



ÓBUDAI EGYETEM
ÓBUDA UNIVERSITY

**NEUMANN JÁNOS
INFORMATIKAI KAR**



SZAKDOLGOZAT

**OE-NIK
2025**

**Hallgató neve:
Hallgató törzskönyvi száma:**

**Füzér Levente András
T009148/FI12904/N**

SZAKDOLGOZAT FELADATLAP

Hallgató neve: **Füzér Levente András**
Törzskönyvi száma: T009148/FI12904/N
Neptun kódja: 

A dolgozat címe:

Követői adatok elemzése szociális platformokon big data eszközökkel
Analyzing follower data on social platforms using big data tools

Intézményi konzulens: Sipos Miklós László
Külső konzulens:

Beadási határidő: 2024. december 15.

A záróvizsga tárgyai: Számítógép architektúrák
Big Data és üzleti intelligencia
specializáció

A feladat

A feladat egy olyan adatfeldolgozó rendszer elkészítése, amely képes következtetéseket levonni különböző szociális platformokon lévő profilok analitikái alapján. A rendszer elkészítése előtt járjon utána, hogy milyen ehhez hasonló adatfeldolgozó analitikai rendszerek léteznek, ezek milyen funkciókkal rendelkeznek, és hogy ehhez képest a saját programja pontosan milyen feladatot akar ellátni. A program legyen képes saját adatgyűjtésre Selenium vagy Webscraper segítségével, valamint legyen képes általunk betöltött adatokból új információkat kinyerni, majd ezt olyan formába rendezni, hogy (esetleg PowerBi vagy más eszközök segítségével) képesek legyünk vizualizációkat, riportokat készíteni az adatokból. Az elemzés egy letisztult, felhasználók számára intuitív kezelési felületen történjen. A program célja, hogy az általunk begyűjtött, és a szociális platformok által nyújtott analitikai adatok alapján a program felhasználói olyan következtetéseket tudjanak levonni, amelyekkel kedvezőbb irányba tudják vinni a platformon folytatott tevékenységeiket, jobban megismerhessék a követőiket, és optimálisan alakíthassák a tartalom-stratégiájukat. Az elkészült rendszert tesztelje.

A dolgozatnak tartalmaznia kell:

- a megoldandó feladat pontos leírását és részletes elemzését,
- magas szintű szerkezeti mintát az alkalmazásról,
- az egyes funkciók megvalósításához felhasználható lehetséges technológiákat,
- az elkészített rendszer tervét, blokkdiagramját,
- a megvalósítás menetét, felmerülő problémákat és az azokra adott megoldást,
- a tesztelési tervet, a tesztelés módszertanát és a tesztelés eredményeit,
- az elkészült szoftver felhasználási lehetőségeit,
- az eredmények bemutatását és értékelését a teljes munkára vonatkozóan.



Vámosy Zoltán

Dr. habil. Vámosy Zoltán
intézetigazgató

A szakdolgozat elévülésének határideje: **2026. december 15.**
(ÓE HKR 54.§ (10) bekezdés szerint)

A dolgozatot beadásra alkalmasnak tartom:

.....
külső konzulens

Sipos Zoltán
.....
intézményi konzulens



HALLGATÓI NYILATKOZAT

Alulírott Füzér Levente András (DH8G3K) kijelentem, hogy a szakdolgozat / diplomamunka saját munkám eredménye, a felhasznált szakirodalmat és eszközöket azonosíthatóan közöltem. Az elkészült szakdolgozatomban / diplomamunkámban található eredményeket az egyetem és a feladatot kiíró intézmény saját céljára térítés nélkül felhasználhatja.

Budapest, 20 24.12.09.....

Füzér Levente
.....
hallgató aláírása



ÓBUDAI EGYETEM
ÓBUDA UNIVERSITY

Neumann János Informatikai Kar

KONZULTÁCIÓS NAPLÓ

Hallgató neve:

FÜZÉR LEVENTE ANDRÁS

Neptun Kód:



Tagozat:

NAPPALI TAGOZAT

Telefon:

Levelezési cím (pl.: lakcím):



Szakdolgozat címe magyarul:

KÖVETŐI ADATOK ELEMZÉSE SZOCIÁLIS PLATFORMOKON BIG DATA ESZKÖZÖKKEL

Szakdolgozat címe angolul:

ANALYZING FOLLOWER DATA ON SOCIAL PLATFORMS USING BIG DATA TOOLS

Intézményi konzulens:

SIPOS MIKLÓS LÁSZLÓ

Külső konzulens:

-

Alk.	Dátum	Tartalom	Aláírás
1.	2024.02.19.	Félév teendőinek, menetrendjének és mérföldköveinek ismertetése.	
2.	2024.03.11.	Hasonló rendszerek elemzése.	
3.	2024.04.15.	Irodalomkutatás, annak összegzése valamint konzekvenciák meghatározása.	
4.	2024.05.06.	Tervezés, követelmények specifikációja. Véglegesítés. Leadással kapcsolatos teendők egyeztetése.	

A Konzultációs naplót összesen 4 alkalommal, az egyes konzultációk alkalmával kell láttamoztatni bármelyik konzulenssel.

A hallgató a Szakdolgozat I. tantárgy követelményét teljesítette, beszámolóra bocsátható.

A konzulens által javasolt érdemjegy: 4

Intézményi konzulens

Budapest, 2024.05.17.



ÓBUDAI EGYETEM
ÓBUDA UNIVERSITY

Neumann János Informatikai Kar

KONZULTÁCIÓS NAPLÓ

Hallgató neve: FÜZÉR LEVENTE ANDRÁS Neptun Kód: [REDACTED] Tagozat: NAPPALI

Telefon: [REDACTED] Levelezési cím (pl.: lakcím): [REDACTED]

Szakdolgozat címe magyarul:

KÖVETŐI ADATOK ELEMZÉSE SZOCIÁLIS PLATFORMOKON BIG DATA ESZKÖZÖKKEL

Szakdolgozat címe angolul:

ANALYZING FOLLOWER DATA ON SOCIAL PLATFORMS USING BIG DATA TOOLS

Intézményi konzulens: SIPOS MIKLÓS LÁSZLÓ Külső konzulens: -

Alk.	Dátum	Tartalom	Aláírás
1.	2024.09.29.	Fejlesztés tervezett menetének revíziója, szakmai egyeztetés, fejlesztési teendők megkezdése.	
2.	2024.10.20.	Fejlesztés állapotának és haladásának áttekintése. Technikai problémák és módosítások egyeztetése.	
3.	2024.11.17.	Felmerülő problémák egyeztetése és megoldása. Tesztelési fázis megindítása.	
4.	2024.12.01.	Dokumentáció véglegesítése. Leadással kapcsolatos teendők egyeztetése.	

A Konzultációs naplót összesen 4 alkalommal, az egyes konzultációk alkalmával kell láttamoztatni bármelyik konzulenssel.

A hallgató a Szakdolgozat II. tantárgy követelményét teljesítette, védeésre bocsájtható.

A konzulens által javasolt érdemjegy: 5

Intézményi konzulens

Budapest, 2024. 12. 06.

Absztrakt

A szociális platformok megjelenésük óta rohamos fejlődésen mentek keresztül. Az emberek egyre több időt töltenek az interneten, ahol állandó tevékenységeikkel számos adatot generálnak. Ez a digitális tér különösen érdekes, mivel nem csupán információt közvetít, hanem egy képet ad emberi viselkedésről és kapcsolatrendszerről a virtuális világban. A szociális platformok működésének mélyebb megértése érdekes számomra, mivel ezek a platformok jelentős szerepet játszanak az életünkben.

Bárkiben, aki rendszeresen tölt időt az interneten, felmerülhet a kérdés, hogy vajon mitől lesz egy internetes videó sikeres, milyen tényezők játszanak közre egy videó vagy poszt eredményességében, hatásosságában. Tartalomgyártóként vagy vállalkozóként milyen stratégiákat lehet alkalmazni annak érdekében, hogy a tartalmak minél több embert elérjenek. Ebben a szakdolgozatban ezekre a kérdésekre próbálok minél pontosabb választ találni, egy saját fejlesztésű analitikai rendszer segítségével.

