EKONOMICKÁ UNIVERZITA V BRATISLAVE
FAKULTA HOSPODÁRSKEJ INFORMATIKY

Evidenčné číslo: 103004/B/2025/36146835373756676

NÁVRH SYSTÉMU NA NAČÍTANIE ÚDAJOV Z INTERNETOVÝCH STRÁNOK – VYBRANÉ KOMODITY A KURZY

Bakalárska práca

2025 **Adam, Boška**

EKONOMICKÁ UNIVERZITA V BRATISLAVE
FAKULTA HOSPODÁRSKEJ INFORMATIKY

NÁVRH SYSTÉMU NA NAČÍTANIE ÚDAJOV Z INTERNETOVÝCH STRÁNOK – VYBRANÉ KOMODITY A KURZY

Bakalárska práca

Študijný program: hospodárska informatika

Študijný odbor: ekonómia a manažment

Školiace pracovisko: Katedra aplikovanej informatiky

Vedúci záverečnej práce: Jaroslav, Kultan doc., Ing., PhD.

Bratislava 2025 **Adam, Boška**

ABSTRAKT

BOŠKA, Adam: *Návrh systému na načítavanie údajov z internetových stránok - vybrané komodity a kurzy*. – Ekonomická univerzita v Bratislave. Fakulta hospodárskej informatiky; Katedra aplikovanej informatiky. – doc. Ing. Jaroslav Kultan PhD. – Bratislava: FHI, 2025, 51 s.

Bakalárskej práce je vypracovaná na tému Návrh systému na načítavanie údajov z internetových stránok - vybrané komodity a kurzy. Cieľom záverečnej práce bolo vytvoriť systém na sťahovanie a prepisovanie údajov z PDF letákov jednotlivých obchodných reťazcov do štruktúrovanej formy. Jednotlivé časti záverečnej práce boli zamerané na súčasný stav spracovaných údajov z letákov obchodných reťazcov, stanovenie cieľov a následne popísanie metód možných riešení. Výsledkom riešenia danej problematiky je systém na automatizované sťahovanie PDF letákov a na nasledovné prepísanie údajov z neštruktúrovanej podoby do štruktúrovanej vo formáte CSV, ktorá je prezentovaná na stránke bpab.wz.cz.

Kľúčové slová: OCR, web scraping, Python, PDF, obchodný reťazec, leták

ABSTRACT

BOŠKA, Adam: Design of a System for Retrieving Data from Websites – Selected Commodities and Exchange Rates. – University of Economics in Bratislava. Faculty of Economic Informatics; Department of Applied Informatics. – Supervisor: doc. Ing. Jaroslav Kultan, PhD. – Bratislava: FHI, 2025, 51 p.

The bachelor's thesis was developed on the topic of designing a system for retrieving data from websites—selected commodities and exchange rates. The aim of the thesis was to create a system for downloading and transcribing data from PDF flyers of various retail chains into a structured format. Individual parts of the thesis focused on the current state of processed data from retail chain flyers, defining objectives, and subsequently describing methods for possible solutions. The result of addressing this issue is a system for automated downloading of PDF flyers and the subsequent transcription of data from an unstructured format into a structured CSV format, which is presented on the website bpab.wz.cz.

Keywords: OCR, web scraping, Python, PDF, retail chain, flyer

Obsah

Zoznam použitých skratiek – slovenský preklad

OCR – Optical Character Recognition - Optické rozpoznávanie znakov

API – Application Programming Interface - Aplikačné programovacie rozhranie

IP – Internet Protocol - Internetový protokol

PDF – Portable Document Format - Prenosný formát dokumentu

AI – Artificial Intelligence - Umelá inteligencia

HTML – HyperText Markup Language - Hypertextový značkovací jazyk

URL – Uniform Resource Locator - Jednotný lokátor zdrojov

DOM – Document Object Model - Objektový model dokumentu

CSV – Comma-Separated Values - Hodnoty oddelené čiarkami

Zoznam obrázkov

Úvod

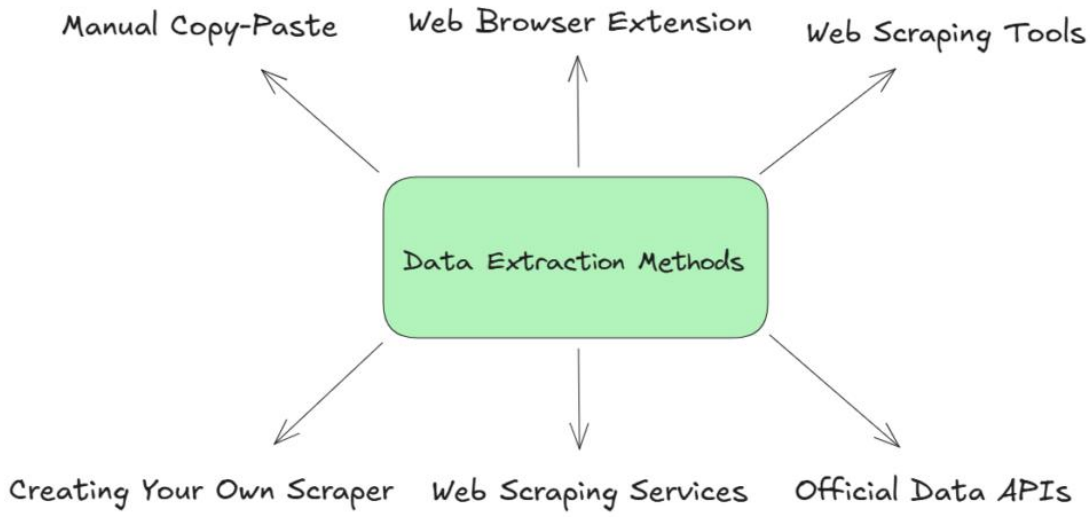
V oblasti obchodu a analýzy trhu je efektívne zhromažďovanie údajov o cenách a vlastnostiach komodít kľúčové. Obchodné reťazce, ako napríklad Billa, Coop Jednota, Kaufland či Lidl, pravidelne zverejňujú svoje akciové ponuky vo forme PDF letákov na svojich webových stránkach. Tieto letáky predstavujú bohatý zdroj údajov, obsahujúci detailné informácie o produktoch, vrátane názvov komodít, ich cien, hmotností, jednotkových cien, zliav a prípadne ďalších špecifikácií. Tieto údaje sú neoceniteľné pre rôzne účely, ako je analýza trhu, porovnávanie cien medzi reťazcami, identifikácia najvýhodnejších ponúk alebo sledovanie cenových trendov v čase. Takéto informácie umožňujú podnikom, analytikom aj spotrebiteľom lepšie pochopiť dynamiku trhu a prijímať informované rozhodnutia.

Napriek vysokej hodnote týchto údajov je ich manuálne spracovanie mimoriadne časovo náročné a náchylné na chyby. Proces zahŕňa ručné prehľadávanie webových stránok jednotlivých reťazcov, vyhľadávanie a sťahovanie aktuálnych PDF letákov, ich čítanie a prepisovanie relevantných údajov do použiteľnej podoby. Tento postup si vyžaduje značné úsilie, precíznosť a pozornosť k detailom, aby sa predišlo chybám, ako je nesprávne zaznamenanie ceny, hmotnosti alebo iných špecifikácií. Takéto chyby môžu viesť k nepresným analýzam, skresleným výsledkom a nesprávnym záverom, čo znižuje spoľahlivosť údajov. Navyše, formát PDF letákov nie je jednotný, čo spracovanie ešte viac komplikuje. Každý reťazec používa odlišné dizajny, štruktúry a spôsoby prezentácie údajov, ako sú rôzne typy písma, rozmiestnenie textu či grafické prvky. Tieto rozdiely znamenajú, že manuálne spracovanie musí byť prispôsobené pre každý reťazec zvlášť, čo zvyšuje náročnosť a časovú náročnosť procesu.

Tieto výzvy jasne poukazujú na potrebu automatizovaného systému, ktorý dokáže efektívne a spoľahlivo získavať údaje z PDF letákov a štruktúrovať ich do podoby vhodnej na ďalšie spracovanie.

1 Načítavanie údajov z internetových stránok a letákov

Schopnosť získavať údaje z webových stránok je praktická zručnosť využiteľná na rôzne účely, ako je zhromažďovanie dát, ich analýza či automatizácia neustále rovnakých činností. Internet ponúka obrovské množstvo informácií, a ak dokážete tieto údaje efektívne extrahovať a vyhodnotiť, získate užitočné poznatky podporujúce rozhodovanie na základe faktov. Pre firmy môže byť získavanie dát kľúčové pri finančných operáciách, ako je nákup alebo predaj. Spôsob extrakcie údajov závisí od vašich potrieb a od toho, ako je konkrétna webová stránka štruktúrovaná. (Koolwal, 2024)



Obrázok 1. spôsoby extrakcie dát z internetu Zdroj : (Koolwal, 2024)

Metódy na získavanie dát z web stránky: (Koolwal, 2024)

- **Manuálne kopírovanie-vkladanie** - Jednoduché skopírovanie údajov z webovej stránky a ich vloženie do dokumentu či tabuľky predstavuje jednu z najľahších metód ich získania.
- **Pomocou rozšírení webových prehliadačov** - Rozšírenia je možné nainštalovať do prehliadača, čím získate možnosť označiť a extrahovať vybrané dátové prvky z webovej lokality.
- **Nástroje web scraping** - Tieto nástroje slúžia na navigáciu po webe a umožňujú vám vybrať a stiahnuť určité údaje podľa vašich kritérií.
- **Oficiálne dátové rozhrania API** - Množstvo webových lokalít disponuje API rozhraniami, ktoré umožňujú prístup k ich údajom v prehľadnom formáte. Používanie API na získavanie dát z webu je výhodné, keďže údaje sú už štruktúrované a pripravené na okamžité využitie. Nie všetky stránky však API ponúkajú a tie, ktoré áno, môžu mať pravidlá obmedzujúce ich použitie.

- **Služby web scrapingu** - Pomocou týchto služieb sú zvládnuté technické detaily webového zoškrabovania a údaje vám poskytnú v preferovanom výstupnom formáte.
- **Vytvorenie vlastnej služby web scraping** - Hlavným obmedzením tejto techniky je blokovanie IP adresy. Pri veľkom zoškrabovaní vás hostiteľ okamžite zastaví blokáciou IP, no pre malé projekty je táto možnosť lacnejšia a jednoduchšie zvládnuteľná.

1.1 Súčasný stav v zahraničí

Súčasný stav zahraničných platforiem pre zhromažďovanie informácií z akciových letákov a zliav je charakteristický rozmanitosťou, technologickou vyspelosťou a globálnym dosahom. Tieto platformy, pôsobiace v desiatkach krajín, od Európy po Áziu, Ameriku a Blízky východ, slúžia miliónom používateľov, ktorí hľadajú výhodné ponuky od veľkých reťazcov, ako sú Walmart, Tesco, Carrefour, Lidl, či lokálnych obchodov. Ich cieľom je uľahčiť plánovanie nákupov, šetriť čas a peniaze a podporovať udržateľnosť prostredníctvom digitálnych letákov.

Platformy ponúkajú intuitívne rozhrania, kde používatelia môžu prehliadať letáky, kupóny a zľavy, triedené podľa kategórií, ako sú potraviny, elektronika, móda alebo domácnosť, prípadne podľa geografickej polohy. Poskytujú doplnkové informácie, vrátane otváracích hodín, adries predajní a vernostných programov, ako sú napríklad Tesco Clubcard a Walmart Rewards. Mobilné aplikácie zvyšujú dostupnosť, umožňujú vytváranie nákupných zoznamov a notifikácie o nových akciách. Tieto funkcie sú navrhnuté tak, aby maximalizovali pohodlie a prispôsobili sa individuálnym potrebám používateľov.

Platformy čelia konkurencii zo strany oficiálnych webových reťazcov, ktoré ponúkajú vlastné letáky, a musia sa vyrovnať s rôznymi regionálnymi preferenciami nakupujúcich. Napriek tomu sa darí prilákať široké publikum dôrazom na jednoduchosť a personalizáciu, keďže ponuky sú prispôsobené záujmom používateľov. Podporujú aj obchodníkov, ktorým pomáhajú zvyšovať návštevnosť predajní. Tieto služby sa stali neoddeliteľnou súčasťou nakupovania, ponúkajúc efektívny spôsob, ako sledovať zľavy a plánovať výhodné nákupy na globálnej úrovni.

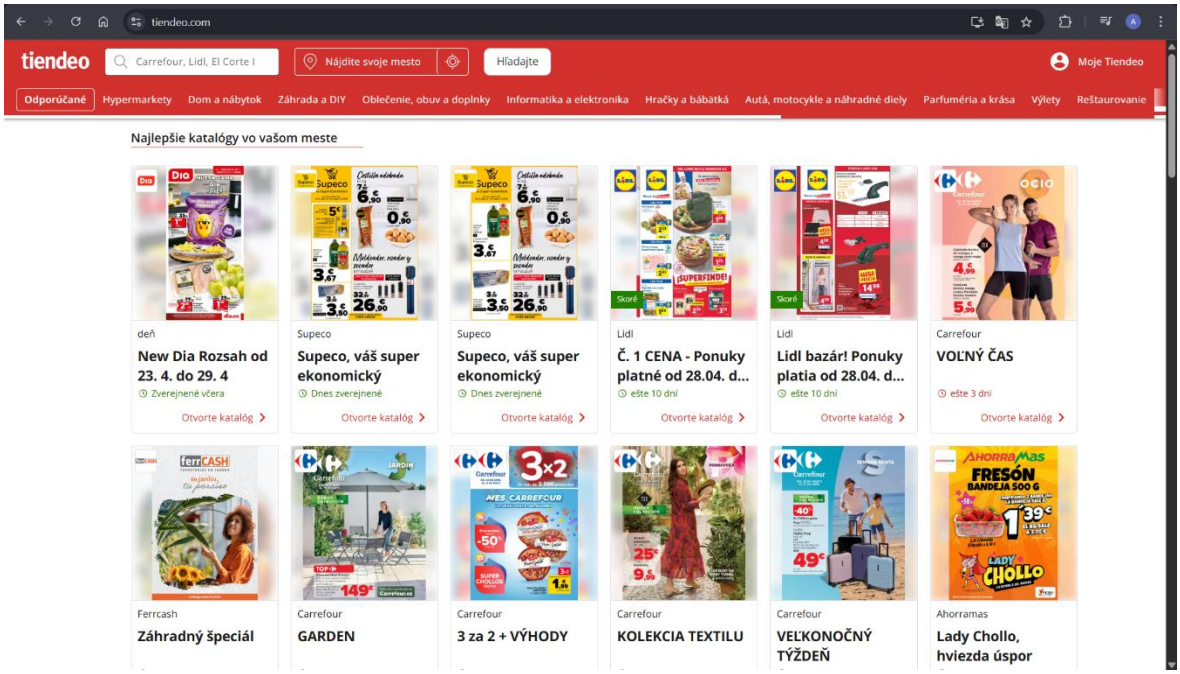
1.1.1 Tiendeo.com

Tiendeo.com je globálna online platforma, ktorá slúži ako centrálny zdroj pre akciové letáky, zľavy a kupóny od širokého spektra obchodných reťazcov a značiek, ako sú Lidl, Tesco, IKEA, Carrefour, Zara či Walmart. Pôsobí vo viac ako 40 krajinách na piatich kontinentoch, vrátane Európy, Latinskej Ameriky, Ázie a Afriky. Tiendeo slúži ako kľúčový

nástroj pre zákazníkov, ktorí chcú šetriť pri nakupovaní, a pre obchodníkov, ktorí hľadajú efektívny spôsob, ako osloviť zákazníkov.

Platforma ponúka jednoduché a prehľadné rozhranie, kde používatelia môžu objavovať aktuálne ponuky, prehliadať digitálne letáky a vyhľadávať zľavy podľa kategórií, ako sú potraviny, oblečenie, elektronika alebo nábytok, prípadne podľa konkrétnych obchodov či lokalít. Tiendeo poskytuje aj doplnkové informácie, napríklad otváracie hodiny predajní, adresy alebo detaily o vernostných programoch, čo pomáha pri plánovaní nákupov. Používatelia majú možnosť vytvárať nákupné zoznamy, ukladať kupóny a dostávať upozornenia na nové akcie.

Tiendeo je dostupné cez web aj mobilnú aplikáciu, ktorá umožňuje pohodlný prístup k obsahu na cestách, vrátane funkcií ako vyhľadávanie obchodov v okolí pomocou geologickej lokácie.



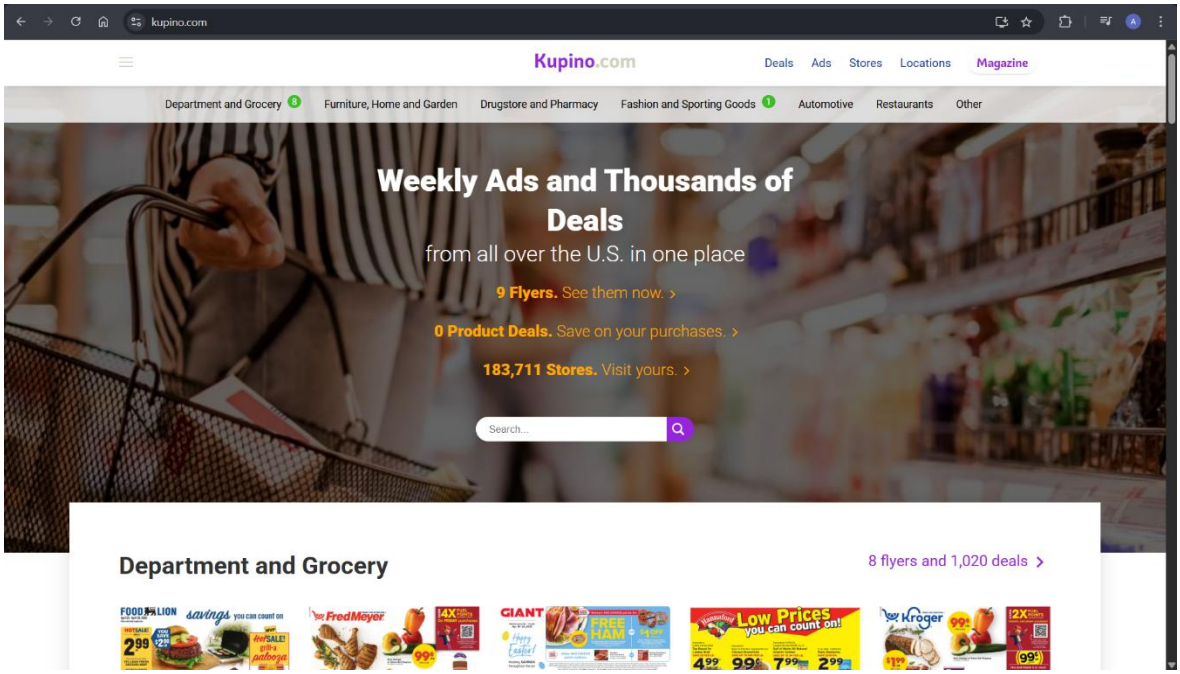
Obrázok 2. webová stránka Tiendeo.com – úvod Zdroj : (Vlastný zdroj)

1.1.2 Kupino.com

Kupino.com je medzinárodná online platforma zameraná na zhromažďovanie a prezentáciu akciových letákov, zliav a kupónov od širokej škály obchodných reťazcov, ako sú Walmart, Aldi, Tesco, Lidl, Target či IKEA. Pôsobí vo viacerých krajinách, vrátane USA, Kanady, Austrálie, Českej republiky a Slovenska, a slúži miliónom používateľov, ktorí hľadajú výhodné ponuky. Platforma, ktorá začala ako lokálny projekt, sa rozrástla na globálny nástroj, ktorý pomáha zákazníkom šetriť peniaze a čas pri plánovaní nákupov.

