÉ PRÁCE	
Školní rok	2023/ 2024
Studijní obor	78-42-M/01 Technické lyceum
Jméno a příjmení	Marek Ruttner
Třída	3. L
Předmět	Kybernetika
Hodnoceno v předmětu	Kybernetika
Téma	SMDM – Social Media Data Model
Obsah práce	1. Vytvoření aplikace pro shromažďování dat o tom, co ovlivňuje úspěch příspěvků na sociálních sítích. 2. Sběr dat s cílem vytvořit model, který bude předpovídat úspěch nových příspěvků. 3. Kombinace vlastního datového modelu a existujícího jazykového modelu s cílem poradit uživateli, jak upravit specifické části příspěvku a zvýšit jeho úspěšnost. 4. Poskytování konkrétních tipů uživateli, jak optimalizovat obsah příspěvku na základě analýzy datových faktorů. 5. Analýza úspěšnosti datového modelu
Zadávající učitel Příjmení, jméno	Švihla Jiří
Podpis zadávajícího učitele	
Termín odevzdání	30. dubna 2024

V Plzni dne: 30. 11. 2023

Mgr. Vlastimil Volák
ředitel školy

Anotace

Tato ročníková práce se zaměřuje na zhodnocení a vzájemné porovnání metod, které by moli být použity v rámci vývoje aplikace určené k shromažďování a analýze dat ze sociálních sítí X̄(dříve Twitter) či Threads, s cílem identifikovat faktory, které ovlivňují úspěšnost jednotlivých příspěvků. Klíčovým cílem je vytvoření datasetu, pomocí nějž bude možné provádět a testovat jednotlivé metody analýzi, které by měla dokázat předpovědět, jak si daný příspěvek povede na síti a objevit možnosti, jak příspěvek upravit pro zvýšení výkonu. Dalším cílem je ověření možnosti zapojení modelu do již existujících velkých jazykových modelů (LLM), pro automatizaci úprav textu příspěvku a podávání konkrétních rad, které povedou ke zvýšení počtu impresí ("lajků").

Prohlášení

„Prohlašuji, že jsem tuto ročníkovou práci vypracoval samostatně a použil literárních pramenů a informací, které cituji a uvádím v seznamu použité literatury a zdrojů informací.“

V Plzni dne: Podpis:

Obsah

Úvod

V době digitální komunikace se sociální sítě staly klíčovým prostředkem pro sdílení informací, názorů a propagaci různých obsahů. Úspěch příspěvků na těchto platformách, zejména na sociálních sítích X (dříve Twitter) a Threads, je nejen indikátorem popularity, ale také významným faktorem při měření úspěšnosti marketingových a komunikačních strategií firem či osob. Tato ročníková práce se zaměřuje na vývoj a implementaci aplikace určené pro sběr a analýzu dat z těchto sociálních sítí, s cílem identifikovat klíčové faktory, které ovlivňují úspěch a viralitu příspěvků.

Cílem práce je prozkoumat možnosti pro předpovídání počtu interakcí s příspěvkem na základě obsahu příspěvku. K zjištění nejoptimálnějšího postupu, který by bylo možné na tu to problematiku využít, bude práce porovnávat několik způsobů. Od ruční datové analýzy až po využití velkých jazykových modelů (LLM), které se v dnešní době již objevují i na poli datové analýzy.

Kritérii pro porovnávání různých přístupů bude přesnost předpověděné hodnoty, rychlost potřebná k vytvoření predikce a možnost zapojení metody do dalších aplikací a případné využití dat k optimalizaci samotného modelu.

V rámci ročníkové práce se taktéž popisuje způsob, kterým získat data z platformy X, jenž budou využita ke tvorbě predikcí. To kromě extrakce dat ze sociální sítě X zahrnuje taktéž preprocessing dat a jejich strukturování,

Cílem práce je vyzkoušení a porovnání různých přístupů, které by mohli být použity pro aplikaci v reálném použití.