A szakdolgozatom célja egy olyan rendszer fejlesztése, amely képes nagy mennyiségű adat automatizált begyűjtésére szociális platformokról. Ezeket az adatokat képes egységes formátumra hozni, feldolgozni, majd elemezni. Végül a levont következtetéseket képes könnyen értelmezhető vizualizációkkal és statisztikákkal alátámasztani.

A szakdolgozatom keretein belül több mint 20 ezer rövid formátumú videó analitikai adatát elemeztem, és hasonlítottam össze egymással. Az elemzéseim során rengeteg érdekességet és szabályszerűséget fedeztem fel.

Abstract

Since their emergence, social platforms have undergone rapid development. People spend more and more time on the internet, where their constant activities generate a wealth of data. This digital space is particularly intriguing because it not only conveys information but also provides insights into human behavior and social connections in the virtual world. Understanding the workings of social platforms on a deeper level is interesting to me, as these platforms play a significant role in our lives.

Anyone who spends regular time on the internet might wonder what makes an online video successful and what factors contribute to the effectiveness and impact of a video or post. As a content creator or entrepreneur, what strategies can be employed to ensure content reaches as many people as possible? In this thesis, I aim to answer these questions as accurately as possible, using a self-developed analytical system.

The goal of my thesis is to develop a system capable of automating the collection of large volumes of data from social platforms. This system can standardize, process, and analyze the data. Finally, it will present conclusions supported by easily interpretable visualizations and statistics.

As part of my thesis, I analyzed and compared the analytical data of over 20,000 short-form videos. Throughout my analyses, I discovered numerous interesting patterns and regularities.

TARTALOMJEGYZÉK

TARTALOMJEGYZÉK	9
1. BEVEZETÉS	12
1.1. Témaválasztás	12
1.2. Célok	12
1.3. A dolgozat felépítése	12
1.3.1. Irodalomkutatás	12
1.3.2. Követelmény specifikáció	12
1.3.3. Tervezés	13
1.3.4. Fejlesztés	13
1.3.5. Tesztelés	13
1.3.6. Elemzés	13
2. IRODALOMKUTATÁS	14
2.1. Hasonló rendszerek	14
2.1.1. Adatgyűjtő rendszerek definiálása	14
2.1.2. Web scraping	14
2.1.3. Web Scraper	15
2.1.4. Selenium	16
2.1.5. Scrapy	16
2.1.6. BeautifulSoup	16
2.1.7. Szociális platformok API-ai	17
2.1.8. Adatgyűjtő rendszerek összehasonlítása	18
2.1.9. Adatfeldolgozó rendszerek	20
2.1.10. Korábban készített adatfeldolgozó rendszerem bemutatása	20
2.1.11. Adatvizualizációs rendszerek	21
2.1.12. Jupyter	21
2.1.13. Python pandas és NumPy könyvtárakkal	21
2.1.14. Microsoft Power BI	22
2.1.15. Az adatvizualizációs rendszerek összehasonlítása	23
2.2. A saját rendszer programlistájának összegzése	24
2.2.1. Az adatgyűjtő program kiválasztása	24
2.2.2. Az adatfeldolgozó program kiválasztása	24
2.2.3. Az adatvizualizációhoz használt program kiválasztása	24
2.3. Kapcsolódó témák	25

2.3.1. A rendszer által vizsgált tartalmak specifikálása és a rövid formátumú videók ismertetése	25
2.3.2. Begyűjthető adatok elemzése az egyes platformok profiljain, és ezek összehasonlítása	25
2.3.3. Begyűjthető adatok elemzése az egyes platformok videóin, és ezek összehasonlítása	26
2.3.4. Mitől lesz sikeres egy rövid formátumú videó, miben lehet mérni a hatásosságot?	26
2.3.5. Kiknek és miért fontos a tudatos tartalomstratégia?	27
2.3.6. Mit jelent az adatvizualizáció, és miért jelentős?	28
2.3.7. Az adatvizualizáció legelterjedtebb típusai	28
2.3.8. Az adatvizualizáció lehetséges hátrányai	30
2.3.9. Az adatvizualizáció alapelvei	30
2.3.10. A web scraping jogszerűsége	32
3. KÖVETELMÉNY SPECIFIKÁCIÓ	33
3.1. Adatgyűjtési folyamat	33
3.2. Adattisztítási folyamat	33
3.3. Tesztelés	33
3.4. Elemzés	33
3.5. Vizualizáció	33
4. TERVEZÉS	34
4.1. Rendszerterv	34
4.1.1. A rendszerterv ábra	34
4.1.2. Az adatgyűjtő alrendszer ismertetése	35
4.1.3. Az adatfeldolgozó alrendszer ismertetése	35
4.1.4. Az elemzési alrendszer adatbázisának felépítése	36
4.1.5. A vizualizációs alrendszer ismertetése	37
4.2. Funkciólista	38
4.3. A fejlesztés tervezett menete	39
4.3.1. A fejlesztés időbeosztásának magyarázata	39
5. FEJLESZTÉS	41
5.1. Adatgyűjtés	41
5.1.1. A Web Scraper felkonfigurálásakor felmerülő problémák	41
5.2. Az adatok transzformálása	42
5.2.1. A projektek létrehozása és előkészületek	42
5.2.2. Single-responsibility principle	42
5.2.3. A program felépítése	42
5.2.4. A transzformációk ismertetése	44
6. TESZTELÉS	46

6.1. Az adatok integritásának ellenőrzése	46
6.2. Unit tesztelés	46
6.2.1. A unit tesztek működése	47
6.2.2. A joindate oszlopot átalakító metódus tesztelése	47
6.2.3. A totalviews oszlopot átalakító metódus tesztelése	47
6.2.4. A subscribers oszlopot átalakító metódus tesztelése	48
6.3. A tesztek eredményeinek összefoglalása	49
7. ELEMZÉS	51
7.1. Az elemzések ismertetése	52
7.1.1. Milyen gyakran érdemes feltölteni?	52
7.1.2. Számít egy videó címének hossza?	53
7.1.3. Számít egy videó címének tartalma?	53
7.1.4. Mi az az engagement rate, és miért jelentős?	54
7.1.5. Számít a biográfia hossza?	56
7.1.6. A „hangok” és zenék használata a videókban	57
7.1.7. A hashtagek definíciója és célja	58
7.1.8. A hashtagek használatának előnyei	58
7.1.9. A hashtagek használatának esetleges hátrányai	58
7.1.10. A hashtagek hatása a megtekintésekre és like-okra	58
7.1.11. Az naprakészség és a trendek követésének jelentősége	59
7.1.12. A pandémia hatása YouTube Shorts videók nézettségére	59
8. ÉRTÉKELÉS	61
8.1.1. A szakdolgozat eredményeinek kiértékelése	61
8.1.2. Egyéb akadályok, melyek kerülőmegoldást igényeltek	61
9. TOVÁBBFEJLESZTÉSI LEHETŐSÉGEK	62
9.1.1. Az adathalmaz növelése	62
9.1.2. Az elemzés és vizualizáció automatizálása	62
9.1.3. A videó tartalmát elemző rendszer	62
9.1.4. Különböző szociális platformokat összehasonlító rendszer	62
9.1.5. Tartalomgyártókat egymással összehasonlító rendszer	62
10. ÖSSZEGZÉS	63
10.1. Záró gondolatok	65
IRODALOMJEGYZÉK	66
ÁBRAJEGYZÉK	69
TÁBLAJEGYZÉK	70

1. BEVEZETÉS

1.1. Témaválasztás

Az internetes videók és ezek sikerességének optimalizálása régóta érdekel. Gyerekkorom óta rendszeresen nézek YouTube-on és más platformokon videóesszét és egyéb formátumú videókat különböző témákban. Rengeteg időt töltöttem a tartalomgyártás folyamatának megismerésével, mivel érdekelt, hogy egyes videók miért sikeresebbek a többinél.