Kupino.com poskytuje prehľadné a intuitívne rozhranie, kde používatelia môžu prehliadať dostupné letáky, triedené podľa kategórií, ako sú potraviny, elektronika, oblečenie, domácnosť alebo záhradné potreby. Okrem letákov ponúka informácie o otváracích hodinách predajní, adresách a vernostných programoch, ako sú Walmart Rewards a Tesco Clubcard. Tieto funkcie uľahčujú plánovanie nákupov a maximalizáciu úspor.

Používatelia môžu filtrovať ponuky podľa lokality, konkrétneho obchodu alebo produktu a vytvárať vlastné nákupné zoznamy. Web aj mobilná aplikácia, dostupná na iOS a Android, umožňujú rýchly prístup k obsahu a využívajú geolokáciu na vyhľadávanie najbližších predajní.



Obrázok 3. webová stránka Kupino.com – úvod Zdroj : (Vlastný zdroj)

1.2 Súčasný stav na Slovensku

V súčasnosti je možné nájsť viaceré slovenské zdroje a platformy, ako Kimbino.sk, KompasZliav.sk, Letakomat.sk či AkcneLetaky.sk, ktoré sa zaoberajú sprístupňovaním informácií z online letákov obchodných reťazcov, vrátane Lidl, Billa, Kaufland, COOP Jednota, ale aj z iných odvetví, ako drogérie či elektronika. Tieto platformy sú primárne navrhnuté pre spotrebiteľov, ktorí hľadajú jednoduchý a rýchly spôsob, ako získať prehľad o aktuálnych akciových ponukách, zľavách a cenách rôznych produktov. Ich hlavným účelom je uľahčiť používateľom plánovanie nákupov, porovnávanie ponúk medzi reťazcami a efektívne využívanie zliav na každodenné potreby, ako sú potraviny, kozmetika či domáce potreby. Tieto služby šetria čas a peniaze, umožňujú prehľadné sledovanie akcií a podporujú informované rozhodovanie pri nákupoch.

Platformy sa však líšia vo svojej technickej vyspelosti, funkčnosti a schopnosti poskytovať dáta v podobe využiteľnej mimo spotrebiteľského kontextu. Väčšina z nich sa zameriava na vizuálnu prezentáciu informácií, často vo forme PDF dokumentov alebo obrázkov letákov. Tento obsah je síce prehľadný, no zostáva uzavretý v grafickej forme, bez hlbšieho spracovania, ako je štruktúrované extrahovanie dát o cenách, produktoch či zľavách. Niektoré platformy ponúkajú doplnkové funkcie, napríklad vyhľadávanie podľa kategórií, filtrovanie podľa lokalít alebo informácie o otváracích hodinách a vernostných programoch, ako je BILLA Bonus karta, Lidl Plus karta a Kaufland Card , no pokročilé analytické nástroje absentujú.

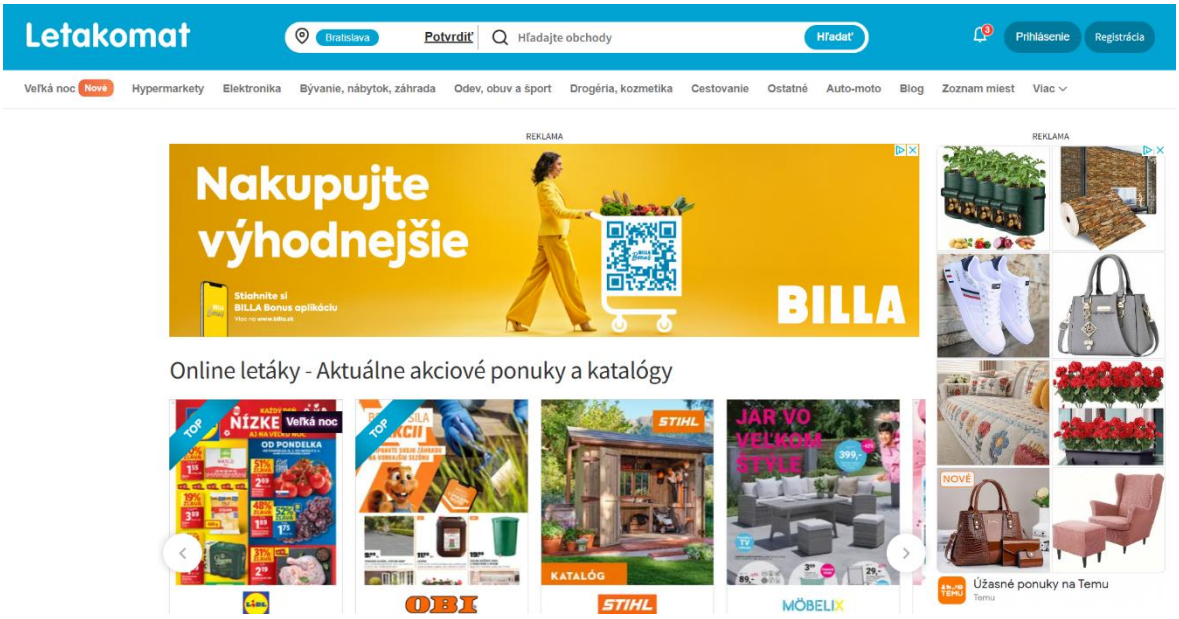
Tento stav odráža širší vývoj digitalizácie na slovenskom trhu, kde sa v maloobchode kladie väčší dôraz na jednoduchosť a dostupnosť pre koncových používateľov než na technickú infraštruktúru podporujúcu automatizáciu či pokročilé analýzy. Platformy zvyčajne fungujú ako sprostredkovatelia, ktorí zhromažďujú letáky priamo od reťazcov alebo ich webov, no nevyužívajú moderné technológie, ako sú automatické extrakčné nástroje, strojové učenie či programové rozhrania API, ktoré by umožnili hlbšiu integráciu dát do iných systémov. Napriek tomu tieto platformy efektívne pokrývajú potreby bežných spotrebiteľov, no ich potenciál pre podnikateľské či analytické využitie zostáva obmedzený, čo naznačuje priestor pre budúci rozvoj a inováciu v oblasti spracovania dát z letákov.

1.2.1 Letákomat.sk

Letákomat.sk patrí medzi najznámejšie platformy na Slovensku, ktoré sa špecializujú na zozbieravanie letákov od rôznych obchodných reťazcov, vrátane Lidl, Billa, Kaufland a COOP Jednota. Táto stránka slúži ako centralizovaný archív, kde sú letáky dostupné vo formáte PDF, pričom väčšina obsahu je preberaná priamo z oficiálnych webových stránok jednotlivých reťazcov. Používateľom ponúka intuitívne rozhranie, v rámci ktorého si môžu vybrať konkrétny obchodný reťazec, prezerať aktuálne aj budúce letáky a filtrovať ich podľa dátumu platnosti alebo regiónu, čo reflektuje rôznorodosť ponúk na lokálnej úrovni. Tento prístup z nej robí obľúbený nástroj pre spotrebiteľov, ktorí hľadajú jednoduchý spôsob, ako získať prehľad o akciách bez potreby navštevovať viaceré webové stránky.

Napriek svojej popularite však Letákomat.sk neposkytuje žiadnu formu automatizovaného spracovania údajov. Informácie o cenách, produktoch či zľavách zostávajú uzamknuté v grafickom formáte PDF dokumentov, čo obmedzuje ich využitie na manuálne prezeranie. Platforma neumožňuje vyhľadávanie konkrétnych komodít naprieč letákmi, porovnávanie cien jednotlivých produktov ani export dát do štruktúrovanej podoby.

Z technického hľadiska ide skôr o pasívny nástroj na distribúciu letákov než o aktívny systém schopný extrakcie a analýzy údajov.



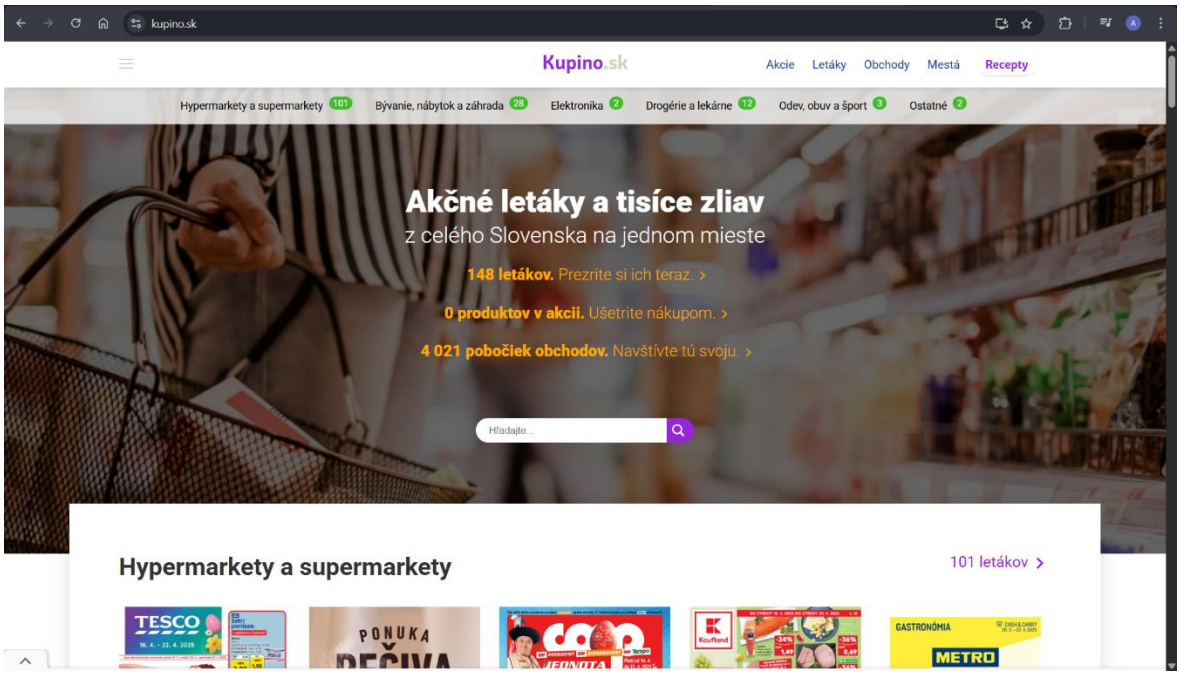
Obrázok 4. webová stránka Letakomat.sk – úvod Zdroj : (Vlastný zdroj)

1.2.2 Kupino.sk

Kupino.sk je populárna slovenská platforma zameraná na poskytovanie aktuálnych letákov a akciových ponúk od obchodných reťazcov ako Lidl, Billa, COOP Jednota, Kaufland a ďalších. Webová stránka a mobilná aplikácia ponúkajú používateľom prehľadný prístup k najnovším zľavám, čím šetria čas aj peniaze pri plánovaní nákupov. Kupino.sk sa vyznačuje jednoduchým a intuitívnym rozhraním, ktoré umožňuje rýchle prehliadanie letákov, vyhľadávanie akcií a lokalizáciu predajní.

Hlavnou funkciou platformy je zhromažďovanie a pravidelná aktualizácia letákov, ktoré pokrývajú aktuálne, budúce aj predchádzajúce akcie. Používatelia si môžu prezerať letáky v digitálnej podobe, filtrovať ich podľa obchodu alebo kategórie (napr. potraviny, elektronika) a získavať detailné informácie o produktoch vrátane cien a dátumov platnosti. Kupino.sk navyše poskytuje adresy a otváracie hodiny predajní, čo uľahčuje plánovanie nákupov v konkrétnej lokalite.

Technologicky platforma využíva web scraping a manuálne spracovanie na zhromažďovanie údajov z oficiálnych webov reťazcov, čím zabezpečuje aktuálnosť informácií. Kupino.sk je bezplatná a nevyžaduje registráciu, čo zvyšuje jej dostupnosť. Okrem letákov ponúka aj tipy na šetrenie a inšpirácie pre domácnosť.



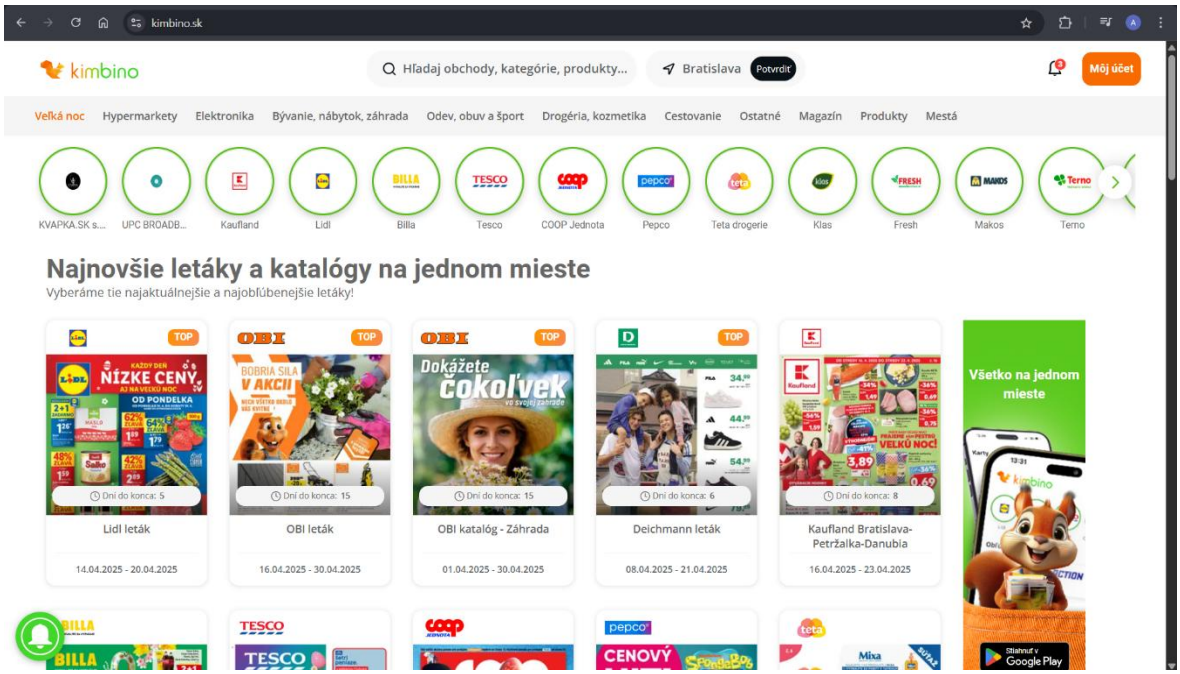
Obrázok 5. webová stránka Kupino.sk – úvod Zdroj : (Vlastný zdroj)

1.2.3 Kimbino.sk

Kimbino.sk je popredný slovenský webový portál zameraný na zhromažďovanie, spracovanie a prezentáciu akciových letákov a zliav od širokej škály obchodných reťazcov, vrátane Lidl, Billa, Kaufland, COOP Jednota, Tesco, Terno a ďalších. Je materskou stránkou platformy Kimbino.com. Stránka poskytuje používateľsky prívetivé rozhranie, ktoré umožňuje rýchle prehliadanie aktuálnych, pripravovaných aj archívnych letákov. Letáky sú prehľadne rozdelené podľa kategórií, ako sú potraviny, nápoje, drogéria, elektronika či oblečenie, a podľa lokalít, čo uľahčuje vyhľadávanie ponúk v konkrétnom regióne alebo meste. Okrem letákov ponúka Kimbino.sk praktické informácie o otváracích hodinách predajní, adresách a kontaktoch. Používatelia môžu vyhľadávať zľavy na konkrétne produkty, ako sú mliečne výrobky alebo elektronika, a porovnávať ponuky medzi rôznymi reťazcami.

Medzi kľúčové výhody patrí široké pokrytie obchodných reťazcov, intuitívne ovládanie a dostupnosť mobilnej aplikácie Kimbino, ktorá umožňuje okamžitý prístup k letákom a akciám kdekoľvek. Web zároveň ponúka tipy na nákupy, sezónne akcie a inšpirácie, napríklad odporúčania na vianočné darčeky alebo letné grilovanie.

Kimbino.sk je ideálnym nástrojom pre zákazníkov, ktorí chcú šetriť čas a peniaze pri nakupovaní.