Pořadí	Téma
1.	Politika
2.	Zprávy
3.	Lifestyle
4.	Finance
5.	Kultura

Tabulka 1: Výsledky průzkumu nejsledovanějších témat na X

1 Sběr dat

Pro vytvoření statistiky či modelu, který bude schopný predikce počtu interakcí, je nutné sebrat dostatečné množství dat a tím vytvořit strukturovaná data, která jsou využitelná jak ve statistice, při použití vlastní neuronové sítě, tak pro dotrénováni (fine-tunning) LLM.

1.1 Výběr dat

Pro potřeby sběru dat byl na platformě X vytvořen zcela nový uživatelský účet, který není zatížen předchozí digitální stopou. Účet sleduje pouze účty, které tématicky odpovídají účtům, které označili osoby vyplňující dotazník, za nejvíce sledované a schopné nejvíce udržet uživatele u čtení příspěvků.

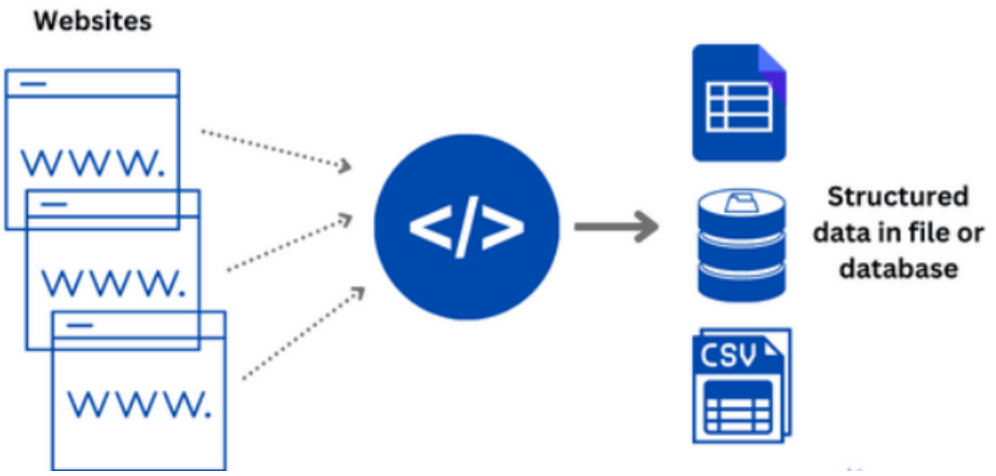
Dotazník vyplnilo celkem 40 osob, přičemž po očištění výsledků průzkumu se jako nejsledovanější témata objevila politika, finance a lifestyle. Výsledky celého průzkumu naleznete v Tabulce 1.

1.2 Sběr dat z X

1.2.1 Teorie

Sběr dat probíhal z webové aplikace platformy X pomocí tzv. data resp. web scraping. Tato metoda umožňuje získat data i z programu, který sám o sobě neumožňuje export dat mimo program.

Web scraping využívá upraveného webového prohlížeče, který se dá spustit pomocí scriptu a automatizovaně se ovládá pomocí předem nastavených parametrů. Ilustrace fungování web scrapingu viz Obrázek 1.