A big data szakirányon sok új adatfeldolgozási eszközzel és elemzési módszertannal gazdagodtam, és ezeket a szakdolgozat keretein belül egy olyan témában akartam felhasználni, amellyel szívesen foglalkozom.

1.2. Célok

A célom az volt, hogy az egyetemen tanultakat felhasználva kialakítsak egy olyan rendszert, amely képes nagy adatmennyiségeket automatizáltan begyűjteni. Ezeket képes megtisztítani, azaz riportkész állapotba helyezni, majd ezután különböző elemzések segítségével érdekes és hasznos szabályszerűségeket felfedezni.

1.3. A dolgozat felépítése

1.3.1. Irodalomkutatás

Ebben a részben már létező, a fejlesztendő programhoz hasonló rendszereket hasonlítok össze egymással. Elemzem ezeket a rendszereket különböző szubjektív szempontok alapján, majd táblázatokkal szemléltetem az egyes rendszerek előnyeit, hátrányait. Emellett ezekben a rendszerekben olyan funkciókat, karakterisztikákat keresek, amelyeket a saját programomba szeretnék majd beleépíteni. Mivel a tervezett rendszer három különböző funkciót ellátó részből áll, ezért mindhárom alrendszert külön részletezem ebben a részben.

Itt fejtem még ki az egyéb, szakdolgozathoz releváns témákat. Részletesebben tárgyalom, hogy az egyes szociális platformokon milyen adatok elérhetőek adatgyűjtésre. Ismertetem, hogy milyen típusú tartalmakat fogok elemezni az egyes platformokon. Ezek mellett néhány olyan témát érintek, amelyek alapvető ismeretét fontosnak tartom ahhoz, hogy a fejlesztendő rendszerem működését és célját érthetővé tegyem.

1.3.2. Követelmény specifikáció

Ebben a rövid fejezetben pontosítom a rendszeremmel szemben állított elvárásokat. A fejlesztés és tesztelés részeket az itt felállított standardok szerint végzem.

1.3.3. Tervezés

Ebben a részben ismertetem a rendszer fejlesztésének ütemtervét, itt határozom meg a rendszertervet, a funkciólistát, valamint az adatbázis táblarajzát. Az egyes komponensek működését is részletesen elmagyarázom.

1.3.4. Fejlesztés

Ebben a fejezetben dokumentálom az adatgyűjtő program felkonfigurálását, a gyűjtendő adatok kiválasztásának folyamatát, az adattisztító program fejlesztésének folyamatát, valamint az adattranszformációk menetét és a vizualizációk előkészítését.

1.3.5. Tesztelés

A tesztelés részben megbizonyosodom a teljes rendszerem helyes működéséről, unit tesztek és egyéb tesztelési módszerek segítségével.

1.3.6. Elemzés

Az elemzés részben több mint 20 ezer YouTube videó analitikai adatait elemzem. A meglátásaimat részletesen kifejtem, majd ezeket számos külső forrással és statisztikai adatokkal alátámasztom.

2. IRODALOMKUTATÁS

2.1. Hasonló rendszerek

Először is fontos leszögeznünk, hogy pontosan milyen rendszereket értünk hasonló rendszerek alatt. A tervezett analitikai rendszer több alrendszerre bontható le. A folyamat az adatok begyűjtésével kezdődik, majd ezek az adatok feldolgozásra kerülnek. A feldolgozást pedig több elemzés követi, melyek eredményeit vizualizációkkal illusztrálom. Az alábbi alfejezetekben tehát adatgyűjtő, adattisztító és adatvizualizáló rendszereket fogok összehasonlítani. Ezt követően pedig kiválasztom hogy melyeket fogom alkalmazni a saját rendszeremben. A végleges programlistát a 2.2-es összegző fejezetben foglaltam össze.

2.1.1. Adatgyűjtő rendszerek definiálása

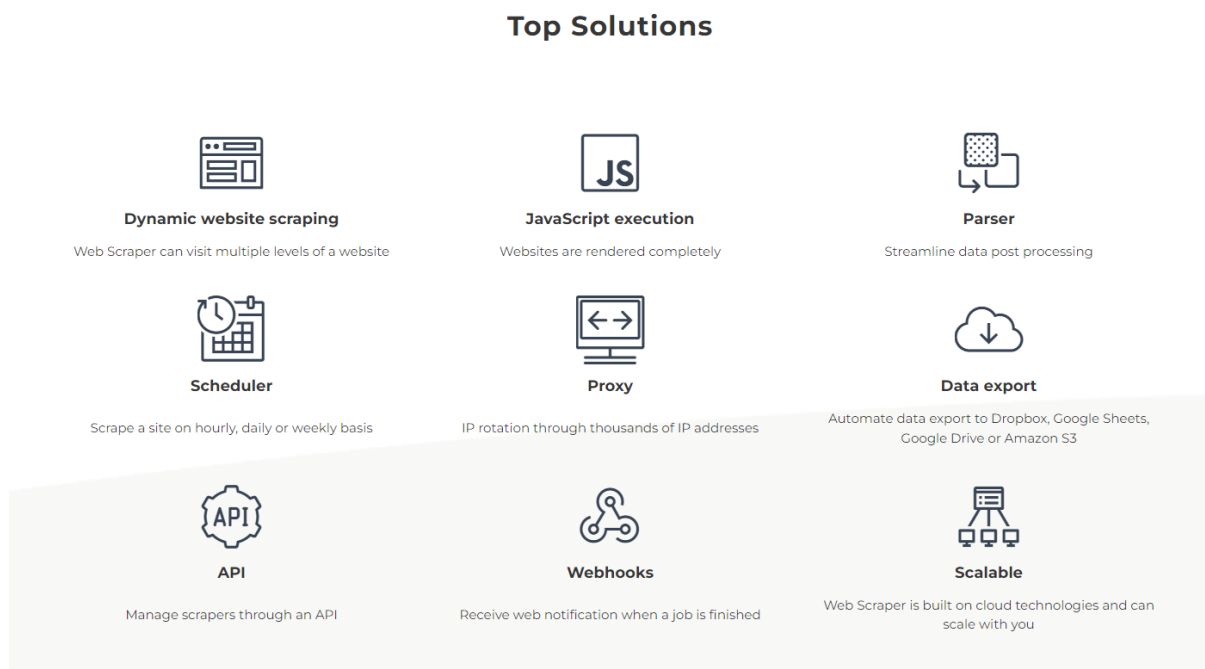
A begyűjtött adatok mennyisége és típusa határozza meg, hogy milyen részletes elemzést tudunk végezni. Minél nagyobb mennyiségű, és minél több típusú adatot tudunk begyűjteni, annál pontosabb képet alkothatunk a felhasználókról, és a viselkedésükről. A szociális platformokon rengeteg adat áll rendelkezésünkre, viszont ezek begyűjtése automatizálás nélkül szinte lehetetlen. Az adatgyűjtő programot tehát az alapján választottam ki, hogy ezt a feladatot melyik tudja a legegyszerűbben és leghatékabban elvégezni.

2.1.2. Web scraping

[1] A szociális platformok felhasználói hatalmas mennyiségű adatot generálnak a tevékenységeikkel. Ez a nagy mennyiségű adat minden weboldalon megtalálható, és különböző alkalmazások számára elérhető. Az adatokat a keresőmotoroknak köszönhetően manuálisan is be lehet gyűjteni, viszont ez a megközelítés nem skálázható nagyobb adatmennyiség esetén. Számos automatizált megközelítés létezik a webes adatok kinyerésére, ezek egyike a web scraping. [2] A web scraping egy olyan adatgyűjtési technika, amely weboldalokról nyer ki adatokat, majd lementi őket egy fájlba vagy adatbázisba, további feldolgozás céljából.