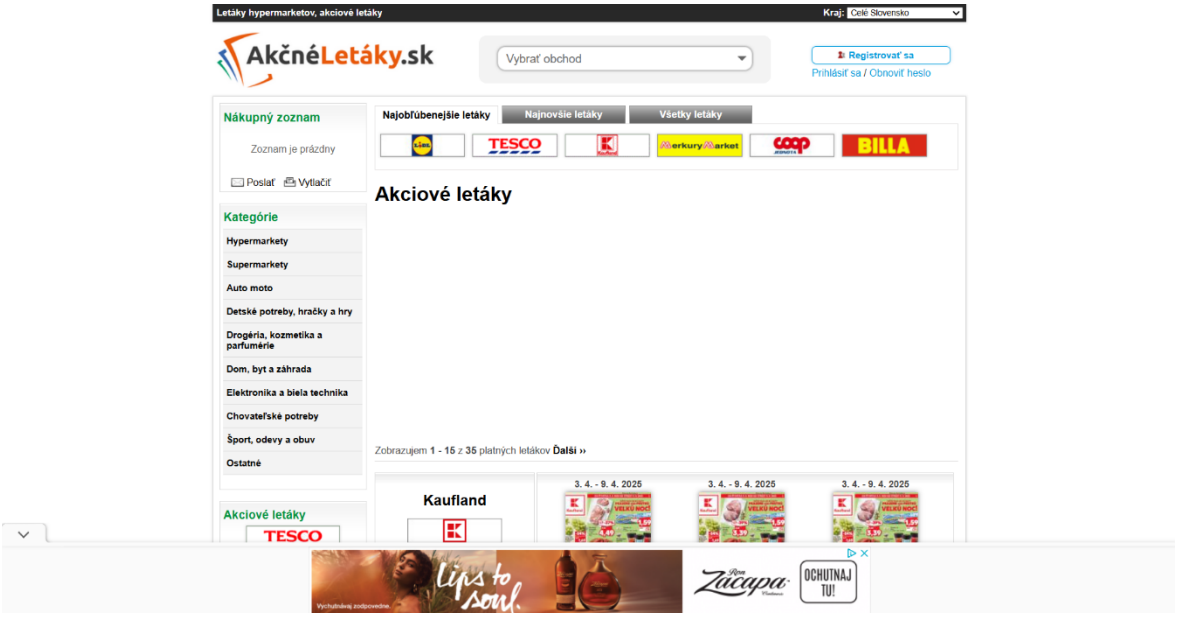


Obrázok 6. webová stránka Kimbino.sk – úvod Zdroj : (Vlastný zdroj)

1.2.4 AkčnéLetáky.sk

Ďalšou rozšírenou platformou na Slovensku je AkčnéLetáky.sk, ktorá si kladie za cieľ podobne ako Letákomat.sk sprístupniť spotrebiteľom prehľad aktuálnych akcií a zliav od rôznych obchodných reťazcov. Stránka pokrýva široké spektrum reťazcov, vrátane Lidl, Billa, Kaufland a COOP Jednota, čím sa snaží uspokojiť potreby širokého okruhu používateľov. Používatelia majú možnosť triediť letáky podľa kategórií, ako sú potraviny, elektronika, drogéria či iné segmenty, čo uľahčuje orientáciu v ponukách a zameranie sa na konkrétne typy produktov. Letáky sú dostupné vo formáte PDF, ktorý je štandardným spôsobom distribúcie, no platforma ide o krok ďalej tým, že pridáva základné informácie o časovej platnosti akcií. V niektorých prípadoch dokonca zvýrazňuje vybrané ponuky, čím upriamuje pozornosť na atraktívne zľavy alebo sezónne produkty.

Napriek týmto vylepšeniam nie je zrejmé, že by AkčnéLetáky.sk využívalo pokročilé technológie na automatické spracovanie obsahu letákov. Informácie zostávajú uväznené v grafickej podobe PDF dokumentov, čo znamená, že akékoľvek ďalšie využitie dát si vyžaduje manuálnu prácu. Pre bežných spotrebiteľov je výhodou jednoduché a prehľadné rozhranie, ktoré umožňuje rýchly prístup k letákom bez zbytočných komplikácií.



Obrázok 7. webová stránka AkčnéLetáky.sk – úvod Zdroj : (Vlastný zdroj)

Popularita portálov s letákmi obchodných reťazcov, ako sú Tiendeo.com, Kupino.com, Letakomat.sk, Kupino.sk, Kimbino.sk či AkčnéLetáky.sk, výrazne rastie vďaka digitálnej transformácii. Tieto platformy umožňujú spotrebiteľom jednoduchý prístup k aktuálnym akciám a zľavám, čím zefektívňujú plánovanie nákupov. Digitalizácia zmenila spôsob, akým ľudia pristupujú k informáciám o cenách, pričom online dostupnosť letákov nahrádza tradičné tlačené formáty. Mobilné aplikácie a webové rozhrania týchto portálov ponúkajú intuitívne funkcie, ako sú vyhľadávanie podľa kategórií, filtrovanie podľa obchodov alebo notifikácie o nových ponukách, čo zvyšuje ich atraktivitu.

Napriek popularite čelia tieto platformy viacerým výzvam. Presnosť a aktuálnosť údajov sú kľúčové, pretože chybné informácie môžu viesť k strate dôvery používateľov. Tieto portály významne prispievajú k transparentnosti cien a efektívnemu nakupovaniu, čím podporujú informované rozhodnutia spotrebiteľov.

2 Cieľ práce

Hlavným cieľom tejto bakalárskej práce je navrhnúť a implementovať automatizovaný systém na efektívne získavanie údajov o vybraných komoditách, ich cenách a špecifikáciách z internetových stránok vybraných obchodných reťazcov, konkrétne Billa, Coop Jednota, Kaufland a Lidl. Tento systém bude zameraný na automatické sťahovanie PDF letákov, ktoré tieto reťazce pravidelne zverejňujú na svojich webových stránkach. Následne systém extrahuje relevantné údaje, ako sú názvy komodít, ich ceny, hmotnosť, jednotkové ceny, prípadne ďalšie špecifikácie uvedené v letákoch, a transformuje ich do štruktúrovanej podoby vhodnej na ďalšie spracovanie.

Prvým krokom systému je vytvorenie mechanizmu na automatické vyhľadávanie a sťahovanie aktuálnych PDF letákov z webových stránok uvedených reťazcov. Tento proces zahŕňa analýzu štruktúry webových stránok a implementáciu robustného programu, ktorý dokáže identifikovať odkazy na letáky aj v prípade zmien v dizajne alebo štruktúre stránok.

Druhý krok systému je za pomoci niekoľkých umelých inteligencií, prepísať potrebné údaje zo stiahnutých PDF letákov do textového súboru.

V tretom kroku systém využije techniky spracovania PDF dokumentov na extrakciu textových údajov. Tieto údaje budú najprv uložené v textovom súbore ako surové dáta, čím sa zabezpečí ich základná archivácia a dostupnosť pre ďalšie spracovanie.

Posledným krokom je transformácia surových dát do štruktúrovanej podoby vo formáte CSV a následná prezentácia týchto dát na web stránke. Systém bude navrhnutý tak, aby dokázal spracovať rôzne formáty letákov a prispôsobiť sa prípadným rozdielom v štruktúre údajov medzi jednotlivými reťazcami. Dôraz bude kladený na presnosť extrahovaných údajov a minimalizáciu chýb, napríklad pri rozpoznávaní cien alebo špecifikácií produktov.

3 Nástroje a metódy riešenia

Na tvorbu systému na načítavanie údajov z internetových stránok pre vybrané obchodné reťazce je potrebné preskúmať a vykonať niekoľko postupných činností, ktoré zahŕňajú analýzu požiadaviek, návrh riešenia a jeho praktickú implementáciu. V kapitole sú detailne popísané viaceré nástroje a technológie, ktoré môžu významne napomôcť pri jednotlivých krokoch procesu, aby bolo možné dosiahnuť požadovaný výsledok efektívne a spoľahlivo. Tieto nástroje pokrývajú široké spektrum úloh, od získavania dát až po ich spracovanie a uloženie.

V konkrétnej časti práce sa rozoberajú systémy a techniky, ktoré slúžia na získavanie cesty k požadovaným údajom na určitej webovej stránke, vrátane metód, ako identifikovať a extrahovať relevantné informácie z jej štruktúry. Okrem toho sa opisujú aj nástroje na automatické sťahovanie súborov, ktoré umožňujú efektívne získavanie dát bez manuálneho zásahu. Ďalej sa venuje pozornosť procesu načítavania údajov z jedného súboru, ich spracovaniu a následnému prepisu do novovytvoreného textového súboru, pričom sa kladie dôraz na presnosť a použiteľnosť výsledných dát.

3.1 Umelá inteligencia

Umelá inteligencia, často označovaná skratkou AI, je oblasť informatiky, ktorá sa zaoberá vývojom systémov schopných vykonávať úlohy, ktoré by za normálnych okolností vyžadovali ľudskú inteligenciu. Patrí sem napríklad schopnosť učiť sa z dát (strojové učenie), rozpoznávať obraz a reč, porozumieť prirodzenému jazyku, riešiť problémy, rozhodovať sa alebo komunikovať v zmysluplnej forme. AI modely napodobňujú niektoré kognitívne funkcie človeka a nachádzajú uplatnenie v rôznych oblastiach – od priemyslu a zdravotníctva cez dopravu a finančný sektor až po každodenné používanie v smartfónoch či domácich asistentoch.

V súčasnosti sú najrozšírenejšie modely tzv. úzkej AI, ktoré sú zamerané na riešenie konkrétnych úloh, ako napríklad automatické rozpoznávanie hlasu, odporúčacie systémy alebo chatboty. Pokročilejšie systémy, ako sú veľké jazykové modely (napr. GPT, Gemini či Claude), majú schopnosť generovať text, odpovedať na otázky alebo pracovať s rôznymi typmi údajov (text, obraz, video). Tieto systémy sa rýchlo rozvíjajú a čoraz viac sa približujú k tomu, čo nazývame všeobecná umelá inteligencia – teda AI so schopnosťou samostatne uvažovať, učiť sa a riešiť rôzne typy problémov naprieč doménami. Napriek obrovskému potenciálu umelá inteligencia prináša aj výzvy a otázky týkajúce sa etiky, bezpečnosti,

zodpovednosti či vplyvu na trh práce a spoločnosť ako celok. Preto je dôležité pristupovať k jej vývoju a využívaniu zodpovedne a premyslene. (OpenAI, 2023)

3.1.1 Gemini

Gemini je pokročilý systém umelej inteligencie vyvinutý spoločnosťou Google DeepMind, ktorý predstavuje novú generáciu takzvaných veľkých jazykových modelov so schopnosťou pracovať s rôznymi formátmi ako obrázok, zvuk a iné. Na rozdiel od starších modelov, ktoré boli primárne zamerané na spracovanie textu, Gemini je od základov navrhnutý tak, aby dokázal pracovať s rôznymi typmi vstupov – vrátane textu, obrázkov, zvuku, videa či programovacieho kódu. To z neho robí univerzálny nástroj vhodný pre široké spektrum úloh, od tvorby textov a analýzy obrázkov až po generovanie softvérového kódu či vyhodnocovanie zložitých dátových vstupov.

Gemini je výsledkom prepojenia dvoch významných výskumných tímov – Google Brain a DeepMind – ktorých spoločným cieľom bolo vytvoriť model, ktorý sa nielen vyrovná konkurencii v podobe modelov GPT od OpenAI, ale ju v mnohých aspektoch aj predbehne. Výsledkom tejto spolupráce je model, ktorý sa nielen dokáže učiť z kontextu a generovať koherentné a logické odpovede, ale je aj schopný analyzovať obrazové a zvukové dáta, a tým lepšie pochopiť komplexné situácie či zložité otázky.

Jednou z hlavných výhod Gemini je jeho schopnosť pracovať s mimoriadne veľkým objemom informácií naraz. Najnovšie verzie modelu, ako napríklad Gemini 1.5, dokážu spracovať kontext s rozsahom až jedného milióna tokenov, čo znamená, že môžu naraz analyzovať celé knihy, rozsiahle zdrojové kódy alebo veľké množstvo dokumentov bez toho, aby sa stratila kontinuita informácií. Okrem výkonnosti sa kládol veľký dôraz aj na bezpečnosť, spoľahlivosť a etické zásady, pričom model prechádza rozsiahlymi testami pred jeho nasadením do reálnych aplikácií. Gemini je už integrovaný v rôznych službách spoločnosti Google, ako sú Bard, Workspace alebo Android zariadenia, a postupne sa stáva súčasťou bežného digitálneho prostredia, v ktorom uľahčuje prácu miliónom ľudí po celom svete. (OpenAI, 2023)

3.1.2 Chat GPT

ChatGPT je jazykový model umelej inteligencie vyvinutý spoločnosťou OpenAI, ktorý patrí medzi najznámejšie a najrozšírenejšie systémy tohto druhu na svete. Je založený na architektúre GPT (Generative Pre-trained Transformer), ktorá umožňuje modelu porozumieť prirodzenému jazyku a generovať plynulé, gramaticky správne a obsahovo zmysluplné odpovede na základe vstupného textu. Modely GPT sa trénujú na obrovskom

množstve dát z internetu, kníh, článkov, webových stránok či programového kódu, vďaka čomu majú široké všeobecné vedomosti a rozvinuté jazykové schopnosti.

ChatGPT je navrhnutý na interakciu v štýle konverzácie, čo znamená, že používateľ s ním môže viesť dialóg podobný komunikácii s človekom. Je schopný odpovedať na otázky, vysvetľovať zložité pojmy, generovať texty, prekladať medzi jazykmi, sumarizovať dokumenty, pomáhať pri programovaní alebo dokonca tvoriť kreatívny obsah, ako sú básne, príbehy či marketingové texty. Vďaka týmto schopnostiam sa ChatGPT uplatňuje v mnohých oblastiach, od vzdelávania a výskumu až po zákaznícku podporu, obchod, marketing alebo softvérový vývoj.

Model je dostupný cez webové rozhranie, ako aj cez API, čo umožňuje jeho integráciu do rôznych aplikácií a systémov. Od uvedenia prvej verzie GPT-3 sa technológia neustále vyvíja a zdokonaľuje. Najnovšie verzie, ako GPT-4 a GPT-4 Turbo, majú lepšiu schopnosť porozumieť kontextu, udržať konzistentnú komunikáciu a pracovať s rozsiahlejšími vstupmi. ChatGPT je známy aj tým, že dokáže generovať vysoko kvalitné odpovede vo veľmi krátkom čase, pričom sa snaží byť nápomocný, jasný a informatívny. S rastúcimi možnosťami prispôsobenia, multimodality a rozšíreniami sa ChatGPT stáva dôležitým nástrojom pre každodenné využitie umelej inteligencie. (OpenAI, 2023)

3.2 Web scraping

Web scraping je proces získavania údajov z webových stránok. Tieto údaje sa zbierajú a potom prevádzajú do formátu, ktorý je pre používateľa praktickejší, napríklad do tabuľky alebo cez API. Hoci je možné web scraping vykonávať ručne, zvyčajne sa uprednostňujú automatizované nástroje, pretože sú cenovo dostupnejšie a pracujú rýchlejšie. Napriek tomu web scraping často nie je jednoduchou úlohou. Webové stránky sa líšia štruktúrou a formátom, čo vedie k tomu, že nástroje na scraping majú rôznu funkčnosť a vlastnosti. (Perez, 2023)

Teoreticky je web scraping proces získavania údajov akýmkoľvek spôsobom, ktorý nezahŕňa interakciu programu s API. Najčastejšie sa to robí vytvorením automatizovaného programu, ktorý kontaktuje webový server, vyžiada si údaje (zvyčajne vo forme HTML a ďalších súborov tvoriacich webové stránky) a následne tieto údaje spracuje, aby z nich vyextrahoval potrebné informácie. V praxi web scraping zahŕňa širokú škálu programovacích techník a technológií, ako sú analýza údajov, spracovanie prirodzeného jazyka či zabezpečenie informácií. (Mitchell, 2018)

Web scraping najčastejšie využívajú webové stránky na porovnávanie cien. Tie pomocou neho zhromažďujú ponuky rovnakých produktov či služieb z rôznych zdrojov a prehľadne ich zoradia na jednom mieste, aby si zákazníci mohli ľahko vybrať tú najlepšiu ponuku. Ďalšími častými používateľmi sú firmy zaoberajúce sa prieskumom trhu, ktoré takto sledujú napríklad diskusie na sociálnych sieťach a online fórach. Nemôžeme zabudnúť ani na webové stránky, ktoré zbierajú voľné pracovné miesta z rôznych portálov, a tiež na internetové roboty a vyhľadávače ako Google, Bing či Safari, ktoré analyzujú obsah webových stránok a určujú ich poradie vo výsledkoch vyhľadávania. (Pawłowski, 2025)

Pomocou web scrapingu je možné z internetu získať širokú škálu dát. Dokáže zbierať textové informácie, ako sú opisy produktov, ich ceny, kontaktné údaje a hodnotenia zákazníkov, ale aj vizuálny obsah, napríklad obrázky a videá. V závislosti od toho, na čo sa dáta použijú, sa dá zamerať na konkrétne typy informácií, ako sú ponuky nehnuteľností, vývoj na burze, voľné pracovné miesta, dáta pre prieskum trhu alebo ceny leteniek. Web scraping sa využíva aj na získavanie potenciálnych zákazníkov (lead generation), na analýzu sentimentu (nálad) v príspevkoch na sociálnych sieťach, na zhromažďovanie správ pre agregátory obsahu, na získavanie obsahu médiami a na zber vedeckých dát pre akademický výskum. (Barton, 2024)

3.3 Technológia optického rozpoznávania znakov

V ére digitálnej transformácie je potreba efektívneho získavania a analýzy údajov dôležitejšia než kedykoľvek predtým. Technológia OCR prináša zásadnú zmenu v spôsobe, akým spracovávame a organizujeme obrovské objemy textových dát obsiahnutých v obrázkoch. Jej využitie siaha od rýchleho vyhľadania konkrétnej faktúry medzi tisíckami dokumentov bez potreby prehľadávania papierov až po automatizáciu vkladania údajov rozpoznaním informácií z faktúr a ich prenosom do tabuliek. Firmám umožňuje vytvárať prehľadné a vyhľadávateľné databázy, zefektívňovať procesy a zvyšovať produktivitu, pričom minimalizuje riziko chýb spôsobených ľudským faktorom. Najnovšie sa OCR uplatňuje pri rozpoznávaní dopravných značiek v autonómnych vozidlách či pri čítaní ŠPZ kamerami na mýtnych bránach. Technológia má korene na začiatku 20. storočia, keď Emanuel Goldberg skonštruoval zariadenie na čítanie tlačených znakov a ich premenu na telegrafický kód. S rozvojom digitálnych počítačov v 50. a 60. rokoch sa OCR začala využívať na digitalizáciu tlačených dokumentov, napríklad bankových šekov. (Smith, 2023)