Obrázek 1: Ilustrace web scrapingu

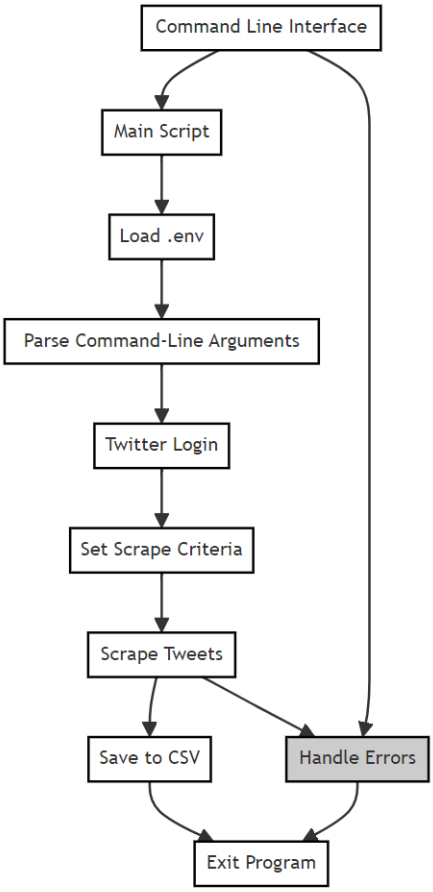
1.2.2 Implementace

V rámci ročníkové práce byl web scraping implementován pomocí aplikace napsané v jazyce Python s použitím knihovny Selenium. Vyhledávání na stránce je řešeno vyhledáváním dle XPath elementů, které jsou poté scrapovány do CSV souboru. Dále jsou podrobněji popsány jednotlivé části programu a na Obrázku 2 je rozkresleno, jak postupně probíhají jednotlivé funkce programu.

Skript `__main__.py` funguje jako vstupní bod, kde je využita knihovna *argparse* pro parsování argumentů z příkazové řádky a bezpečné spravování uživatelských přihlašovacích údajů, ať už skrze proměnné prostředí nebo interaktivní výzvy. Pro citlivé informace, jako jsou přihlašovací údaje pro přihlášení na X, je dynamicky načítána konfigurace *.env*, což zajišťuje bezpečné zacházení s uživatelskými daty. Skript inicializuje a koordinuje proces sběru dat, s elegancí zvládá scénáře s chybami a umožňuje bezproblémové ukončení po dokončení či přerušení.

Jádro funkcionality sběru dat je soustředěno do třídy Twitter_Scraper v souboru twitter_scraper.py. Tato třída využívá Selenium WebDriver pro interakci s webovým rozhraním X, automatizuje procesy přihlašování a naviguje do relevantních částí webu podle specifikovaných kritérií sběru. Extrahuje data z příspěvků, zahrnující informace o uživatelích, obsah, metriky o dosahu, interakce s příspěvkem a další, zatímco efektivně řeší dynamické načítání obsahu skrze kontrolované scrollování. Třída je navržena s důrazem na odolnost, aby čelila běžným výzvám web scrapingu, jako jsou problémy s načítáním prvků a neočekávané změny ve struktuře stránky.

Doplňkové komponenty, jako je třída Scroller, se zabývají správou scrollovacího chování, nezbyt-



Obrázek 2: Diagram ilustrující průběh programu

ného pro dynamické načítání dodatečných tweetů, které jsou na rozhraní Twitteru prezentovány. Třída Tweet se specializuje na parsování a extrakci podrobných informací z jednotlivých tweetů, včetně zpracování speciálního obsahu, jako jsou emodži, a získávání dalších podrobností o autorovi tweetu, je-li to potřebné.

Třída Progress nabízí vizuální reprezentaci průběhu sběru dat, čímž zlepšuje uživatelský zážitek poskytováním okamžité zpětné vazby o stavu operace. Tato funkce je zvláště cenná při dlouhodobých sběrech dat, poskytuje jasný přehled o průběhu a zbývající práci.

Celý program na extrakci dat z platformy X je napsán s ohledem na tvorbu předem strukturovaných dat ve formátu CSV, který lze následně zpracovávat pomocí funkcí v rámci scriptů použitých v postupech dále v ročníkové práci. Taktéž tento způsob obchází nutnost použití Twitter API, které je momentálně nedostupné a tudíž je toto jediná možnost, jak získat aktuální data.

1.3 Preprocessing

Přesto, že data o příspěvcích vyexportovaná jsou v CSV souboru základním způsobem strukturovaná, je nutné provést několik úprav, před prováděním samotných analýz. Jakožto nejzásadnější bylo změnění jazyka, tedy přeložení všech textů příspěvků do angličtiny, což ve výsledku usnadní práci s textem pomocí složitějších metod, kde nebude nutné řešit znaky české abecedy. Dalším krokem je odstranění sloupců, které nejsou používané při provádění analýzy, pokud nebude uvedeno jinak.

2 Ruční statistika

3 Statistika dle výskytu nejpoužívanějších slov

4 Vektorizace slov a použití lineární regrese

5 Vlastní neuronová síť

6 Analýza pomocí LLM

7 Závěr a zhodnocení