2.1.3. Web Scraper

[3] A Web Scraper egy könnyen használható, ingyenes böngésző bővítmény, amelynek letisztult felületén könnyedén lehet felkonfigurálni a scraper-eket, amelyek automatikusan begyűjtik az adott weboldal adatait. Az adatok egy CSV vagy Excel formátumú táblázatként letölthetők.



2.1-es ábra: A Web Scraper fontosabb funkciói

- A bármely böngészőbe installálható bővítmény teljesen ingyenes. Az ingyenes verzió tartalmaz JavaScript integrációt, CSV és XLSX exportálási lehetőséget.
- Rendelkezik emellett egy fizetős verzióval, amely az előbb említett formátumok mellett JSON formátumban is képes exportálni. A fájl export mellett tud Dropbox-ba, Google Sheets-be, Google Drive-ba és Amazon S3-ba is exportálni.
- A manuálisan indítható scraping mellett rendelkezik időzített scraping-gel, valamint képes egyszerre három adatgyűjtési folyamat párhuzamos futtatására.
- A fizetős verzió igénybe vehető egy hetes próbaidőre, egy alkalommal.

2.1.4. Selenium

[4] A Selenium egy Python alapú automatizálható eszköz, amely számos feladatot képes ellátni. Az egyik ilyen feladat a web scraping.

- A Selenium használata ingyenes, hiszen egy nyílt forráskódú szoftver, azaz a felhasználók személyes és kereskedelmi célokra is szabadon használhatják az Apache License 2.0 szerint.
- Rendelkezik JavaScript integrációval, ezért kifejezetten jól kezeli a JavaScript-el készült dinamikus weboldalakot.
- Képes böngésző automatizálásra, azaz képes kérdőíveket kitölteni előre megadott adatokkal, gombokat nyomkodni, vagy letekerni oldalakon. Ezzel a funkcióval olyan adatokat is el tudunk érni, amelyeket statikus HTML elemzéssel nem tudnánk.
- Működik Chrome-mal, Firefox-al, Safari-val és Edge-el is
- Rendelkezik párhuzamos futtatási opcióval, azaz egyszerre több adatgyűjtési folyamatot tud véghez vinni.

2.1.5. Scrappy

[5] A Scrappy egy nyílt forrású kollaboratív keretrendszer weboldalakról történő adatgyűjtéshez. Hatalmas fejlesztői közösséggel rendelkezik, több mint 43.000 GitHub-os csillagozással és 18.000 StackOverflow-os kérdéssel. Ez azért jelentős, mert a fejlesztés folyamán felmerülő problémák megoldása és kérdések megválaszolása sokkal egyszerűbb, ha egy aktív fejlesztői közösséghez tudunk fordulni.

- Széleskörű az operációs rendszer kompatibilitása. Létezik Windows-os, Linux-os, Mac-es és BSD-s verziója is
- Nagyméretű, aktív fejlesztői közösség
- Könnyen integrálható Python alapú rendszerekbe

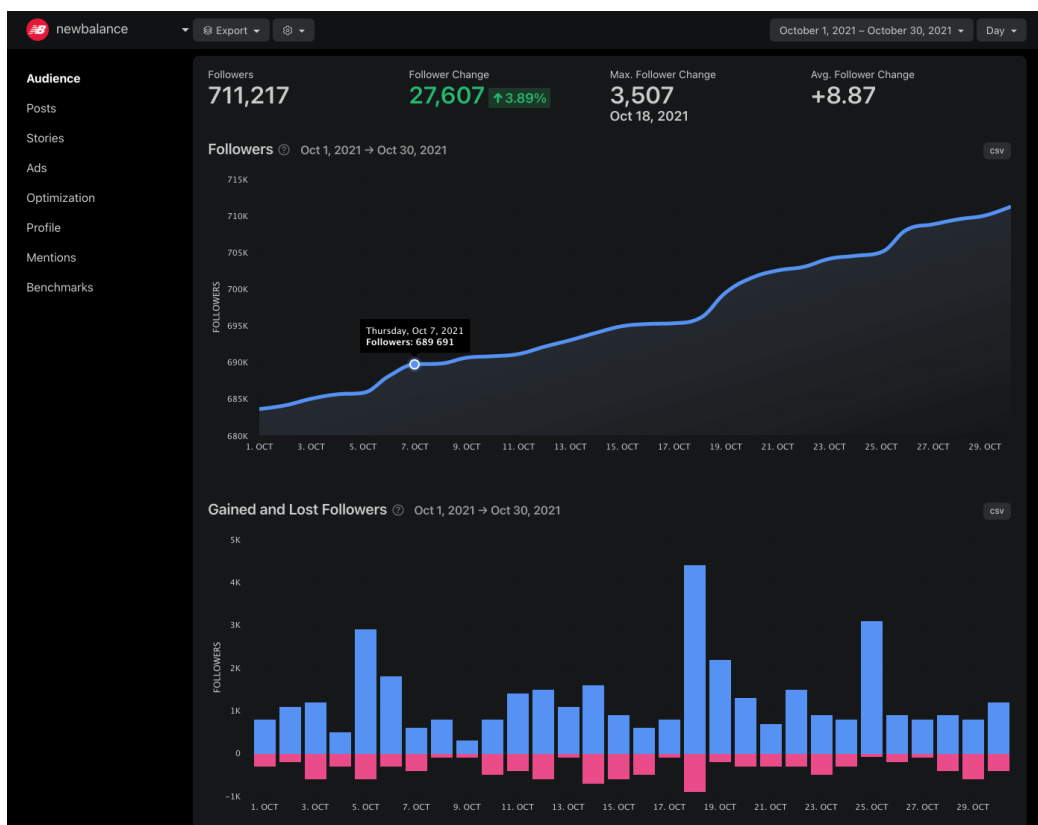
2.1.6. BeautifulSoup

[6] A BeautifulSoup egy Python csomag amely HTML és XML fájlokból való adatkinyerésre lett fejlesztve.

- Kiválóan működik HTML és XML fájlformátumokkal, mivel specifikusan fa adatstruktúrák elemzésére hozták létre
- Működik a Python beépített HTML parser-ével, de harmadik féltől származó parser-ekkel is kompatibilis

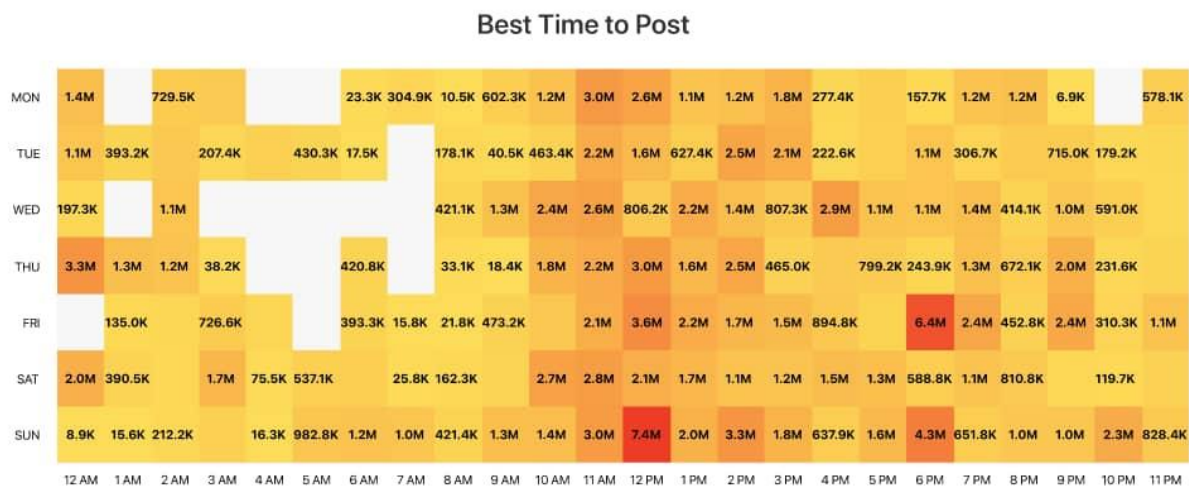
2.1.7. Szociális platformok API-ai

A web scraping mellett van lehetőség egyenesen a közösségi média platformoktól megszerezni a szükséges adatokat. [7] Egy példa az ilyen megvalósításra a Minter.io által fejlesztett Instagram Analytics Tool, ami az Instagram Graph API segítségével képes teljesítmény mérőszámokat lekérni. Ezeket az adatokat egy könnyen kezelhető, intuitív felhasználói felületen lehet megtekinteni. Rendelkezik beépített vizualizációs eszközökkel is, amelyekkel grafikonokon, időre lebontva lehet elemezni a követők tevékenységét. (2.2-es ábra)