Optické rozpoznávanie znakov (OCR) je elektronický proces premeny textu z obrázkov na digitálne kódovaný text prostredníctvom špecializovaného softvéru. Pomocou

OCR softvéru môže počítač previesť naskenovaný papier, digitálnu fotografiu textu alebo iný obrazový text na údaje, ktoré sú pre stroj čitateľné a editovateľné. (Russell, 2023)

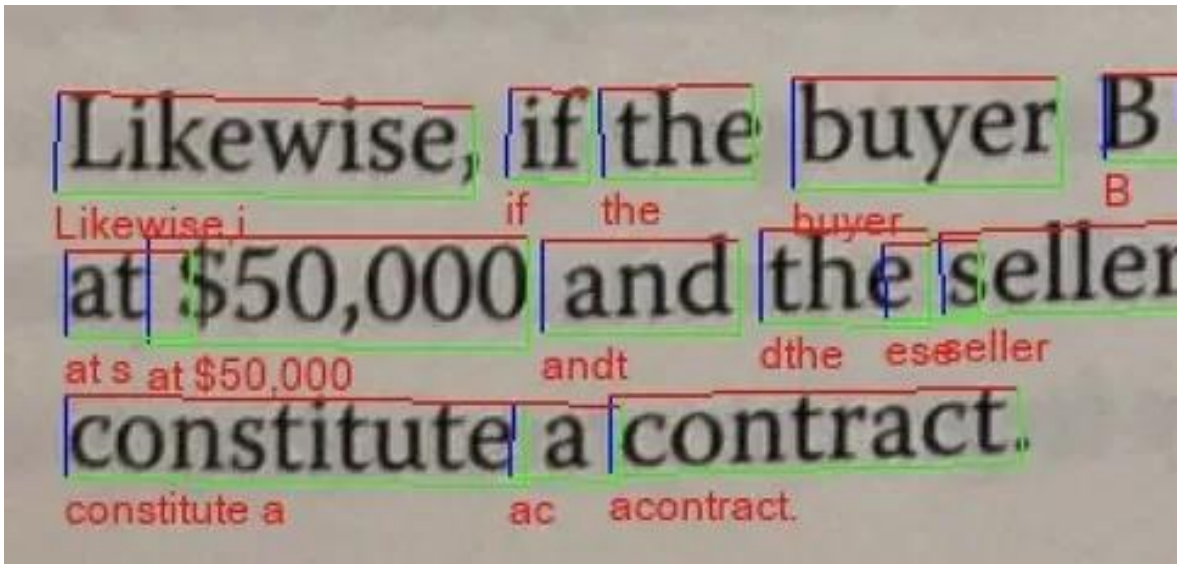
OCR technológia má 3 časti, ktoré na seba nadväzujú: (Russell, 2023)

- otvorenie a skenovanie dokumentu v softvéri OCR,
- rozpoznanie dokumentu v softvéri OCR
- uloženie dokumentu vytvoreného OCR vo formáte podľa vášho výberu.

OCR preberie text z obrázkov – či už je to kniha, účtenka alebo starý list – a premení ho na údaje, ktoré počítač dokáže prečítať, upraviť a vyhľadať. Táto technológia konvertuje ručne aj strojovo písaný text zo zdrojov, ako sú obrázky, videá či naskenované dokumenty vo formáte PDF, do digitálneho formátu pripraveného na úpravy (Potrimba, 2023).

Postupnosť v akej OCR pracuje na načítanie údajov (Potrimba, 2023):

1. **Predspracovanie obrázka**: Po zachytení obrázka prechádza fázou predspracovania, ktorá zvyšuje jeho kvalitu a pomáha pri rozpoznávaní. Tento krok môže zahŕňať úpravu veľkosti, zdokonalenie kontrastu, binarizáciu, odstránenie šumu a iné techniky.
2. **Detekcia textu**: Špeciálne navrhnutý model hlbokého učenia, vytrénovaný na veľkých dátových súboroch obrázkov a textu, umožňuje technológii počítačového videnia rozpoznať časti vstupného obrázka, ktoré pravdepodobne obsahujú text. Tento proces je spravidla zásadný.
3. **Analýza rozloženia**: Keď sú textové oblasti odhalené, technológia počítačového videnia analyzuje ich rozvrhnutie, aby určila štruktúru a poradie textu na obrázku. Tento proces podporuje zachovanie zmyslu a organizuje výsledok pre ľahšiu čitateľnosť, no nie je súčasťou každého OCR systému.
4. **Rozpoznávanie textu**: Oblasť s detekovaným textom je analyzovaná modelom rozpoznávania textu využívajúcim hlboké učenie, ktoré integruje konvolučné a rekurentné neurónové siete. Tento systém rozpoznáva jednotlivé znaky a slová na obrázku a premieňa ich na text prístupný pre počítače.
5. **Jazykový model**: Posledný výstup podlieha ďalšiemu spracovaniu na odstránenie šumu, nápravu pravopisných chýb a zlepšenie celkovej správnosti. Predpokladaná postupnosť znakov môže zahŕňať chyby, obzvlášť pri dlhých alebo neštandardných slovách. Jazykové modely, pôsobiace ako textové procesory, vylepšujú výstup odhadovaním pravdepodobnosti slovných kombinácií podľa vstupného obrázka, pričom sa aplikujú štatistické modely aj pokročilé techniky hlbokého učenia.



Obrázok 8. fungovanie OCR softvéru na text v knihe Zdroj : (Potrimba, 2023)

Optické rozpoznávanie znakov je kľúčovou disciplínou v oblasti umelej inteligencie, počítačového videnia, analýzy vzorov a strojového učenia. Ide o jednu z prvých oblastí, ktoré sa začali skúmať v rámci umelých technológií, a dnes je považovaná za vysoko rozvinutú technológiu. Skratka OCR označuje softvér schopný elektronicky rozoznávať text – či už písaný, alebo tlačený – v digitálnych obrázkoch alebo fyzických dokumentoch, ako sú naskenované listy, a transformovať ho do podoby čitateľnej pre počítače na ďalšie spracovanie. Tento proces je známy aj ako „textové rozpoznávanie“. V jednoduchosti, OCR softvér umožňuje premenu obrázkov a dokumentov do prehľadávateľného formátu. Medzi nástroje OCR patria napríklad extraktory textu, konvertory PDF do textového súboru či funkcia vyhľadávania v obrázkoch od Googlu. (Boesch, 2023)

Algoritmy optického rozpoznávania znakov (OCR) rozoznávajú tlačený alebo rukou písaný text v naskenovaných dokumentoch a snímkach scén, pričom ho transformujú do textového formátu čitateľného pre počítače. Spoločne s optickými skenermi umožňujú tieto OCR softvéry konvertovať fyzické papierové dokumenty na digitálne súbory, čím uľahčujú ich ďalšie spracovanie. (Konovalchuk, 2024)

Postup techniky OCR osvedčenými 4 operáciami: (Konovalchuk, 2024)

1) **Získavanie obrazu** – Technológia OCR používa optický skener na zachytenie neupraviteľného textu z rôznych dokumentov (napríklad ploché skeny firemných archívov, zábery textu zo scén získané vonkajšou kamerou a pod.) a prevádza ho do binárneho formátu čitateľného pre stroje. Binarizácia môže prebiehať napríklad tak, že čiernym pixelom sa priradí hodnota „1“ a bielym „0“.
2) **Predspracovanie** – Softvér OCR spracováva zdrojové obrázky na súhrnnej úrovni, aby bol text lepšie čitateľný a šum bol minimalizovaný alebo úplne odstránený. Tento

proces môže zahŕňať rôzne metódy, ako napríklad korekciu sklonu, analýzu usporiadania alebo rozdelenie znakov na segmenty.

3) **Rozpoznávanie textu** – Systém prehľadáva obsah obrázka, aby našiel zhluky pixelov, ktoré zrejme predstavujú jednotlivé znaky, a zaradí ich do príslušných kategórií. Podľa zvolenej metódy (porovnávanie vzorov alebo analýza prvkov) následne softvér porovnáva tieto znaky so štandardizovanými OCR šablónami či predchádzajúcimi modelmi, prípadne využíva algoritmy strojového učenia na určenie charakteristík opakujúcich sa pixelových zoskupení.
4) **Koncové spravovanie** – Po dokončení spracovania systém OCR premení získané textové dáta na jednoduchý znakový súbor alebo, v prípade sofistikovanejších riešení, na PDF súbor s anotáciami, ktorý zachováva pôvodné usporiadanie strany. Súčasný OCR softvér dokáže vytvárať veľmi presné výsledky, no používatelia môžu jeho presnosť ďalej vylepšiť, napríklad úpravou výstupu algoritmu pomocou dodatočného trénovania s novými textovými dátami.

3.3.1 Analýza hardvérových prostriedkov OCR

Techniky OCR potrebujú na presné a kvalitné rozpoznávanie znakov obrázky vo vysokom rozlíšení alebo kvalite, kde je text jasne odlíšený od pozadia. Spôsob vytvárania obrázkov zohráva kľúčovú úlohu v úspešnosti a presnosti OCR, pretože výrazne ovplyvňuje ich kvalitu. Obrázky zo skenerov zvyčajne zabezpečujú vysokú presnosť a spoľahlivosť OCR, zatiaľ čo fotografie z fotoaparátov sú menej vhodné kvôli vplyvom prostredia či vlastnostiam zariadenia. (Hamad and Kaya, 2016)

Texty majú odlišné rozmery. Niektoré sú krátke, ako značenie na cestách, iné sú dlhšie, napríklad popisy k videám. Pri vyhľadávaní textu je potrebné zohľadniť jeho polohu, veľkosť a dĺžku, čo vedie k vysokej výpočtovej náročnosti (Hamad and Kaya, 2016).

Kurzívne a skriptové písma môžu spôsobiť prekrývanie znakov, čo bráni hladkému priebehu kľúčových OCR procesov, ako je segmentácia. Rôzne písma prinášajú značné rozdiely v rámci jednej triedy znakov a tvoria viaceré vzorové podskupiny, čo sťažuje presné rozoznanie pri veľkom počte tried znakov. (Hamad and Kaya, 2016)

Aj keď latinské jazyky obsahujú mnoho znakov, jazyky ako japončina, čínština a kórejčina majú oveľa väčší počet tried znakov. Arabské jazyky používajú prepojené znaky, ktoré menia svoj tvar v závislosti od kontextu. V hindčine sa slabiky vytvárajú spojením abecedných písmen do tisícov variácií. Pri multijazyčnom prostredí je OCR v

naskenovaných dokumentoch významnou výskumnou otázkou, pretože práca so zložitými symbolmi je komplikovanejšia. (Hamad and Kaya, 2016)

3.3.2 Analýza softvérových prostriedkov OCR

Softvérové faktory presnosti technológie OCR: (Rao, 2024)

1) **Pokročilé algoritmy** : Pre dosiahnutie presnejšieho rozpoznávania textu je vhodné využiť OCR softvér, ktorý disponuje sofistikovanými algoritmami a technológiami strojového učenia.
2) **Pravidelné aktualizácie** : Pravidelné aktualizácie OCR softvéru, ktoré obsahujú nové vylepšenia a opravy chýb, sú nevyhnutné pre dosiahnutie optimálneho výkonu a presnosti rozpoznávania.
3) **Vylepšenie obrazu**: Použitím nástrojov na optimalizáciu obrazu pred spracovaním OCR je možné výrazne zlepšiť čitateľnosť textu v dokumentoch, čo prispieva k presnejšiemu rozpoznávaniu.
4) **Segmentácia**: Efektívne oddelenie textu od vizuálnych prvkov v dokumente umožňuje softvéru OCR lepšie sa sústrediť na textový obsah, čo vedie k zlepšeniu presnosti rozpoznávania.

Správnym riešením týchto faktorov a zavedením overených postupov dokážu organizácie zlepšiť presnosť OCR a zvýšiť celkovú efektivitu svojich procesov spracovania dokumentov. (Rao, 2024)

Softvér OCR rozpoznáva znaky tak, že ich tvary porovnáva s tvarmi uloženými v jeho databáze. Na identifikáciu slov využíva blízkosť jednotlivých znakov a snaží sa obnoviť pôvodné rozloženie stránky. Najvyššia presnosť sa dosahuje pri použití zreteľných a kvalitných skenov z dobre zachovaných originálov, avšak klesá, ak kvalita originálu alebo skenovania nie je dostatočná. (Gregersona, 2025)

Rozdelenie OCR softvér na dve druhy: (Konovalchuk, 2024)

- **Jednoduchý softvér na rozpoznávanie optických znakov a slov –** Tento druh softvéru OCR analyzuje zachytené textové obrázky porovnávaním s vopred nastavenými šablónami, ktoré zodpovedajú konkrétnym vzorom textových obrázkov. Porovnávanie môže prebiehať buď po jednotlivých znakoch, alebo po celých slovách. Keďže rôznorodosť rukopisov by si vyžadovala obrovské množstvo šablón v databázach, tieto systémy sú schopné spracovať iba text vytvorený strojom.
- **Inteligentný softvér na rozpoznávanie znakov a slov –** Pokročilý softvér OCR využíva umelú inteligenciu, najmä neurónové siete, namiesto spoliehania sa na

prednastavené textové vzory. Tieto modely je možné vycvičiť na rozsiahlych dátových sadách, čo im umožňuje identifikovať text z obrázkov bez potreby manuálne vytvorenej heuristiky.

3.4 Programovací jazyk Python a prislúchajúce knižnice

Python je programovací jazyk, ktorý umožňuje objektovo-orientované, funkcionálne aj imperatívne programovanie. Vďaka svojej prehľadnosti a jednoduchosti je skvelou voľbou pre nováčikov. Primárne ide o skriptovací jazyk, no možno ho preložiť do binárneho formátu, ktorý počítač dokáže spracovať. Hlavným prínosom je, že na vytvorenie programu stačí menej riadkov kódu než pri použití jazykov ako C/C++ či Java. (Sloan, 2016)

3.4.1 PyPDF2

Python knižnica PyPDF2 zjednodušuje prácu s PDF súbormi a je ideálna na vývoj automatizovaných aplikácií, ktoré spracúvajú PDF dokumenty, či už ide o webové alebo stolové programy. Umožňuje extrahovať údaje, ako sú texty či obrázky, a vytvárať nové PDF dokumenty. Podporuje funkcie ako rozdelenie dokumentov na jednotlivé stránky, ich spájanie, orezávanie alebo zlúčenie viacerých stránok do jednej. Okrem toho dokáže získať metadáta dokumentu, napríklad názov, autora či počet stránok. Knižnica umožňuje aj úpravy PDF, vrátane pridávania hesiel, nastavenia zobrazenia alebo vkladania vlastných údajov. PyPDF2 obsahuje viaceré triedy a funkcie, medzi hlavné patria PdfFileReader a PdfFileWriter, ktoré slúžia na čítanie a zápis PDF súborov. Trieda PageObject umožňuje manipulovať s obsahom stránky, ako sú texty, obrázky či iné objekty. Trieda DocumentInformation sprístupňuje a upravuje metadáta dokumentu, ako sú názov alebo autor. Trieda ContentStream zase umožňuje pracovať s obsahovými prúdmi dokumentu, vrátane textu a obrázkov. PyPDF2 je všestranná a výkonná knižnica, ktorá uľahčuje vytváranie, úpravu a správu PDF dokumentov pomocou širokej škály nástrojov a funkcií. (Makka, 2023)

3.4.2 requests

Knižnica requests, postavená na protokole HTTP/1.1, uľahčuje komunikáciu s webom bez zbytočných komplikácií. Nie je potrebné ručne pridávať parametre do URL ani kódovať údaje pre POST požiadavky. Funkcie ako keep-alive či automatická správa HTTP pripojení fungujú úplne automaticky vďaka integrovanej knižnici urllib3. Vďaka requests nemusíme neustále riešiť kódovanie parametrov, či už ide o GET alebo POST požiadavky. (Chandra and Varanasi, 2015)

3.4.3 Pandas

Pandas patrí medzi najobľúbenejšie knižnice Pythonu v dátovej vede a analýze. Je to open-source knižnica, ktorá umožňuje efektívne spracovanie dát podobných tabuľkám, vrátane rýchleho načítania, úpravy, zosúlaďovania a spájania údajov. Poskytuje širokú škálu funkcií a metód, ktoré zefektívňujú analýzu a predspracovanie dát. Obsahuje nástroje a dátové štruktúry na rýchle a jednoduché čistenie a analýzu údajov. Pandas sa často kombinuje s numerickými knižnicami ako NumPy a SciPy, analytickými nástrojmi ako statmodels či scikit-learn a vizualizačnými knižnicami ako matplotlib. Na rozdiel od NumPy, ktoré je optimalizované pre homogénne numerické polia, je Pandas určený na prácu s tabuľkovými a rôznorodými dátami. Python s Pandas sa využíva v mnohých odvetviach, vrátane akademického výskumu, financií, ekonómie, štatistiky, bioinformatiky, medicínskeho výskumu a ďalších. (Gupta and Bagchi, 2024)

3.5 Nástroje na uchovávanie získaných údajov

Extrahované údaje, ako sú názvy produktov, ceny, hmotnosti či zľavy, je potrebné efektívne uskladniť, aby boli dostupné pre ďalšie spracovanie a analýzu. Existuje viacero spôsobov uchovávania údajov, ktoré sa líšia v závislosti od štruktúry dát, požiadaviek na prístupnosť a rozsahu spracovania. Medzi najbežnejšie metódy patria databázové systémy, tabuľkové systémy a textové systémy.

Databázové systémy umožňujú štruktúrované uchovávanie údajov s vysokou efektivitou pri vyhľadávaní a analýze. Tabuľkové systémy sú vhodné na jednoduché spracovanie a vizualizáciu dát, najmä ak ide o menší objem údajov. Textové systémy, hoci menej štruktúrované, poskytujú flexibilitu a jednoduchosť pri ukladaní dát, čo je výhodné v počiatočných fázach spracovania. Výber vhodnej metódy závisí od požiadaviek na rýchlosť prístupu, škálovateľnosť a typ analýz, ktoré sa plánujú vykonať.

Kombinácia viacerých metód môže byť výhodná, napríklad použitie databázy na hlavné údaje a textových súborov na archiváciu pôvodných dát. Tieto metódy spolu umožňujú efektívne uchovávanie a správu údajov.

3.5.1 Databázové systémy

Databázové systémy sú navrhnuté na efektívne uchovávanie a správu štruktúrovaných údajov, ako sú názvy produktov, ceny, hmotnosti či zľavy. Umožňujú rýchle vyhľadávanie, triedenie a analýzu dát pomocou SQL dotazov, čo je ideálne pre komplexné analýzy, napríklad porovnávanie cien komodít v čase.

- **MySQL -** sú obľúbené pre svoju jednoduchosť a širokú komunitnú podporu, vhodné na menšie až stredne veľké projekty
- **PostgreSQL -** ponúka pokročilé funkcie, ako sú geografické dáta, a je ideálne pre väčšie aplikácie
- **SQLite -** je ľahká databáza bez servera, vhodná pre malé projekty
- **Oracle -** Výkonný relačný databázový systém, ideálny pre veľké podnikové aplikácie, s robustnou podporou SQL a vysokou škálovateľnosťou

3.5.2 Tabuľkové systémy

Tabulkové systémy sú jednoduchšie a vhodné na uchovávanie menšieho objemu dát s dôrazom na prehľadnosť a základnú analýzu. Umožňujú vizualizáciu údajov v riadkoch a stĺpcoch, čo uľahčuje rýchle výpočty, ako sú súčty cien komodít.

- **Microsoft Excel -** je najrozšírenejší nástroj, podporuje formáty .xlsx a ponúka funkcie ako filtre, grafy a makrá, ideálne na rýchlu analýzu
- **Google Sheets -** je cloudová alternatíva, umožňuje zdieľanie a spoluprácu v reálnom čase, vhodná pre tímovú prácu
- **LibreOffice Calc -** je open-source riešenie, kompatibilné s formátmi .ods a .csv, vhodné pre používateľov hľadajúcich bezplatný nástroj

3.5.3 Textové systémy

Textové systémy ukladajú údaje v jednoduchých formátoch, ako sú .txt alebo .doc, a sú nenáročné na zdroje, no menej štruktúrované, čo obmedzuje ich využitie na komplexné analýzy. Sú ideálne na archiváciu surových dát, napríklad pôvodných textov z letákov.

- **Notepad -** je základný editor na Windows, jednoduchý na rýchle poznámky, no bez pokročilých funkcií
- **Microsoft Word -** Výkonný editor na tvorbu dokumentov, podporuje formátovanie, tabuľky a obrázky, no nie je zadarmo
- **Google Docs -** Webový editor, zdarma, vhodný na tímovú prácu a jednoduché dokumenty, uložené v cloude
- **LibreOffice Writer -** Zdarma editor, podobný Wordu, s funkciami na formátovanie a tabuľky, no menej intuitívny

4 Výsledky práce

Načítavanie údajov z PDF dokumentov do textového súboru je kľúčovým procesom pri automatizovanom spracovaní informácií získaných z internetových zdrojov. Tento postup zahŕňa získanie PDF súborov z webových stránok, ich uloženie na lokálny disk a extrakciu textového obsahu, ktorý je následne zapísaný do textového formátu. Výsledkom sú excelovské súbory obsahujúce štruktúrovaný text, ktorý je prezentovaný na webovej stránke.

4.1 Rozbor stránky na vyhľadávanie a stiahnutie letáku

Ako prvý krok bol pre každú webovú stránku potravinárskeho obchodu vytvorený vlastný program na vyhľadávanie PDF letáku, ktorý následne stiahne na disk. Tento program je navrhnutý tak, aby najprv analyzoval štruktúru webovej stránky a identifikoval kľúčové prvky, ako sú HTML tagy obsahujúce odkazy na PDF súbory. Pomocou techník webového scrapingu, využitím knižníc BeautifulSoup a requests v Pythone, sa prehľadávajú DOM štruktúry stránok, čím sa zefektívňuje proces vyhľadávania.

Automatizácia sťahovania zahŕňa aj správu výnimiek, ako sú nefunkčné odkazy alebo obmedzenia prístupu, ktoré môžu vyžadovať použitie hlavičiek HTTP či oneskorenie požiadaviek, aby sa predišlo blokovaniu zo strany servera.

4.1.1 Billa

Sťahovanie PDF letáku z webovej stránky billa.sk začína prístupom na stránku https://www.billa.sk/letaky-a-akcie/aktualny-letak, ktorá slúži ako vstupný bod pre získanie akciových letákov. Pomocou HTTP požiadavky sa načíta zdrojový kód stránky, ktorý sa analyzuje na prítomnosť odkazov na podstránky s letákmi.

Následne sa vykoná ďalšia HTTP požiadavka na získanú podstránku, kde sa v jej zdrojovom kóde vyhľadá priamy odkaz na PDF súbor. Tento odkaz je označený reťazcom “downloadPdfUrl":“. Algoritmus opäť identifikuje začiatok a koniec odkazu, čím získa finálnu URL adresu PDF súboru. Počas oboch krokov sa kontroluje úspešnosť požiadaviek, aby sa predišlo chybám, ako sú nefunkčné odkazy alebo obmedzenia zo strany servera.

PDF súbor sa stiahne pomocou streamovanej požiadavky, ktorá umožňuje sťahovanie po menších častiach, čím sa šetrí pamäť a zvládajú aj väčšie súbory. Súbor sa ukladá lokálne pod názvom billa_letak.pdf s poradovým číslom. Po úspešnom stiahnutí sa proces potvrdí, inak sa oznámi chyba. Tento automatizovaný postup zjednodušuje získavanie letákov z billa.sk.

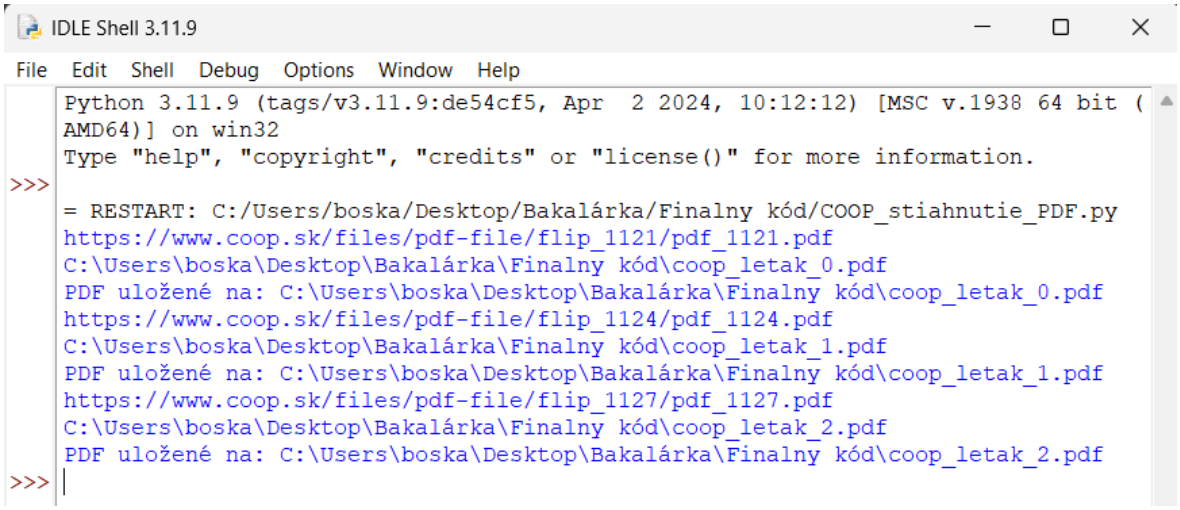
Obrázok 9. odpoveď IDLE Schell na stiahnutie Billa letáku Zdroj: (Vlastný zdroj)

Obrázok 10. stiahnutý Billa PDF leták na lokálnom úložisku Zdroj : (Vlastný zdroj)

4.1.2 COOP Jednota

Proces sťahovania PDF letáku z webovej stránky coop.sk, začína načítaním stránky https://www.coop.sk/sk/pdfflip/1 pomocou knižnice requests v Pythone. Najprv sa načíta zdrojový kód stránky prostredníctvom HTTP požiadavky a overí sa jej úspešné načítanie. V zdrojovom kóde sa vyhľadajú odkazy na PDF súbory, ktoré sú označené špecifickým reťazcom source=". Algoritmus postupne identifikuje začiatok a koniec každého odkazu, čím získa čisté URL adresy PDF súborov. Tieto odkazy sa ukladajú do zoznamu, pričom sa zabezpečí, aby sa duplicitné odkazy nezaznamenávali, čím sa zamedzí zbytočnému sťahovaniu rovnakých dokumentov.

Následne sa každý získaný odkaz použije na stiahnutie príslušného PDF súboru. Sťahovanie prebieha po častiach, aby bolo efektívne a zvládalo aj väčšie súbory. Súbory sa ukladajú na lokálne úložisko s názvami, ktoré zahŕňajú poradové číslo, napríklad coop_letak_0.pdf, coop_letak_1.pdf a podobne, čím sa zabezpečí ich prehľadná organizácia. V ďalších kapitolách výsledkov práce sa stiahnuté letáky COOP Jednota rozdelia na COOP Jednota Supermarket a COOP Jednota Tempo. Počas sťahovania sa kontrolujú potenciálne chyby, ako sú nefunkčné odkazy alebo sieťové problémy, a v prípade úspechu sa potvrdí uloženie súboru. Ak nastane problém, chyba sa oznámi. Tento automatizovaný proces zjednodušuje získavanie akciových letákov z coop.sk a umožňuje ich ďalšie spracovanie, ako je extrakcia obsahu.



Obrázok 11. odpoveď IDLE Schell na stiahnutie COOP Jednota letáku Zdroj : (Vlastný zdroj)

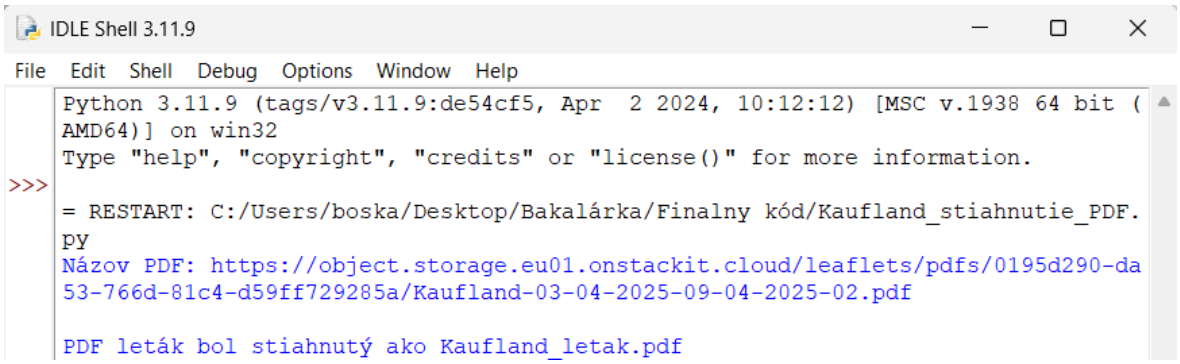
Obrázok 12. stiahnutý COOP Jednota Supermarket PDF leták na lokálnom úložisku Zdroj : (Vlastný zdroj)

Obrázok 13. stiahnutý COOP Jednota Tempo PDF leták na lokálnom úložisku Zdroj : (Vlastný zdroj)

4.1.3 Kaufland

Proces získavania PDF letáku z webovej stránky kaufland.sk začína návštevou stránky https://predajne.kaufland.sk/aktualna-ponuka/letak.html, kde sú dostupné informácie o aktuálnych akciových ponukách. Tento web slúži ako hlavný zdroj pre získanie letákov, ktoré obsahujú zľavy a špeciálne ponuky reťazca Kaufland. Prvým krokom je načítanie zdrojového kódu stránky prostredníctvom HTTP požiadavky, pričom sa overí, či sa stránka načítala správne, čo je indikované stavovým kódom 200. V HTML kóde sa potom vyhľadá špecifický identifikátor data-parameter="letak-aktualny" data-download-url=, ktorý označuje priamu URL adresu PDF letáku. Pomocou analýzy sa určí začiatok a koniec tejto URL, čím sa získa presný odkaz potrebný pre nasledujúci krok sťahovania.

Potom sa vykoná HTTP požiadavka na získanú URL adresu PDF súboru, pričom sťahovanie je rozdelené na menšie časti, aby bolo efektívne a zvládlo aj väčšie dokumenty. Súbor sa ukladá na lokálne úložisko pod názvom Kaufland_letak.pdf s poradovým číslom. Počas celého procesu sa monitorujú možné chyby, ako napríklad problémy s pripojením alebo neplatné odkazy. V prípade úspešného stiahnutia sa potvrdí, že súbor bol uložený, a ak nastane problém, chyba sa oznámi. Tento automatizovaný postup je navrhnutý tak, aby eliminoval potrebu manuálnej interakcie a zaistil spoľahlivé získavanie letákov z kaufland.sk. Stiahnutý dokument je pripravený na ďalšie použitie, ako je analýza obsahu alebo extrakcia údajov o akciách.



Obrázok 14. odpoveď IDLE Schell na stiahnutie Kaufland letáku Zdroj : (Vlastný zdroj)

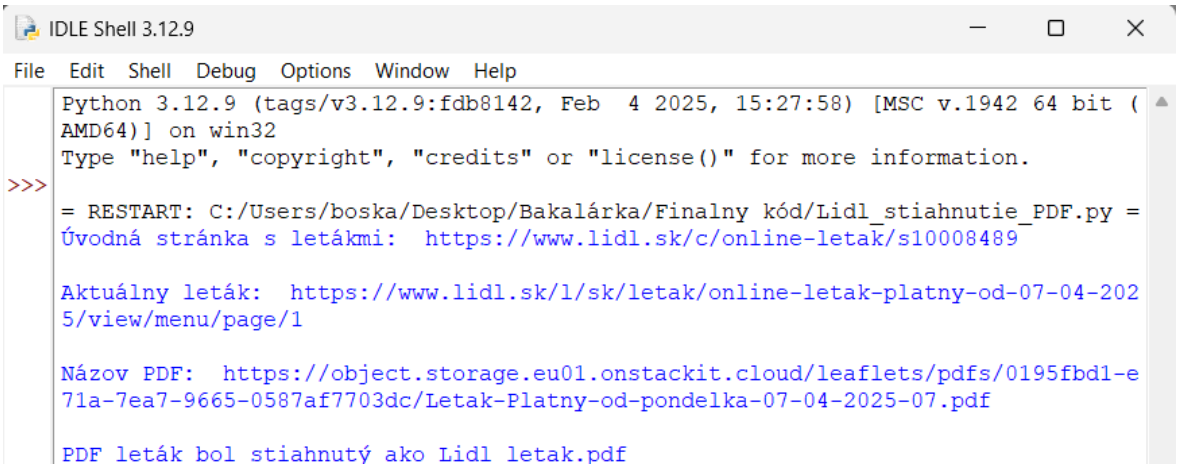
Obrázok 15. stiahnutý Kaufland PDF leták na lokálnom úložisku Zdroj : (Vlastný zdroj)

4.1.4 Lidl

Sťahovanie PDF letáku z webovej stránky lidl.sk začína prístupom na stránku https://www.lidl.sk/c/online-letak/s10008489, kde sú dostupné informácie o aktuálnych letákoch. Prvým krokom je načítanie zdrojového kódu tejto stránky cez HTTP požiadavku a potvrdenie jej úspešného načítania. V kóde sa vyhľadá odkaz na podstránku s online letákom, ktorý je následne upravený tak, aby viedol na konkrétnu sekciu letáku.

Po načítaní podstránky sa počká niekoľko sekúnd, aby sa zabezpečilo úplné vygenerovanie obsahu, a potom sa získa jej zdrojový kód. V tomto kóde sa vyhľadá odkaz na PDF súbor, ktorý je označený špecifickým textom súvisiacim s navigáciou letáku. Odkaz sa extrahuje určením jeho začiatku a konca, čím sa získa priama URL adresa PDF dokumentu. Následne sa súbor stiahne po menších častiach, aby bolo sťahovanie efektívne a zvládlo aj väčšie súbory. Ukladá sa lokálne pod názvom Lidl_letak.pdf s poradovým číslom.

Počas sťahovania sa sledujú možné chyby, ako sú problémy s pripojením alebo neplatné odkazy, a v prípade úspechu sa potvrdí uloženie súboru. Ak nastane problém, používateľ je informovaný. Tento postup kombinuje spracovanie statického aj dynamického obsahu, čím zaisťuje spoľahlivé získavanie letákov z lidl.sk pre ďalšie použitie, ako je analýza akcií.



Obrázok 16. odpoveď IDLE Schell na stiahnutie Lidl letáku Zdroj : (Vlastný zdroj)

Obrázok 17. stiahnutý Lidl PDF leták na lokálnom úložisku Zdroj : (Vlastný zdroj)

4.2 Tvorba textových súborov pomocou umelej inteligencie

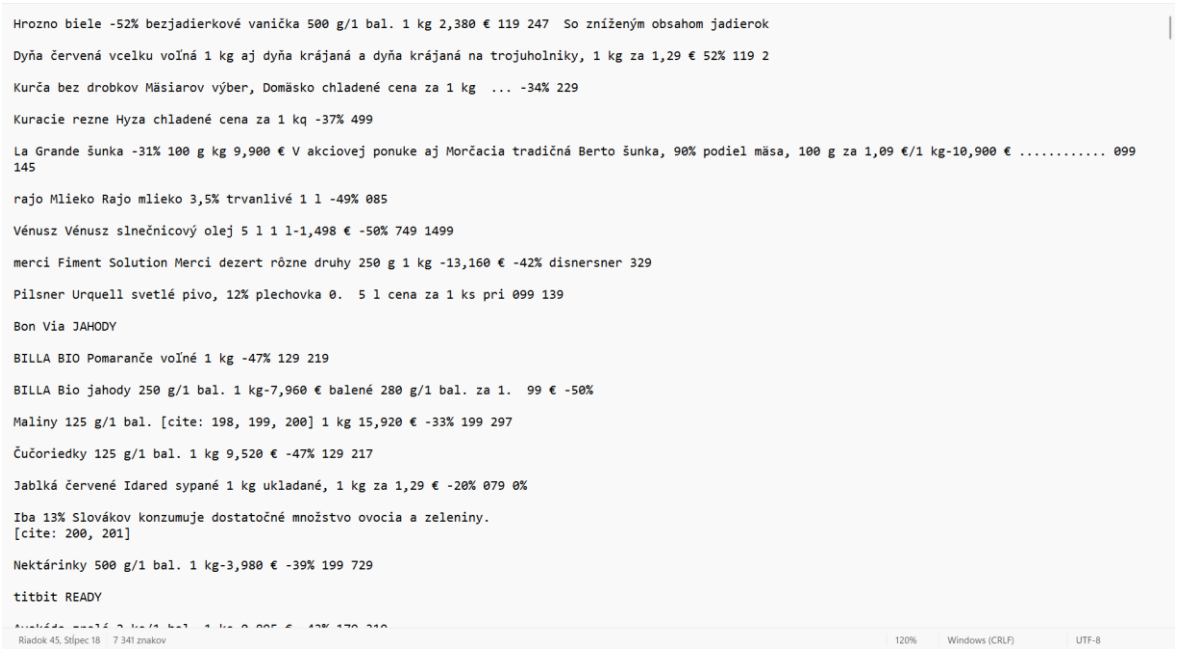
Kapitola sa zaoberá využitím umelej inteligencie (UI) pri tvorbe textových súborov obsahujúcich údaje získané z online letákov obchodných reťazcov, ako sú Lidl, Billa, Kaufland a COOP Jednota. Umelá inteligencia ponúka pokročilé možnosti na automatizáciu spracovania neštruktúrovaných dát, čo je kľúčové pre efektívne načítavanie informácií o komoditách a cenách z letákov, často dostupných vo formáte PDF. Proces zahŕňa niekoľko fáz, vrátane extrakcie textu, rozpoznávania vzorov a štruktúrovania dát do textových súborov vhodných na ďalšie použitie.