2.2-es ábra: Az Instagram Analytics Tool vizualizációs felülete

A felhasználók tevékenységeinek idejét is le lehet kérni, majd ezt egy táblázatban vizualizálva létre lehet hozni egy aktivitási hőtérképet, amely megmutatja, hogy melyik napokon, milyen órákban a legaktívabbak a követők. A lehető legtöbb felhasználó elérésének érdekében ilyenkor érdemes tartalmakat feltölteni. A 2.3-as ábrán látható, hogy hogy néz ki egy ilyen hőtérkép.



2.3-as ábra: Az Instagram Analytics Tool „Best Time to Post” funkciója

A hátulütője az ilyen API-okra épülő analitikai rendszereknek, hogy csak egy adott fiók adatait lehet megtekinteni velük, azaz az analitikák csak az API-ra regisztrált fiók követőiről nyújtanak információt. Emellett, az API használatát kérvényezni kell, és csak az adott szociális platform hozzájárulásával lehet felhasználni fejlesztésre. [8] Ez a kérvényezési folyamat sok időt vesz igénybe, és sok kritériumnak meg kell felelni. Például a Tiktok Research API használatához részletes leírást kell adni magadról, az intézményedről, és az adott kutatási területről, majd csak egy 3-4 hetes kiértékelés után lehet engedélyt szerezni.

2.1.8. Adatgyűjtő rendszerek összehasonlítása

Az összehasonlítás általam választott, szubjektív szempontok szerint történt. Olyan programot kerestem, ami az én specifikus fejlesztési igényeimnek megfelelt.

A kezelhetőség egy fontos szempont az adatgyűjtő rendszer kiválasztásánál. A Python csomagként vagy futtatható kódként letölthető adatgyűjtők, mint például a Scrapy vagy a BeautifulSoup testre szabhatóbbak, specifikus feladatok ellátására fejlesztett változatokkal rendelkeznek, mivel nyílt forráskódúak. Ennek a testreszabhatóságnak az ára viszont hogy nem rendelkeznek egy könnyen kezelhető kezelőfelülettel. Ezzel szemben a Web Scraper egy Google Chrome-hoz percek alatt hozzáadható bővítmény, amely egy letisztult felülettel beüzemeltető.

A második szempont a fejlesztés akadálymentessége. A táblázatban szereplő összes eszköz rendelkezik a fejlesztők weboldalára feltöltött dokumentációkkal, melyek alapszinten leírják az eszközök működését, viszont ezek mellett sok segítséget nyújt, ha az eszköz rendelkezik egy aktív fejlesztői közösséggel, ahol segítséget lehet kérni StackOverflow-on vagy GitHub-on (mint például a fentebb tárgyalt Scrapy közösség).

Fontos megemlíteni ezek mellett az integrálhatóságot. Itt specifikusan a három alrendszer közötti integrálhatóságra gondolok. Az első alrendszer elképzeléseim szerint CSV vagy JSON, vagy ezekhez hasonló adatstruktúráként fogja továbbadni az adatokat a következő alrendszernek, ami egy adatbázis lesz. Ennek az alapszintű integrálhatóságnak mindegyik elemzett eszköz megfelel, hiszen képesek több formátumban exportálni.

Az utolsó szempont ami alapján értékelem az eszközöket, a használati díjuk, valamint a használati feltételeik.

Az alábbi táblázatban az eddig felsorolt és ismertetett adatgyűjtő eszközök összehasonlítása látható. Az összehasonlítás megkönnyítésének érdekében számértékeket rendeltem az adott cellákhoz. A cellákban 1-től 3-ig adtam értéket annak megfelelően, hogy az adott program az adott jellemzőből mennyire erős vagy gyenge. A szakdolgozat további táblázatos összehasonlításai szintén ezt a számozási rendszert használják.

	Kezelőfelület		Fejlesztés könnyűsége		Integrálhatóság a saját rendszerbe		Használati díj / használati feltétel	
Web Scraper	Egyszerű és intuitív	3	Nem igényel fejlesztést	3	Könnyű	3	Részben ingyenes	2
Selenium	Nincs	1	Dokumentáció alapján	1	Időigényes, nehezebb	1	Ingyenes	3
Scrapy	Nincs	1	Dokumentáció, és közösségi fórumok alapján	2	Időigényes, nehezebb	2	Ingyenes	3
Beautiful Soup	Nincs	1	Dokumentáció alapján	1	Időigényes, nehezebb	1	Részben ingyenes	2
Szociális platformok API-ai	Van, és adatvizualizációra is képes	3	Dokumentáció alapján	1	Időigényes, nehezebb	2	Ingyenes, viszont igényelni kell	2

2.1-es táblázat: Az adatgyűjtő rendszerek összehasonlítása

2.1.9. Adatfeldolgozó rendszerek

Az adatok egységesítésére és feldolgozására rengeteg eszköz áll rendelkezésünkre. Itt igazából nem kész programokról, hanem fejlesztői környezetekről beszélünk. Tehát az adatok feldolgozását egy saját fejlesztésű program fogja végezni, amely csatlakoztatva lesz az adatbázisomhoz.

A szakirányos tanulmányaim során eddig főleg relációs adatbázisokkal dolgoztam. Az ilyen adatbázisok lényege, hogy az adatokat táblákban tárolja. Ezek a táblák kapcsolatban, relációban vannak egymással. A sorok és oszlopok feltöltése után lekérdezésekkel lehet hozzáférni az adatokhoz, és módosítani azokat. A táblákat elsődleges és idegen kulcsok segítségével lehet összekapcsolni összetettebb lekérdezések esetén.

2.1.10. Korábban készített adatfeldolgozó rendszerem bemutatása

Az Adattárházak és Üzleti Intelligencia tárgy keretein belül egy három rétegű adattárházat kellett létrehoznom. A félév során több fejlesztői környezetet is használtam, de a legtöbbet a Visual Studio (SQL Server Data Tools bővítménnyel) és az SQL Server Management Studio közös használatával kialakított rendszerrel dolgoztam. Az ebben a rendszerben létrehozott adattárház működése a következő volt:

Először létrehoztam három réteget, azaz adatbázist. Ezek a Stage, Data Warehouse és Data Mart rétegek voltak. Ez a három réteg funkcionalitása tükrözte az adattárházakra jellemző ETL folyamatokat.

[9] Az ETL (Extract, Transform, Load) egy régóta használt adatintegrációs folyamat, amelyet arra használnak, hogy több forrásból származó adatokat egyetlen, összhangban lévő adatkészletbe kombináljanak, majd betöltsék egy adattárházba, data lake-be vagy más célrendszerbe.

Az első (Stage) adatbázisba betöltöttem a forrásfájlokból (CSV kiterjesztésű táblázatok) az adatokat. Ezek az adatok ebben a rétegben még mind azonos adattípussal rendelkeztek. Minden tábla minden oszlopát szöveggént importáltam az adatbázisba, annak érdekében, hogy minél gyorsabban, minél több adatot tudjak beolvasni adattípus ütközések nélkül. Ezek az táblák még nem rendelkeztek elsődleges, és idegen kulcsokkal sem.

Ezután az adatok átkerültek a Data Warehouse, vagy más néven Transform rétegbe. Itt minden oszlophoz beállításra került a megfelelő adattípus, valamint a historizálást elősegítő timestamp típusú oszlopok is létrejöttek. Néhány oszlop törlésre és hozzáadásra került, valamint az oszlopok nagyrészen valamiféle egységes transzformációt végeztem.

Miután minden adattípus beállításra került, kiválogattam, hogy melyik táblák mely oszlopaikra lesz szükségem a harmadik rétegben. Ezt követően az adatokat áttöltöttem a harmadik

adatbázisba, ahol a végleges, transzformált táblák szerepeltek, kapcsolatokkal és kulcsokkal ellátva.

Azért tartom fontosnak ennek a folyamatnak a részletes bemutatását, mivel a szakdolgozatom keretein belül fejlesztett rendszer is hasonló lépésekkel fog rendelkezni. A betöltött táblák ebben az esetben a Web Scraper által begyűjtött adatok lesznek, amelyek hasonló transzformációs folyamatokon fognak átesni, mint amelyeket az előbb mutattam be.

2.1.11. Adatvizualizációs rendszerek

A mai adatalapú világban kiemelt fontossággal bír az összetett adatkészletekből kinyert információk effektív értelmezése. Az adatvizualizációs eszközök kiemelkedő szerepet játszanak ebben a folyamatban, hiszen átalakítják a nyers adatokat látványos grafikonokká, diagramokká és irányítópultokká. Ezek az eszközök elősegítik az összefüggések és trendek észrevételét, valamint segítenek az informált döntéshozatalban. Sok eszköz áll rendelkezésünkre az adatvizualizáció terén.

2.1.12. Jupyter

A Jupyter egy webes adatvizualizációs eszköz, amely lehetővé teszi a felhasználók számára, hogy létrehozzanak és megosszanak dokumentumokat, amelyek vizualizációkat, egyenleteket, szöveget és élő kódot tartalmaznak. A Jupyter ideális az adattisztításhoz és átalakításhoz, statisztikai modellezéshez, numerikus szimulációhoz, interaktív számításhoz és gépi tanuláshoz.

- Interaktív fejlesztési környezettel rendelkezik, ahol egyszerre lehet lekérdezéseket futtatni, és ezek eredményét valós időben vizualizálni.
- Több programozási nyelvvel működik. Alapvetően Pythonos működésre fejlesztették, viszont támogatja az R és Julia nyelveket is, amelyeket egy interfészen lehet használni akár egyszerre is.
- A Jupyter Notebook letisztult, jegyzetfüzet szerű kialakításának köszönhetően könnyen átlátható.

2.1.13. Python pandas és NumPy könyvtárakkal

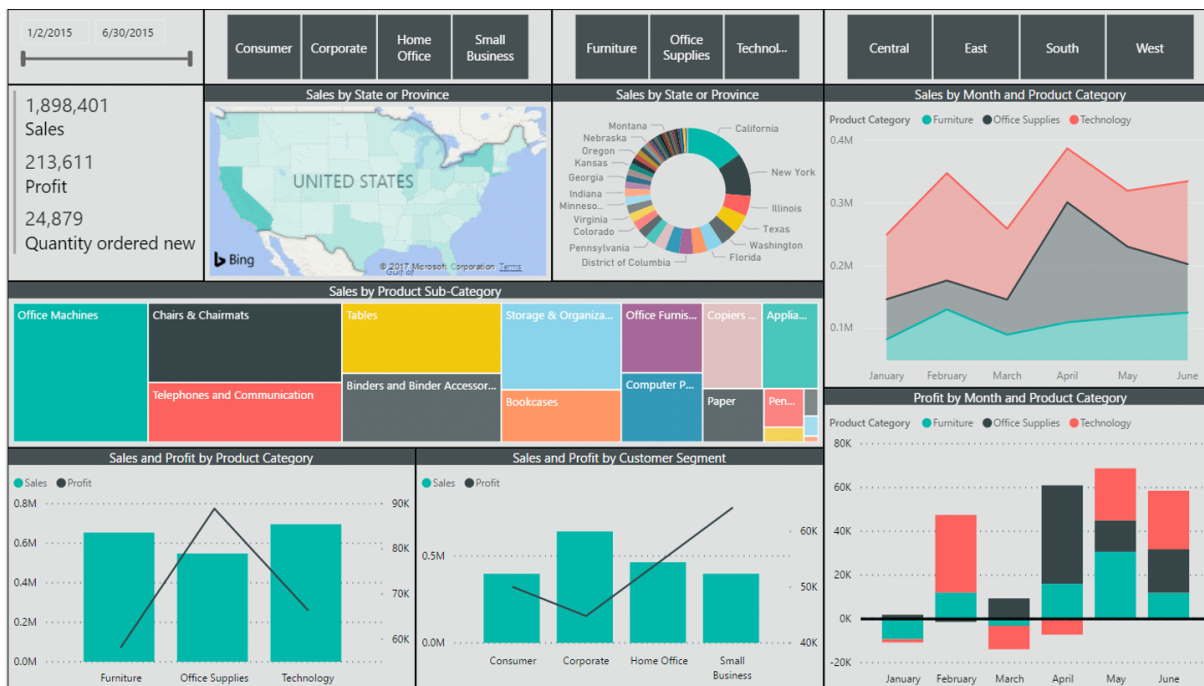
A Python széles körben használt az adatelemzés terén, és olyan erős adatelemző könyvtárakkal rendelkezik, mint a pandas és a NumPy. Ezek a könyvtárak kiterjedt funkcionalitást kínálnak nagy adathalmazok kezelésére, számítások végrehajtására és összetett lekérdezések készítésére.

- A pandas és NumPy könyvtárak strukturált adatok feldolgozására lettek kifejlesztve. Optimalizált statisztikai függvényeikkel egyszerűen és gyorsan képesek szűrni, rendezni, és csoportosítani nagyobb mennyiségű adatot is.

2.1.14. Microsoft Power BI

Power BI egy üzleti analitikai eszköz, amelyet a Microsoft fejlesztett ki. Alkalmas részletgazdag, látványos jelentések létrehozására, valamint adatmodellezésre és transzformálásra.

- A betöltött adatok táblákba kerülnek, majd a táblák között kapcsolatokat lehet létrehozni. A Power Query segítségével adattisztításra és transzformálásra is van lehetőség.
- Jól integrálható más Microsoft programokkal, mint például a Microsoft SQL Server Management Studio-val
- Képes interaktív, valós idejű riportok és vizualizációk készítésére
- Széleskörű adatbázis támogatással rendelkezik. Működik Teradatával, Salesforce-al, PostgreSQL-el, Oracle-el és még sok más adatbázissal.
- Sokféle forrásból képes adatokat importálni. Képes fogadni Excel munkalapokat, SQL Server-eket, CSV és JSON kiterjesztésű szöveges dokumentumokat és webcímeket is.



2.4-es ábra: Power BI dashboard minta

2.1.15. Az adatvizualizációs rendszerek összehasonlítása

Az alábbi táblázatban (2.2.) hasonlítom össze a három általam választott legfontosabb szempont alapján az előbb ismertetett adatvizualizációs eszközöket:

	Egyszerűség		Integrálhatóság		Használati díj	
Jupyter	Alapszintű Python tudást igényel	2	Szűkebb körű integrálhatóság a PowerBI-hoz képest	2	Ingyenes	3
Python	Python tudást igényel	1	Python alapú programokkal könnyen integrálható	2	Ingyenes	3
PowerBI	Intuitív kezelőfelület	3	Könnyen integrálható, főleg Microsoft programokkal	3	Ingyenes	3

2.2-es táblázat: Az adatvizualizáló rendszerek összehasonlítása

2.2. A saját rendszer programlistájának összegzése

2.2.1. Az adatgyűjtő program kiválasztása

A szempontok mérlegelése után a Web Scraper mellett döntöttem, mivel egyszerű és könnyen kezelhető a grafikus felhasználói kezelőfelülete, ingyenes, és könnyen integrálható az rendszerem adatfeldolgozó részébe, mivel csak a Web Scraperből exportálható CSV fájlokat kell betöltenem az adatbázisba. Egy ekkora mennyiségű adathalmaz begyűjtése nem könnyű, viszont egyszerű. Ha egyszer sikerül felkonfigurálni a scrapert, akkor az utána több tízezer sornyi adatot képes begyűjteni. Úgy érzem a Web Scraper tökéletesen el tudja látni ezt a feladatot, és indokolatlan egy ennél összetettebb rendszer használata.