4.2.1 Billa

V rámci testovania schopností umelej inteligencie pri spracovaní údajov z PDF letákov bol použitý model Chat GPT na prepis údajov z letáku obchodného reťazca Billa. Cieľom bolo overiť presnosť a efektivitu modelu pri extrakcii údajov, ako sú názvy produktov, akciové ceny, percentuálne zľavy, ceny pred akciou a jednotkové ceny a ich transformáciu do textového formátu.

Model Chat GPT preukázal schopnosť správne identifikovať väčšinu kľúčových informácií, ako sú názvy produktov a akciové ceny, jednotkové ceny, percentuálne hodnoty akcie a ceny pred akciou. Pri niektorých položkách, však došlo k určitým nepresnostiam. Model nesprávne interpretoval číselné hodnoty ceny, neuviedol percentuálnu zľavu, akciové a pred akciové ceny potravín, ktoré boli uvedené v letáku a nesprávne prepísal údaj o hmotnosti, kde uviedol kq namiesto kg.

Pre ďalšiu prácu s údajmi je potrebné použíť inú metódu alebo spôsob na prepis údajov z Billa letákov do textového súboru.



Obrázok 18. prepísanie Billa leták pomocou AI do textového súboru Zdroj : (Vlastný zdroj)

4.2.2 COOP Jednota Supermarket

V rámci testovania automatizovaného spracovania údajov z letákov bol model umelej inteligencie Gemini využitý na prepis údajov z letáku obchodného reťazca COOP Jednota Supermarket. Leták obsahoval zoznam akciových produktov, kde boli uvedené len názvy produktov, percentuálna hodnota akcie a akciová cena. Cieľom bolo overiť schopnosť

modelu extrahovať a štruktúrovať údaje, ako sú názvy produktov, ceny a hmotnosti, do textového formátu.

Silnou stránkou pri tomto druhu spracovania bolo správne rozpoznanie názvov produktov, percentuálnej hodnoty akcie a akciových cien, najmä pri položkách s jednoduchou štruktúrou. Model dokázal normalizovať jednotkové ceny, ktoré boli v celom dokumente správne identifikované. Problémy nastali pri položkách, kde pred názov produktu vložil určitú časť akciovej ceny.

Model Gemini ukázal dobré výsledky pri prepise údajov z letáku COOP Jednota, no pre dosiahnutie vyššej presnosti je potrebné zamerať sa na lepšiu optimalizáciu vstupných dát.

Obrázok 19. prepísanie COOP Jednota Supermarket leták pomocou AI do textového súboru Zdroj : (Vlastný zdroj)

4.2.3 COOP Jednota Tempo

V rámci testovania automatizovaného spracovania údajov z letákov bol model umelej inteligencie Chat GPT využitý na prepis údajov z letáku obchodného reťazca COOP Jednota Supermarket. Leták obsahoval zoznam akciových produktov, kde boli uvedené len názvy produktov, percentuálna hodnota akcie a akciová cena. Cieľom bolo overiť schopnosť modelu extrahovať a štruktúrovať údaje, ako sú názvy produktov, hmotnosti, akciové a jednotkové ceny do textového formátu.

Pri tomto druhu spracovania bolo vo väčšine prípadov správne rozpoznanie názvov produktov, percentuálnej hodnoty akcie a akciových cien, najmä pri položkách s jednoduchou štruktúrou. Model dokázal normalizovať jednotkové ceny, ktoré boli v časti

dokumente správne identifikované. Problémy nastali pri položkách s jednotkovými cenami, kde nie sú zapísané akciové ceny pre dané produkty.

Model Chat GPT ukázal dobré výsledky pri prepise údajov z letáku COOP Jednota, no pre dosiahnutie vyššej presnosti je potrebné zamerať sa na lepšiu optimalizáciu vstupných dát a prácu s nimi.

Obrázok 20. prepísanie COOP Jednota Tempo leták pomocou AI do textového súboru Zdroj : (Vlastný zdroj)

4.2.4 Kaufland

V rámci testovania automatizovaného spracovania údajov z PDF letákov bol model umelej inteligencie Gemini využitý na prepis údajov z letáku obchodného reťazca Kaufland. Leták obsahoval zoznam akciových produktov, kde boli uvedené údaje, ako názov potravín, akciová cena potravín, percentuálna hodnota akcie, cena pred akciou. Cieľom bolo zhodnotiť schopnosť modelu Gemini extrahovať a štruktúrovať údaje o produktoch, cenách a hmotnostiach do textového formátu.

Model Gemini úspešne identifikoval názvy produktov akciové ceny, percentuálne hodnoty akcie a ceny pred akciou. Model správne extrahoval hlavnú cenu, ale neinterpretoval jednotkovú cenu v plnom rozsahu, pravdepodobne kvôli nejednoznačnému formátovaniu. Hlavným nedostatkom bolo nedostatočné spracovanie určitých položiek, kde sú uvedené len názvy produktov bez ďalších doplnkových informácií ako akciová cena, percentuálna hodnota akcia a cena pred akciou.

Pri tomto druhu prepisovania údajov je potrebné použiť iný spôsob pre dosiahnutie správnej presnosti.

Obrázok 21. prepísanie Kaufland leták pomocou AI do textového súboru Zdroj : (Vlastný zdroj)

4.2.5 Lidl

V rámci testovania automatizovaného spracovania údajov z PDF letákov bol model umelej inteligencie Gemini použitý na prepis údajov z letáku obchodného reťazca Lidl. Leták obsahoval akčné ponuky, ako názov produktu, akciová cena, percentuálna hodnota akcie alebo podmienky akcie, jednotková cena a cena pred akciou. Cieľom bolo posúdiť schopnosť modelu Gemini extrahovať a štruktúrovať údaje o produktoch, cenách a špecifických podmienkach do textového súboru.

Spôsob extrahovania pomocou Gemini dokázal identifikovať názvy produktov a akciové ceny a ceny pred akciou. Problémy nastali pri položkách s dodatočnými podmienkami, ako potrebné množstvo na dosiahnutie akciovej ceny, ktoré model často ignoroval, pretože neboli jednoznačne štruktúrované. Taktiež akciová cena pri mlieku nebola správne spojená s podmienkou nákupu, čo viedlo k neúplnej interpretácii.

Model Gemini preukázal, že pri prepise, sú potrebné ďalšie vylepšenia pre správny zápis údajov z Lidl letáku.

Obrázok 22. prepísanie Lidl leták pomocou AI do textového Zdroj : (Vlastný zdroj)

4.3 Tvorba textových súborov

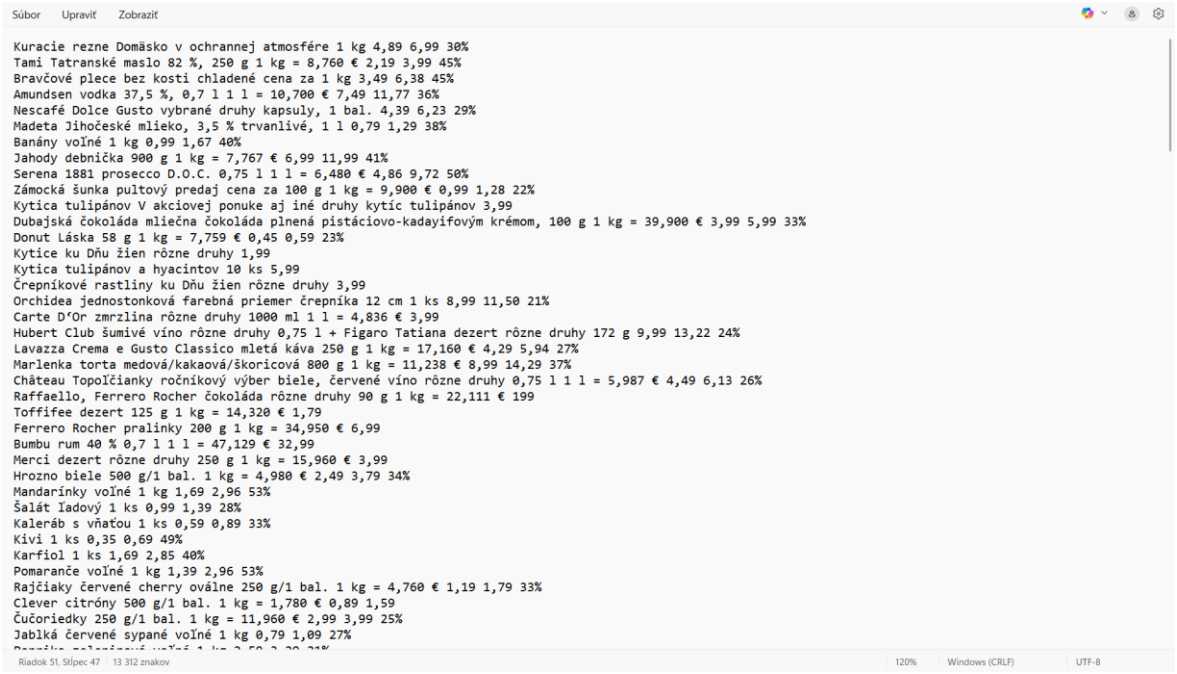
Proces načítavania údajov z PDF letákov do textového súboru predstavuje základnú časť navrhnutého systému na spracovanie informácií z vybraných obchodných reťazcov ako sú Billa, Coop Jednota, Lidl a Kaufland. Tento proces zahŕňa otvorenie každého dokumentu a získanie jeho textového obsahu. Text sa z PDF vyberá v podobe, v akej je v dokumente usporiadaný, pričom sa zachováva poradie a štruktúra informácií, ako sú názvy a ďalšie informácie. Cieľom tejto časti je získať čo najpresnejší textový výstup, ktorý odráža informácie obsiahnuté v pôvodnom dokumente. Po extrakcii je text uložený do textového súboru.

Každý PDF dokument je spracovaný samostatne, pričom výsledný textový súbor obsahuje neštruktúrované údaje. Pre lepšiu prehľadnosť môžu byť do textu vložené, napríklad nové riadky, ktoré uľahčujú rozlíšenie jednotlivých častí, napríklad stránok alebo sekcií. Tieto súbory tak predstavujú surový textový materiál, ktorý zachytáva údaje v podobe, aká bola v PDF, a sú pripravené na ďalšiu fázu spracovania.

4.3.1 Billa

Textový súbor vytvorený z PDF letáku Billa slúži ako kľúčový medzistupeň v procese spracovania údajov o komoditách a cenách. Obsahuje surové údaje extrahované z PDF dokumentu, vrátane názvov produktov, ich vlastností, akciových cien, predakciových cien , jednotkových cien a percentuálnych zliav. Dáta sú organizované v riadkoch, kde každý riadok alebo skupina riadkov reprezentuje jeden produkt s informáciami o názve, hmotnosti , jednotkovej cene, akciovej cene a percentuálnej zľave. Formát súboru je čiastočne

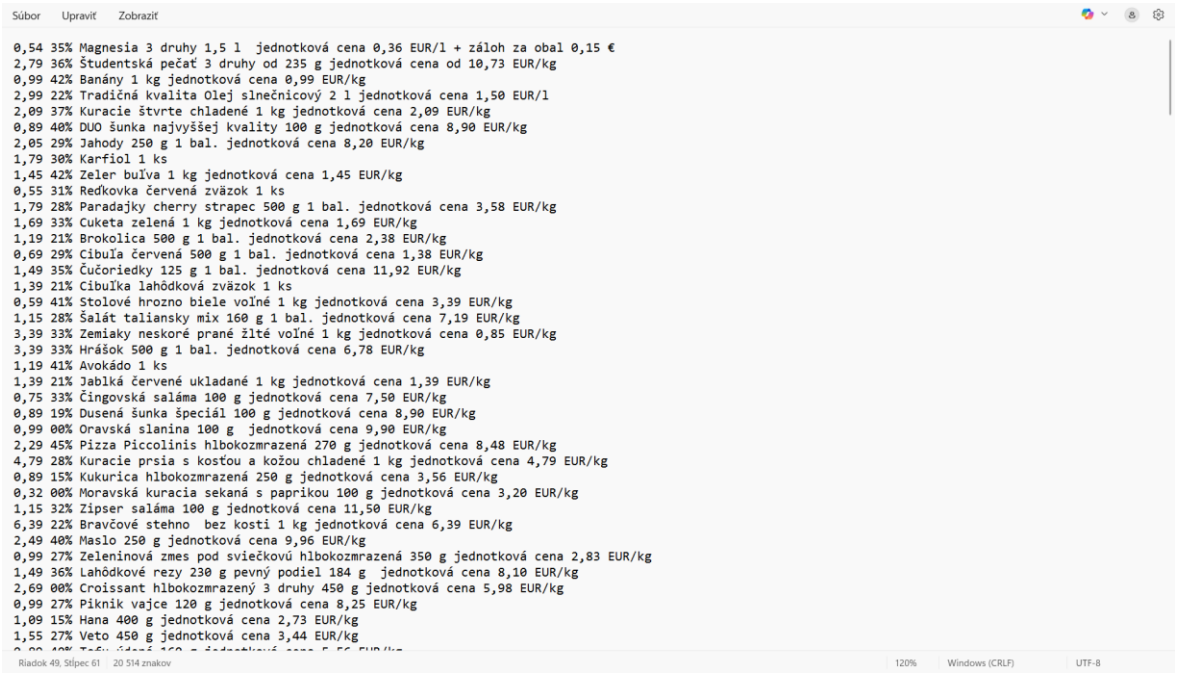
štruktúrovaný a obsahuje nekonzistencie, ako sú rôzne spôsoby zápisu cien, chýbajúce údaje či nejednotné oddeľovače. Tieto nekonzistencie vznikajú v dôsledku zložitého rozloženia PDF letákov, ktoré môžu obsahovať tabuľky, stĺpce alebo obrázky. Hlavnou úlohou súboru je poskytnúť surový textový výstup po extrakcii z PDF, ktorý slúži ako základ pre vytvorenie štruktúrovaného CSV súboru, ktorý bude prezentované na web stránke.



Obrázok 23. prepísanie Billa letáku do textového súboru Zdroj : (Vlastný zdroj)

4.3.2 COOP Jednota Supermarket

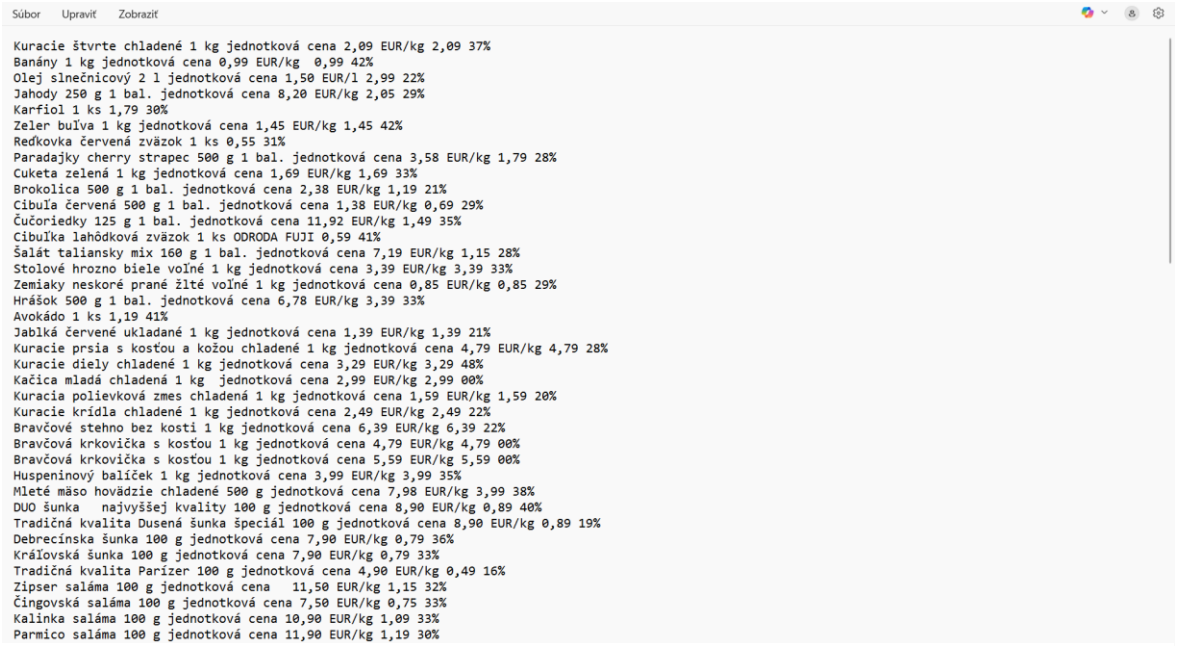
Textový súbor obsahuje surové údaje extrahované z PDF letáku Coop Jednota Supermarket, slúži ako kľúčový medzistupeň v procese spracovania údajov o komoditách a cenách. Súbor má čiastočne štruktúrovaný formát, kde každý riadok zvyčajne reprezentuje jeden produkt s informáciami o názve, hmotnosti, jednotkovej cene, akciovej cene a percentuálnej zľave. Napriek tomu obsahuje nekonzistencie, ako sú rôzne zápisy cien a nepravidelné označenia produktov. Jeho úlohou je slúžiť ako dočasné úložisko dát pred ich vyčistením a konverziou do štruktúrovaného CSV súboru, ktorý bude prezentované na web stránke.



Obrázok 24. prepísanie COOP Jednota Supermarket letáku do textového súboru Zdroj : (Vlastný zdroj)

4.3.3 COOP Jednota Tempo

Tento textový súbor obsahuje surové údaje extrahované z PDF letáku Coop Jednota Tempo, vrátane názvov produktov, hmotností/objemov, akciových cien, jednotkových cien a percentuálnych zliav. Údaje sú čiastočne štruktúrované v riadkoch, kde každý riadok reprezentuje produkt, no obsahujú nekonzistencie, ako rôzne formáty cien alebo chýbajúce zľavy, ktoré neboli v PDF letáku uvedené. Slúži ako medzistupeň medzi extrakciou z PDF a štruktúrovaným CSV výstupom, ktorý bude prezentované na web stránke.



Obrázok 25. prepísanie COOP Jednota Tempo letáku do textového súboru Zdroj : (Vlastný zdroj)

4.3.4 Kaufland

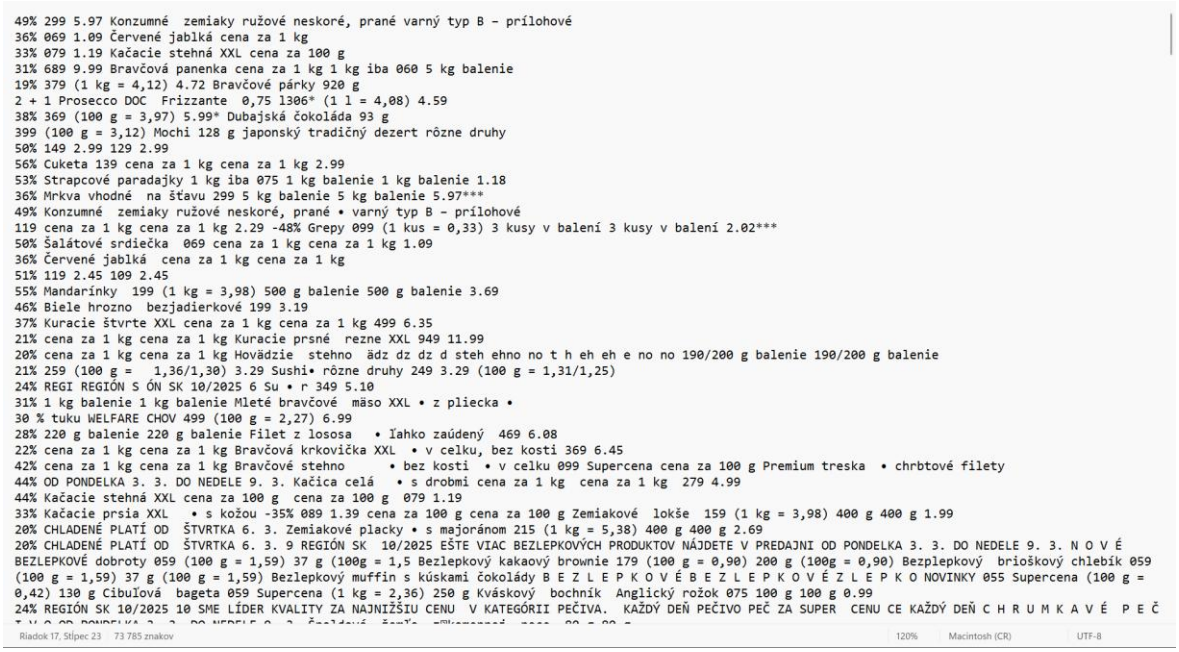
Textový súbor z PDF letáku Kaufland. Obsahuje zoznam produktov, ich popisy, hmotnosti/objemy, akciové a bežné ceny, percentuálne zľavy, jednotkové ceny a informácie o dostupnosti. Súbor obsahuje aj marketingové texty a informácie o Kaufland Card zľavách, kupónoch a súťaži o vstupenky na MS v hokeji 2025. Údaje sú čiastočne štruktúrované v sekciách, no obsahujú nekonzistencie, ako rôzne formáty cien, chybné spojenia potravín s cenami a akciami či opakovania textu. Formátovanie je náročné na spracovanie kvôli chybám z extrakcie PDF, v dôsledku toho nie je možné vytvoriť štruktúrovaný výstup vo formáte .CSV.

-36% 1,09 0,69 -33% 1,48 0,99 -31% 17,41 11,99 -45% 19,99 10,99 -40% 1,09 0,65 0,65 0,65 -39% 1,08 BEŽNÁ CENA 0,99 0,99 -52% 1,19 2,07 -42% -40% 7,15 4,29
Vaječné cestoviny rôzne druhy 200 - 250 g (=1 kg od KC: 2,60 / BC: 4,32) * Smotana na šľahanie 33 % 180 ml (=1 l 5,50) Sušienky rôzne druhy 146 - 220 g
(=1 kg od KC: 4,50 / A: 5,41) * Zrnková káva rôzne druhy 1 kg Eidam 45 % pultový predaj 100 g (=1 kg 6,50) Bylinný likér rôzne druhy 33 - 35 % alk. 0,7 l
(=1 l 17,13) Jablko červené 1 kg Bravčová krkovička bez kosti, v celku pultový predaj 1 kg AKCIOVÁ PONUKA PLATÍ DO NEDELE 9. 3. 2025 OD ŠTVRTKA 6. 3.
2025 DO STREDY 12. 3. 2025 č. 10 OD ŠTVRTKA 6 Na základe online prieskumu marketingového programu Dôveryhodné značky vyhodnoteného agentúrou NIQ v roku
2024. Sme NAJDÔVERYHODNEJŠÍ PREDAJCA POTRAVÍN Podľa štúdie spoločnosti KPMG analyzujúcej zákaznícke skúsenosti Slovákov za rok 2024. V KATEGÓRII REŤAZCE S
POTRAVINAMI. SME ZÁKAZNÍCKY SME ZÁKAZNÍCKY SME ZÁKAZNÍCKY NAJOBĽÚBENEJŠIA ZNAČKA S1-SK-KW10-LFT-TITULKA-1030-11 1 – TERAZ IBA V KAUFLANDE – – TERAZ IBA V
KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V
KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V
KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V
KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V
KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V
KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V
KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V
KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – Niektoré výrobky môžu byť v ponuke iba do vypredania zásob. Ceny v
eurách. Právo na opravu chýb v písaní a počítaní a tlačových chýb vyhradené. 1,89 3,69 -48% Hrozno biele bezkôstkové 500 g balenie (=1 kg 3,78) OD ŠTVRTKA
6. 3. S2-SK-KW10-LFT-TOP_PON_1_1_L-1020-11 2 – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ
IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ
IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ
IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ
IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ
IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ
IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ
IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – – TERAZ IBA V KAUFLANDE – Predaj v obvyklom množstve pre
jednu domácnosť. Kaufland Card cena platí pri predložení karty Kaufland Card. *A – jednotková cena platná pre akciu, BC – jednotková cena platná pre bežnú
cenu, KC – jednotková cena platná s Kaufland Card. 0,64 1,22 -47% Svetlé pivo 12 % 0,5 l (=1 l 1,28) + záloh za obal 0,15 OD ŠTVRTKA 6. 3. ** Naša najnižšia
predajná cena za posledný rok je najnižšia predajná cena vybraných výrobkov v predajniach Kaufland na Slovensku v období od 6. 3. 2024 do 12. 3. 2025 okrem
individuálne zlacnených výrobkov v našich jednotlivých predajniach napríklad z dôvodu blížiaceho sa dátumu minimálnej trvanlivosti alebo dátumu spotreby.
S3-SK-KW10-LFT-TOP_PON_1_2_R-1020-11 3 Kaufland je NAJOBĽÚBENEJŠÍM miestom na nákup ovocia a zeleniny. Podľa prieskumu agentúry NMS Market Research
Slovakia v Slovenskom národnom paneli v októbri 2024. Kaufland je VHODNÉ NA ŠŤAVU OD ŠTVRTKA 6. 3. ČERSTVÉ OVOCIE A ZELENINA 1,19 -39% 1,29 1,97 -34%
Zrelé avokádo 1 kus 0,45 -46% 0,24 0,24 Kivi 1 kus 1,59 2,29 -30% 1,26 2,39 -46% Hrušky zelené 1 kg Mandarínky 1 kg balenie 0,99 1,68 -41% Citróny 500 g
balenie (=1 kg 1,98) S4-SK-KW10-LFT-OVO_ZEL_1_1_L-1020-11 4 Paradajky cherry strapcové 500 g balenie (=1 kg 3,38) 0,44 0,55 -20% Cesnak biely 100 g (=1 kg
4,40) 2,39 3,59 -33% Paprika biela 1 kg 0,75 1,19 -36% Mrkva 1 kg balenie Čerstvé zo Slovenska 0,49 0,79 -37% Cibuľka lahôdková 1 zväzok -43% 2,99 1,69
1,69 Paradajky S5-SK-KW10-LFT-OVO_ZEL_1_2_R-1020-11 5 Naplňme spoločne školy a škôlky zdravou výživou! Registrácia škôl a škôlok do projektu Čerstvé
hlavičky je otvorená od 6. 3. do 30. 3. 2025 na www.cerstvehlavicky.sk 1,29 1,89 -31% 1,59 2,69 -40% 1,19 1,59 -25% 1,19 1,19 -55% 1,49 2,69 -44% Šalát
s korienkami farebný mix 1 kus Cuketa 1 kg Jablko Pink Lady 1 kg Tekvica maslová 1 kg Čerstvé zo Slovenska OD ŠTVRTKA 6. 3. ČERSTVÉ OVOCIE A ZELENINA ZA
SKVELÉ ČERSTVOSŤ CENY! DOPRAJTE SI S6-SK-KW10-LFT-OVO_ZEL_2_1_L-1020-11 6 14,99 do vyčerpania zásob MIMORIADNA PONUKA 7,99 do vyčerpania zásob MIMORIADNA
PONUKA 2,99 do vyčerpania zásob MIMORIADNA PONUKA 8,99 do vyčerpania zásob MIMORIADNA PONUKA 2,99 do vyčerpania zásob MIMORIADNA PONUKA 11,99 do

Obrázok 26. prepísanie Kaufland letáku do textového súboru Zdroj : (Vlastný zdroj)

4.3.5 Lidl

Textový súbor obsahuje údaje z akciového letáku Lidlu, extrahované z PDF, so zoznamom potravín a drogérie. Zahŕňa názvy produktov, ďalej uvádza ich akciové ceny, predakciové ceny, jednotkové ceny, percentuálne zľavy alebo ponuky pri nákupe viacerých kusov. Údaje sú štruktúrované v celku. Súbor obsahujú nekonzistencie, ako rôzne formáty cien, chybné spojenia potravín s cenami a akciami či opakovania textu. Formátovanie je náročné na spracovanie kvôli chybám z extrakcie PDF, v dôsledku toho nie je možné vytvoriť štruktúrovaný výstup vo formáte .CSV.



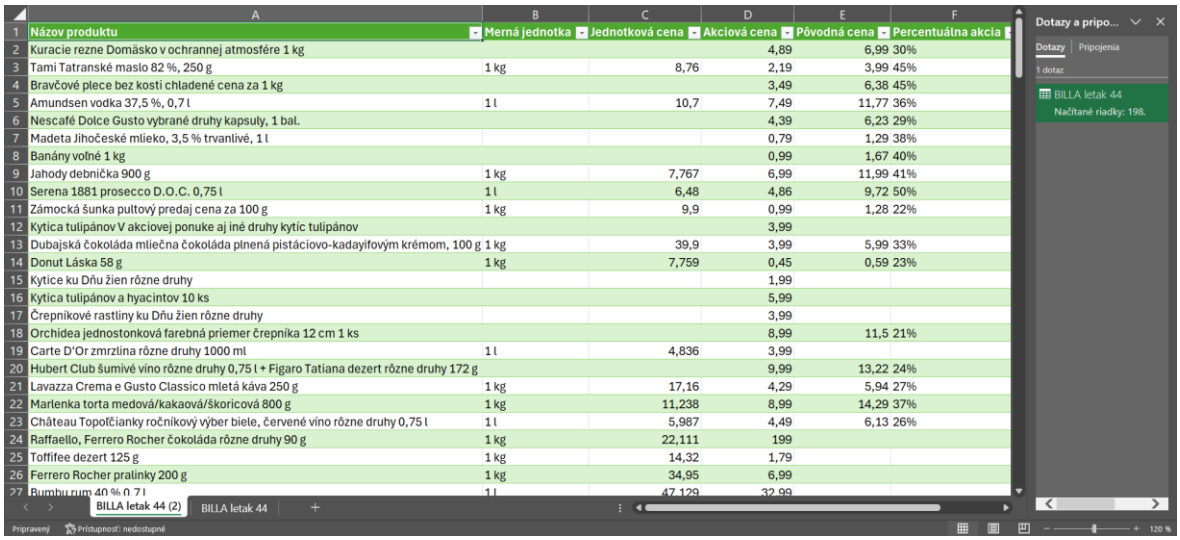
Obrázok 27. prepísanie Lidl letáku do textového súboru Zdroj : (Vlastný zdroj)

4.4 Tvorba tabuľkových súborov

Proces vytvárania štruktúrovaných tabuľkových súborov z extrahovaných dát letákov. Zameriava sa na transformáciu neštruktúrovaných textových údajov do prehľadného formátu vhodného na analýzu. Proces zahŕňa čistenie dát, kde sa odstraňujú chyby z extrakcie PDF. Čistenie dát umožňuje identifikovať a štruktúr dát položky ako názov produktu, hmotnosť, akciovú cenu, percentuálna zľavu a jednotkovú cenu. Výsledné CSV súbory obsahujú stĺpce ako „Názov produktu“, „Akciová cena“, „Pôvodná cena“, „Percentuálna zľava“ „Merná jednotka“ a „Jednotková cena“, čo uľahčuje porovnávanie produktov a cien. Kapitola zdôrazňuje význam automatizácie procesu pomocou Python skriptov, minimalizáciu manuálnej úpravy a presné formátovanie, ktoré sú nevyhnutné pre ďalšie analýzy, ako je porovnávanie cien medzi reťazcami, sledovanie zliav v čase alebo analýza nákupných trendov.

4.4.1 Billa

CSV súbor predstavuje štruktúrovaný výstup spracovania údajov z akciového letáku obchodného reťazca Billa. Obsahuje 198 riadkov údajov o jednotlivých produktoch ponúkaných v letáku Billa a ďalej je rozdelený do šiestich stĺpcov: „Názov produktu“, „Merná jednotka“, „Jednotková cena“, „Akciová cena“, „Percentuálna zľava“ a „Bežná cena“. Tento súbor bol vytvorený s cieľom poskytnúť prehľad štruktúrovaných údajov z neštruktúrovanej formy vo forme PDF letáku. Údaje boli extrahované, vyčistené a automaticky transformované pomocou skriptu.



	A	B	C	D	E	F
1	Názov produktu	Merná jednotka	Jednotková cena	Akciová cena	Pôvodná cena	Percentuálna akcia
2	Kuracie rezne Domäsko v ochrannej atmosfére 1 kg			4,89	6,99	30%
3	Tami Tatranské maslo 82 %, 250 g	1 kg	8,76	2,19	3,99	45%
4	Bravčové plece bez kosti chladené cena za 1 kg			3,49	6,38	45%
5	Amundsen vodka 37,5 %, 0,7 l	1 l	10,7	7,49	11,77	36%
6	Nescafé Dolce Gusto vybrané druhy kapsuly, 1 bal.			4,39	6,23	29%
7	Madeta Jihočeské mlieko, 3,5 % trvanlivé, 1 l			0,79	1,29	38%
8	Banány voľné 1 kg			0,99	1,67	40%
9	Jahody debnička 900 g	1 kg	7,767	6,99	11,99	41%
10	Serena 1881 prosecco D.O.C. 0,75 l	1 l	6,48	4,86	9,72	50%
11	Zámocká šunka puttový predaj cena za 100 g	1 kg	9,9	0,99	1,28	22%
12	Kytica tulipánov V akciovej ponuke aj iné druhy kytíc tulipánov			3,99		
13	Dubajská čokoláda mliečna čokoláda plnená pistáciovo-kadayifovým krémom, 100 g	1 kg	39,9	3,99	5,99	33%
14	Donut Láska 58 g	1 kg	7,759	0,45	0,59	23%
15	Kytice ku Dňu žien rôzne druhy			1,99		
16	Kytica tulipánov a hyacintov 10 ks			5,99		
17	Črepníkové rastliny ku Dňu žien rôzne druhy			3,99		
18	Orchidea jednostonková farebná priemer črepníka 12 cm 1 ks			8,99	11,5	21%
19	Carte D'Or zmrzlina rôzne druhy 1000 ml	1 l	4,836	3,99		
20	Hubert Club šumivé víno rôzne druhy 0,75 l + Figaro Tatiana dezert rôzne druhy 172 g			9,99	13,22	24%
21	Lavazza Crema e Gusto Classico mletá káva 250 g	1 kg	17,16	4,29	5,94	27%
22	Marlenka torta medová/kakaová/škoricová 800 g	1 kg	11,238	8,99	14,29	37%
23	Château Topoľčianky ročníkový výber biele, červené víno rôzne druhy 0,75 l	1 l	5,987	4,49	6,13	26%
24	Raffaello, Ferrero Rocher čokoláda rôzne druhy 90 g	1 kg	22,111	199		
25	Toffifee dezert 125 g	1 kg	14,32	1,79		
26	Ferrero Rocher pralinky 200 g	1 kg	34,95	6,99		
27	Bumbu rum 40 % 0,7 l	1 l	47,129	32,99		

Dotazy a pripo... | Dotazy | Pripojenia | 1 dotaz | BILLA letak 44 | Načítané riadky: 198.