Megjegyzés: Web Scraper mellett szükség szerint az Octoparse nevű program ingyenes verzióját is igénybe fogom venni. Az Octoparse lényegében ugyanúgy működik mint a Web Scraper, azaz egy kódolás nélküli böngészőbe integrálható web scraper, viszont sokkal több oktatói anyaggal rendelkezik, ezért lehetséges hogy effektívebben tudom használni, mint a Web Scraper-t.

2.2.2. Az adatfeldolgozó program kiválasztása

Az adatok tárolására a Microsoft SQL Server Management Studio-t fogom használni. Az adatok feldolgozásához pedig a Visual Studio-t (a SQL Server Data Tools csomaggal felszerelve). Egy ilyen környezetben fejlesztett rendszer képes lesz teljesíteni az adatfeldolgozási feladatokat, emellett nagyon egyszerű lesz a Power BI-al integrálni.

2.2.3. Az adatvizualizációhoz használt program kiválasztása

A feldolgozott adatok vizualizálására a Microsoft Power BI-t fogom használni, mivel ingyenes, képes nagymennyiségű adatot modellezni, transzformálni és vizualizálni, valamint már van tapasztalatom a programmal. Emellett jól integrálható a Microsoft SQL Server Management Studio-val.

2.3. Kapcsolódó témák

2.3.1. A rendszer által vizsgált tartalmak specifikálása és a rövid formátumú videók ismertetése

A rendszerem különböző szociális platformokra feltöltött, videókat, posztokat és profilokat fog elemezni. A videókon belül specifikusan a rövid formátumú tartalmakra (angol kifejezés: short-form content) szeretnék fókuszálni, mivel manapság egyre népszerűbbek az egy-két perces, vagy akár fél percnél is rövidebb formátumú videók. Az elemzett platformok a TikTok, Instagram, YouTube és a Facebook.

A 2016-ban kiadott, és 2018-ban világszerte népszerűvé vált TiktTok sikerének kulcsa elsősorban a rövid formátumú videók bevezetése volt. Ezeket a videókat egy személyre szabott, algoritmus által válogatott videó lejátszó felületen lehet megtekinteni. Ahogy megnyitjuk az alkalmazást, egyből elindul egy videó, amelyről a következőre görgetéssel lehet eljutni. Ez a formátum olyan óriási sikert aratott, hogy az ezt követő években az összes piacvezető platform kénytelen volt beépíteni egy hasonló funkciót a felhasználói felületükbe. A YouTube bevezette a YouTube Shorts-ot, az Instagram az Instagram Reels-t és a Facebook a Facebook Reels-t. Ezek a felületek kisebb eltérésekkel ugyanúgy működnek, mint a TikTok. Ebből következik, hogy nagyjából ugyanazokat az adatokat teszik elérhetővé adatgyűjtés szempontjából.

2.3.2. Begyűjthető adatok elemzése az egyes platformok profiljain, és ezek összehasonlítása

Minden szociális platformon más nyilvános adatok érhetőek el. Ebben a rövid elemzésben azt fogom bemutatni, hogy melyik platformon pontosan milyen adatok érhetőek el a nyilvános felhasználók profiljain és posztjain, azaz miket tudok web scraping-gel begyűjteni.

TikTok profilon elérhető adatok: név, követők száma, követések száma, like-ok száma, biográfia, felhasználó által megadott link

Instagram profilon elérhető adatok: név, követők száma, követések száma, posztok száma, biográfia, felhasználó által megadott link

YouTube profilon elérhető adatok: név, feliratkozók száma, videók száma, biográfia, felhasználó által megadott linkek

Facebook profilon elérhető adatok: név, követők száma, követések száma, biográfia, felhasználó által megadott linkek, névjegyben megadott adatok

Összegzés: Mind a négy platformon nagyjából hasonló adatok érhetőek el a nyilvános profilokon. A legtöbb adatot Facebook-on a névjegy almenüben lehet megadni, viszont ezek megadása opcionális, azaz nem garantált hogy minden felhasználó teljesen kitölti.

2.3.3. Begyűjthető adatok elemzése az egyes platformok videóin, és ezek összehasonlítása

Fontos leszögezнем, hogy kizárólag a „short-form” tartalmakat elemzem ebben a szekcióban, azaz Instagram-on és Facebook-on a Reels videók adatmezőit, YouTube-on pedig a Shorts videók adatmezőit. (YouTube-on az egy percnél rövidebb videók automatikusan Shorts videóként kerülnek feltöltésre, valamint a YouTube Shorts felületén csak az ilyen videók jelennek meg.)

TikTok videók alatt elérhető adatok: feltöltő neve, feltöltés ideje, megtekintések száma, like-ok száma, kommentek száma, lementések száma, videó leírása, hashtag-ek, komment szekció tartalma

Instagram Reels videók alatt elérhető adatok: feltöltő neve, feltöltés ideje, megtekintések száma, like-ok száma, kommentek száma, lementések száma, videó leírása, hashtag-ek, komment szekció tartalma

YouTube Shorts videók alatt elérhető adatok: feltöltő neve, megtekintések száma, like-ok száma, kommentek száma, videó leírása, hashtag-ek, komment szekció tartalma

Facebook Reels videók alatt elérhető adatok: feltöltő neve, megtekintések száma, like-ok száma, kommentek száma, megosztások száma, videó leírása, hashtag-ek, komment szekció tartalma

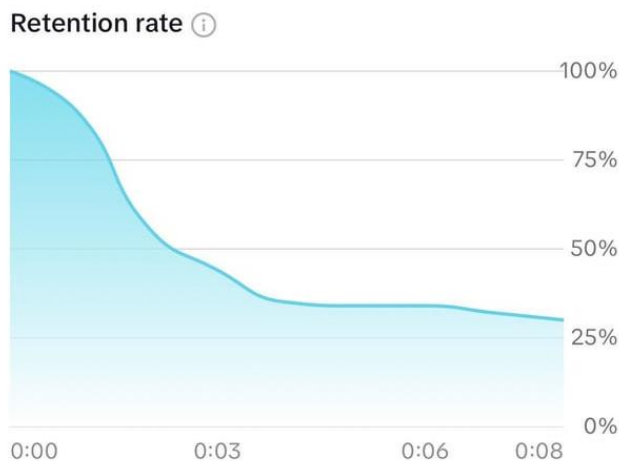
Összegzés: Minden platform majdnem megegyező adatmezőkkel rendelkezik, viszont Facebook-on nem teszik elérhetővé a feltöltés idejét, de elérhető a megosztások száma, ami fontos adat lehet ha a videó sikerességét szeretnénk mérni.

2.3.4. Mitől lesz sikeres egy rövid formátumú videó, miben lehet mérni a hatásosságot?

Természetesen a megtekintések, like-ok és kommentek száma egyértelmű indikátora annak, hogy mennyire sikeres egy videó, viszont más, nem olyan egyértelmű mérőszámok is fontosak.

Ezek közül talán a legfontosabb az „average viewer retention rate”, azaz az átlagos nézettségi idő. Ha egy videónak két másodperc az átlagos nézettségi ideje, az azt jelenti, hogy a nézők nagyrésze két másodperc után tovább tekert. A nézettségi arányt egy olyan grafikonon is lehet ábrázolni, amelynek vízszintes tengelye a videó kezdete óta eltelt idő másodpercekben, a függőleges tengelye pedig a videót addig az időpontig megtekintő felhasználók százalékos aránya az összes videót megtekintő felhasználóhoz képest.