BILLA letak 44 (2) | BILLA letak 44 | Pripravený | Prístupnosť: nedostupné | 120 %

Obrázok 28. prepísanie Billa letáku do tabuľkového súboru Zdroj : (Vlastný zdroj)

4.4.2 COOP Jednota Supermarket

CSV súbor predstavuje štruktúrovaný dataset obsahujúci údaje o akciových ponukách produktov z letáku obchodného reťazca COOP Jednota Supermarket. Súbor je výsledkom spracovania neštruktúrovaných textových dát extrahovaných z PDF letáku a slúži na prehľadné zobrazenie ponúkaného sortimentu. Obsahuje 293 riadkov údajov a je rozdelený do piatich stĺpcov: „Akciová cena“, „Percentuálna zľava“, „Názov produktu“, „Jednotková cena“ a „Merná jednotka“. Tento formát umožňuje jednoduché porovnávanie cien, zliav a jednotkových cien jednotlivých produktov. Tento súbor bol vytvorený s cieľom poskytnúť prehľad štruktúrovaných údajov z neštruktúrovanej formy vo forme PDF letáku. Údaje boli extrahované, vyčistené a automaticky transformované pomocou skriptu, aby boli konzistentné a pripravené na ďalšie využitie, napríklad na webových stránkach na ich prezentáciu.

	A	B	C	D	E	F
1	Akciová cena	Percentuálna akcia	Názov produktu	Jednotková cena	Merná jednotka	
2	0,54	0,35	Magnesia 3 druhy 1,5 l	0,36	l	
3	2,79	0,36	Študentská pečať 3 druhy od 235 g	10,73	kg	
4	0,99	0,42	Banány 1 kg	0,99	kg	
5	2,99	0,22	Tradičná kvalita Olej slnečnicový 2 l	1,5	l	
6	2,09	0,37	Kuracie štvrte chladené 1 kg	2,09	kg	
7	0,89	0,4	DUO šunka najvyššej kvality 100 g	8,9	kg	
8	2,05	0,29	Jahody 250 g 1 bal.	8,2	kg	
9	1,79	0,3	Karfiol 1 ks			
10	1,45	0,42	Zeler bulva 1 kg	1,45	kg	
11	0,55	0,31	Reďkovka červená zväzok 1 ks			
12	1,79	0,28	Paradajky cherry strapec 500 g 1 bal.	3,58	kg	
13	1,69	0,33	Cuketa zelená 1 kg	1,69	kg	
14	1,19	0,21	Brokolica 500 g 1 bal.	2,38	kg	
15	0,69	0,29	Cibuľa červená 500 g 1 bal.	1,38	kg	
16	1,49	0,35	Čučoriedky 125 g 1 bal.	11,92	kg	
17	1,39	0,21	Cibuľka lahôdková zväzok 1 ks			
18	0,59	0,41	Stolové hrozno biele voľné 1 kg	3,39	kg	
19	1,15	0,28	Šalát taliansky mix 160 g 1 bal.	7,19	kg	
20	3,39	0,33	Zemiaky neskoré pranéžlté voľné 1 kg	0,85	kg	
21	3,39	0,33	Hrášok 500 g 1 bal.	6,78	kg	
22	1,19	0,41	Avokádo 1 ks			
23	1,39	0,21	Jablká červené ukladané 1 kg	1,39	kg	
24	0,75	0,33	Čingovská saláma 100 g	7,5	kg	
25	0,89	0,19	Dusená šunka špeciál 100 g	8,9	kg	
26	0,99	0	Oravská slanina 100 g	9,9	kg	
27	2,29	0,45	Pizza Piccolinis hlbokozmrazená 270 g	8,48	kg	

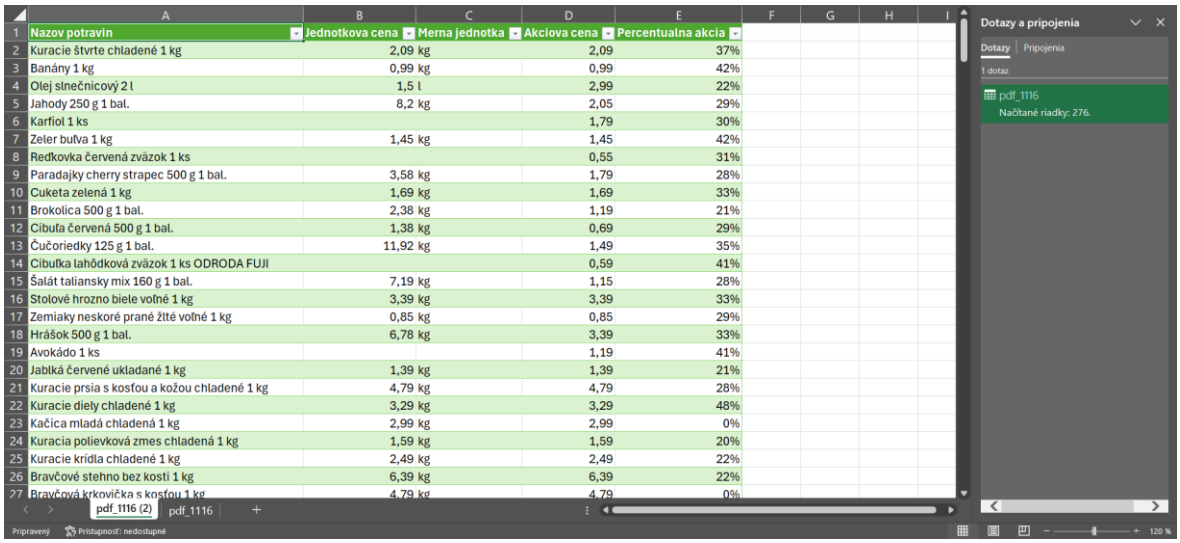
Dotazy a pripojenia | Dotazy | Pripojenia | 1 dotaz | pdf_1114 | Načítané riadky: 293.

pdf_1114 (2) | pdf_1114 | Pripravený | Prístupnosť: nedostupné | 120 %

Obrázok 29. prepísanie COOP Jednota Supermarket letáku do tabuľkového súboru Zdroj : (Vlastný zdroj)

4.4.3 COOP Jednota Tempo

Tabuľkový súbor predstavuje štruktúrovaný dataset obsahujúci údaje o akciových ponukách produktov z letáku obchodného reťazca COOP Jednota Tempo. Súbor je výsledkom spracovania neštruktúrovaných textových dát extrahovaných z PDF letáku a slúži na prehľadné zobrazenie sortimentu z akciového letáku. Obsahuje 276 riadkov údajov a je rozdelený do piatich stĺpcov: „Názov produktu“, „Jednotková cena“, „Merná jednotka“, „Akciová cena“ a „Percentuálna akcia“. Tento formát podporuje jednoduchú zobrazenie cien, zliav a jednotkových cien produktov. Tento súbor bol vytvorený s cieľom poskytnúť prehľad štruktúrovaných údajov z neštruktúrovanej formy vo forme PDF letáku. Údaje boli extrahované, vyčistené a automaticky transformované pomocou skriptu, aby boli konzistentné a pripravené na ďalšie využitie, napríklad na webových stránkach na ich prezentáciu.



	A	B	C	D	E
1	Nazov potravin	Jednotkova cena	Merna jednotka	Akciova cena	Percentualna akcia
2	Kuracie štvrte chladené 1 kg	2,09	kg	2,09	37%
3	Banány 1 kg	0,99	kg	0,99	42%
4	Olej slnečnicový 2 l	1,5	l	2,99	22%
5	Jahody 250 g 1 bal.	8,2	kg	2,05	29%
6	Karfiol 1 ks			1,79	30%
7	Zeler bulva 1 kg	1,45	kg	1,45	42%
8	Reďkovka červená zväzok 1 ks			0,55	31%
9	Paradajky cherry strapec 500 g 1 bal.	3,58	kg	1,79	28%
10	Cuketa zelená 1 kg	1,69	kg	1,69	33%
11	Brokolica 500 g 1 bal.	2,38	kg	1,19	21%
12	Cibuľa červená 500 g 1 bal.	1,38	kg	0,69	29%
13	Čučoriedky 125 g 1 bal.	11,92	kg	1,49	35%
14	Cibuľka lahôdková zväzok 1 ks ODRODA FUJI			0,59	41%
15	Šalát taliansky mix 160 g 1 bal.	7,19	kg	1,15	28%
16	Stolové hrozno biele voľné 1 kg	3,39	kg	3,39	33%
17	Zemiaky neskoré prané žlté voľné 1 kg	0,85	kg	0,85	29%
18	Hrášok 500 g 1 bal.	6,78	kg	3,39	33%
19	Avokádo 1 ks			1,19	41%
20	Jablká červené ukladané 1 kg	1,39	kg	1,39	21%
21	Kuracie prsia s kosťou a kožou chladené 1 kg	4,79	kg	4,79	28%
22	Kuracie diely chladené 1 kg	3,29	kg	3,29	48%
23	Kačica mladá chladená 1 kg	2,99	kg	2,99	0%
24	Kuracia polievková zmes chladená 1 kg	1,59	kg	1,59	20%
25	Kuracie krídla chladené 1 kg	2,49	kg	2,49	22%
26	Bravčové stehno bez kosti 1 kg	6,39	kg	6,39	22%
27	Bravčová krkovička s kosťou 1 kg	4,79	kg	4,79	0%

Obrázok 30. prepísanie COOP Jednota Tempo letáku do tabuľkového súboru Zdroj : (Vlastný zdroj)

5 Diskusia

Uvedená práca je zameraná na načítanie z rôznych internetových stránok, hlavne letákov potravinových reťazcov. Získané údaje napomôžu vyhľadávaniu najlepších cien jednotlivých produktov. V danej práci je vytvorený spôsob ukladania údajov vo forme veľkého množstva malých súborov, pričom v každom súbore sú údaje z jedného letáku a jedného reťazca. Druhým spôsobom ukladania, je ukladanie do veľkých súborov, napríklad podľa jednotlivých reťazcov, alebo do jedného veľkého súboru, kde sú údaje o všetkých reťazcoch a veličinách. Takýto spôsob uloženia údajov je možné spracovať pomocou metód a nástrojov z oblasti Big Data. Preto je možné danú prácu ďalej rozširovať s využitím teórie Big Data a následne využívať podrobnejšie údaje pre štatistické spracovanie a optimalizáciu výberu nákupného košíka.

Systém automaticky sťahuje PDF letáky z webových stránok reťazcov, ako sú Billa, Coop Jednota, Kaufland a Lidl, a extrahuje údaje o produktoch, vrátane cien, hmotností a zliav. Budúci rozvoj by mohol zahŕňať integráciu s relačnými databázami, ako je PostgreSQL, pre efektívne ukladanie a rýchle vyhľadávanie údajov, alebo implementáciu algoritmov strojového učenia na predikciu cenových výkyvov a automatickú kategorizáciu produktov. Ďalší krok by mohol súvisieť aj s vývoj rozhraní API pre integráciu s externými analytickými nástrojmi. Hlavné nedostatky systému spočívajú v obmedzenej schopnosti spracovať neštruktúrované PDF dokumenty a citlivosti na zmeny v štruktúre webových stránok.

Efektívnosť systému vyplýva z jeho schopnosti plne automatizovať získavanie a štruktúrovanie údajov, čím výrazne znižuje čas a náklady oproti manuálnemu spracovaniu.

Záver

V bakalárskej práce bol navrhnutý a podrobne opísaný systém na automatizované načítavanie a spracovanie údajov z PDF letákov obchodných reťazcov, so zameraním na vybrané komodity a ich ceny. Práca sa sústredila na analýzu neštruktúrovaných dát získaných z letákov reťazcov Lidl, COOP Jednota, Kaufland a Billa, ich transformáciu do štruktúrovanej podoby vo formáte CSV. Tento proces mal za cieľ uľahčiť spotrebiteľom porovnávanie ponúk a podporiť efektívnejšie rozhodovanie pri nákupe, ako aj poskytnúť základ pre ďalšie analýzy ponuky jednotlivých reťazcov.

V priebehu práce boli identifikované kľúčové výzvy spojené so spracovaním neštruktúrovaných dát. Medzi hlavné problémy patrilo nekonzistentné formátovanie textu, chyby vzniknuté pri extrakcii z PDF a potreba manuálnej kontroly na zabezpečenie presnosti údajov. Na spracovanie boli použité regulárne výrazy v programovacom jazyku Python, ktoré umožnili automatizovanú identifikáciu a štruktúrovanie položiek, ako sú názvy produktov, hmotnosti, objemy, bežné a akciové ceny, percentuálne zľavy a jednotkové ceny.

Navrhnutý systém preukázal schopnosť efektívne spracovať neštruktúrované dáta a transformovať ich do použiteľnej štruktúrovanej podoby, čím splnil stanovené ciele práce. Napriek tomu existuje priestor na ďalšie zlepšenia. Ďalšom kroku by mohol byť rozšírený o ďalšie obchodné reťazce, komodity alebo analýzy, napr. sledovanie sezónnych trendov v cenách. Práca tak poskytuje pevný základ pre budúce rozšírenia a zlepšenia v oblasti automatizovaného spracovania dát z letákov, čím prispieva k zefektívneniu analýzy ponuky na trhu.

Zoznam literatúry

(Koolwal, 2024). KOOLWAL, Manthan. *How To Extract Data From Any Website* [online]. Dostupné na: https://www.scrapingdog.com/blog/how-to-extract-data-from-website/

(Potrimba, 2023). POTRIMBA, Petru. What is Optical Character Recognition (OCR)? [online]. [cit. 2024-12-20]. Dostupné na: https://blog.roboflow.com/what-is-optical-character-recognition-ocr/

(Smith, 2023). SMITH, Craig S.. *What Is OCR (Optical Character Recognition) Technology?* [online]. Dostupné na: https://www.forbes.com/sites/technology/article/what-is-ocr-technology/

(Russell, 2023). RUSSELL, John. *Optical Character Recognition (OCR): An Introduction* [online]. [cit. 2024-12-20]. Dostupné na: https://guides.libraries.psu.edu/OCR

(Gregersona, 2025). GREGERSEN, Erik. *OCR* [online]. Dostupné na: https://www.britannica.com/technology/OCR

(Boesch, 2023). BOESCH, Gaudenz. *Optical Character Recognition (OCR)* [online]. [cit. 2024-12-20]. Dostupné na: https://viso.ai/computer-vision/optical-character-recognition-ocr/

(Konovalchuk, 2024). KONOVALCHUK, Nikolay. *OCR algorithms: types, operation & best solutions* [online]. [cit. 2024-12-17]. Dostupné na: https://www.itransition.com/computer-vision/ocr-algorithm

(Hamad and Kaya, 2016). HAMAD, Karez Abdulwahhab – KAYA, Mehmet. A Detailed Analysis of Optical Character Recognition Technology. IN *International Journal of Applied Mathematics Electronics and Computers* [online]. 1.12.2016, č. 4, s. 244-249. ISSN 2147-8228. Dostupné na: https://dergipark.org.tr/tr/download/article-file/236939

(Perez, 2023). PEREZ, Martin. *What is Web Scraping and What is it Used For?* [online]. [cit. 2025-1-2]. Dostupné na: https://www.parsehub.com/blog/what-is-web-scraping/

(Barton, 2024). BARTON, David. *What is web scraping?* [online]. [cit. 2024-12-20]. Dostupné na: https://blog.apify.com/what-is-web-scraping/

(Pawłowski, 2025). PAWŁOWSKI, Piotr. *Čo je web scraping a ako ho využiť v podnikaní?* [online]. [cit. 2024-12-17]. Dostupné na: https://firmbee.sk/ako-pouzivat-web-scraping-v-podnikani

(Mitchell, 2018). MITCHELL, Ryan. *Web Scraping with Python* [elektronický zdroj]. 2.vyd. Sebastopol: O'Reilly Media, Inc. 4, 2018. 306 s. ISBN 978-1-491-98557-1. Dostupné na:https://edu.anarcho-copy.org/Programming%20Languages/Python/Web%20Scraping%20with%20Python,%202nd%20Edition.pdf

(OpenAI, 2023). *ChatGPT* (Mar 14 version) [Large language model]. https://chat.openai.com/chat

(Gupta and Bagchi, 2024). GUPTA, Pramod – BAGCHI, Anupam. Essentials of Python for Artificial Intelligence and Machine Learning [elektronický zdroj]. 1 vyd. Cham: Springer Cham. 2, 2024. 509 s. ISBN 978-3-031-43725-0. Dostupné na: https://link.springer.com/chapter/10.1007/978-3-031-43725-0_5

(Sloan, 2016). SLOAN, Kelly. *Python, PyGame and Raspberry Pi Game Development* [elektornický zdroj]. 1. vyd. Berkeley: Apress, 2016. 198 s. ISBN 978-1-4842-2517-2 Dostupné na: https://link.springer.com/chapter/10.1007/978-1-4842-2517-2_2

(Makka, 2023). MAKKA, Shanthi et al. A GUI Based Application for PDF Processing Tools Using Python & CustomThinker. In *International Journal for Research in Applied Science & Engineering Technology* [online]. Hyderabad, Vardhaman College of Engineering, 1.2023, č. 4, s. 1613-1618. ISSN 2321-9653. Dostupné na: https://d1wqtxts1xzle7.cloudfront.net/104365877/a-gui-based-application-for-pdf-processing-tools-using-python-and-customtkinter-libre.pdf?1689738693=&response-content-disposition=inline%3B+filename%3DA_GUI_Based_Application_for_PDF_Processi.pdf&Expires=1745423057&Signature=QRnBhn41~jnyHgDbm5r6~Nn1ekhxGXOC~yOTfQyTkkxPhSiBgmAJhWrTbMFXLA7CaPCxdREb5q094ljdYHf1rJulfNPWcDARrpdQb-VUd1sQoufh8bowfw7fCSxiH8gzEu6vDZ-IwCqXN-KQP-wfZWRDHucvjgLWK57ujgrhPyDHkl5j6docbJhUHch3THbbmZpNTvS-SjhOQksZXyQkpZenyT9uIT-mMsU2RyexXyDgYAWQiMsI4FvB~i9l61NfWAadnjvTGrjZgji3S6rTQct7JgJWqs8RKASn34dv4PKGrGBEg~B47sUZ1VS2u0B67Of2pSabtUNQkEGDRzN80A__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

(Chandra and Varanasi, 2015). CHANDRA, Rakesh Vidya – VARANASI, Bala Subrahmanyam. *Python Requests Essentials* [elektornický zdroj]. 1. vyd. Birmingham: Published by Packt Publishing Ltd., 2015. 134 s. ISBN 978-1-78439-541-4. Dostupné na: https://www.programmer-books.com/wp-content/uploads/2019/04/Python-Requests-Essentials.pdf