Az alábbi képen látható a TikTok-ba beépített TikTok Analytics Tool viewer retention rate grafikonja:



2.5-ös ábra: A TikTok beépített analitikai felületének „Viewer retention rate” grafikonja

[10] Az évek során több kutatás kimutatta hogy a modern ember koncentrációs képessége bizonyos mértékben csökkent, ami nagyrészt a környezetünk rohamos változásával, illetve egyéb más okokkal magyarázható.

Az emberek figyelmének megragadása és megtartása a legfontosabb eleme egy videó hatásosságának. Ezt legegyszerűbben figyelemfelkeltő hang és videó effektekkel, élénk színekkel, és gyors, célratörő beszédstílussal lehet elérni.

2.3.5. Kiknek és miért fontos a tudatos tartalomstratégia?

Minden influenszernek, tartalomgyártónak, vállalkozónak, vagy interneten kampányoló szervezetnek szüksége van egy tartalomstratégiára, ha effektíven és konzisztensen szeretnének elérni egy minél szélesebb közönséget.

A tudatos tartalomstratégia akkor valósul meg, amikor egy tartalomgyártó a közönségének ismeretében szándékosan olyan típusú, stílusú tartalmakat tölt fel, amellyel a célközönségének imponál. Ezeket a feltöltéseket megfelelő platformokon és megfelelő időben végzi. Ha az analitikai adatok alapján tudja, hogy a célközönsége melyik napszakban, melyik órákban aktív az egyes platformokon, akkor lehetősége van megtalálni azt az időszakot, amikor a legtöbb emberhez el tudja juttatni az üzenetét.

Ennek a stratégiának a kidolgozásában játszanak kulcsfontosságú szerepet az analitikai eszközök, mivel valós idejű visszajelzést nyújtanak a feltöltött tartalmak hatásosságáról.

2.3.6. Mit jelent az adatvizualizáció, és miért jelentős?

[11] Az adatvizualizáció különböző információk és adatok grafikus ábrázolása. Egyszerű és célrátörő módszert biztosít a trendek, összefüggések és kiugró értékek észrevételére és megértésére az adatokban. Emellett kiváló eszköz a laikusok felé történő kommunikációra, mivel érthetően, félreértések nélkül képes átadni az információkat.

Az adatszempléltetés kulcsfontosságú szerepet játszik a modern világban, és manapság mindenhol megtalálható. A táblázatok, gráfok, grafikonok, diagramok, hőtérképek, amelyekkel életünk különböző területein találkozunk, mind adatvizualizációs eszközöknek számítanak. Ezeknek a célja mindig egy bizonyos adathalmaz szemléltetése, információközlés.

[11] Az adatvizualizáció fontossága egyszerű: segít az embereknek látni, és jobban megérteni az adatokat. Az egyszerű és összetettebb adatelemzésekre egyaránt igaz, hogy a megfelelő vizualizáció mindenkivel megértetheti az adott helyzetet, szaktudásbeli szintjétől függetlenül.

Ezek az eszközök lehetővé teszik számunkra, hogy a mindennapi életünk során generált és fogyasztott nagy mennyiségű adatot megértsünk. Az üzleti szektorban az adatelemzés segíti a vállalatokat abban, hogy stratégiai döntéseket hozzanak, piaci trendeket azonosítsanak és üzleti teljesítményüket javítsák. Az állami és kutató szférában az adatvizualizáció segíthet az összetett problémák megértésében és a társadalom számára releváns információk átadásában. Az adatelemzést és adatvizualizációt széles körben használják a vállalati vezetőktől a kutatókig, az üzleti elemzőktől az oktatókig, valamint a kormányzati szervektől a non-profit szervezetekig.

Minden olyan szervezet vagy személy, aki fontos döntéshozatalban vesz részt, használhatja ezeket az eszközöket.

2.3.7. Az adatvizualizáció legelterjedtebb típusai

Az adatok szemléltetésére rengeteg eszköz áll rendelkezésünkre. Ezek közül a legelterjedtebbek a következők:

Táblázat: A táblázat egy strukturált formátum, amelyben az adatok cellákban szerepelnek, oszlopokba és sorokba rendezve. Ez talán a legalapvetőbb adatvizualizációs forma, amely lehetővé teszi az adatok könnyű összehasonlítását és rendezését.

Gráf: A gráfok csúcsokból, és ezeket összekötő élekből állnak, melyek különböző entitások egymás közti kapcsolatát szemléltetik. A kapcsolatok lehetnek egy vagy több irányúak.

Térkép: A térképek földrajzi adatokat jelenítenek meg, például országokat, régiókat vagy városokat, és azokhoz kapcsolódó adatokat ábrázolnak. Lehetővé teszik a földrajzi elhelyezkedés alapján történő vizsgálatot és összehasonlítást.

Hő térkép: A hő térképek színek kódolt diagramok, amelyek az adatok értékeit ábrázolják egy adott térben. A magasabb értékekhez forróbb, élénkebb színek tartoznak, míg az alacsonyabb értékekhez hidegebb, halványabb színek.

Vonal diagram: A vonal diagramok vonalakkal ábrázolják az adatok változását időben vagy más változó függvényében. Ez a típusú diagram különösen alkalmas trendek és időbeli változások vizualizálására.

Osztópdiagram: Az osztópdiagramok osztópokat használnak az adatok ábrázolására, ahol az osztópok magassága az adatok értékeit jelképezi. Ezek a diagramok alkalmasak az adatok összehasonlítására és az értékek közötti különbségek megjelenítésére.

Kördiagram (Pie Chart): A kördiagramok körképeket használnak az adatok arányainak ábrázolására. A kördiagramok általában egyszerű arányok, részek és egészek vizualizálására használatosak.

Gantt-diagram: A Gantt-diagramok idővel kapcsolatos feladatokat és eseményeket ábrázolnak egy vízszintes vonalon. Az ilyen diagramok segítenek a projektmenedzsmentben és a feladatok időbeli ütemezésében.

Venn-diagram: A Venn-diagram egy olyan diagram, amely kör alakú régiókat használ az adatok közötti kapcsolatok megjelenítésére. A körök átfedése vagy azonos területen való elhelyezkedésük azt jelzi, hogy az adatok között közös tulajdonságok vagy kapcsolatok vannak.



2.6-os ábra: Példa a különböző adatvizualizációs módszerekre

2.3.8. Az adatvizualizáció lehetséges hátrányai

Első látásra úgy tűnhet, hogy az adatvizualizációnak csupán előnyei vannak, viszont az adatokat a megfelelő háttértudás nélkül, vagy a szemléltetésre alkalmatlan szemléltetési stílus használata miatt könnyen félre lehet érteni. Emellett mindig érdemes ellenőrizni hogy milyen forrásból származnak az adatok, és milyen célból készülhetett a vizualizáció.

Az adatvizualizáció esetleges hátrányai az alábbiak:

- A vizualizáció elfogult vagy pontatlan információkat tartalmazhat
- Egy korreláció nem mindig jelent ok-okozati összefüggést
- Az alapvető üzenetek elveszhetnek a leképzés során

2.3.9. Az adatvizualizáció alapelvei

[12] [13] **Megjegyzés:** Az alábbi alapelveket a bekezdés elején megjelölt források alapján állítottam össze. Egyéb források is hasonló alapelveket, szabályokat tartalmaztak.

1. Az üzenet előtérbe helyezése:

Bár ez magától értetődőnek tűnhet, először az információra és az üzenetre kell összpontosítani, mielőtt a konkrét megvalósításra gondolnánk. Más szóval, nem feltétlenül a végleges vizualizációra, és annak kinézetére kell fókuszálni, hanem arra, hogy milyen alapinformációkat akarunk átadni a vizualizációval. Minden más csak ezután következhet.

2. A megfelelő szoftver használata:

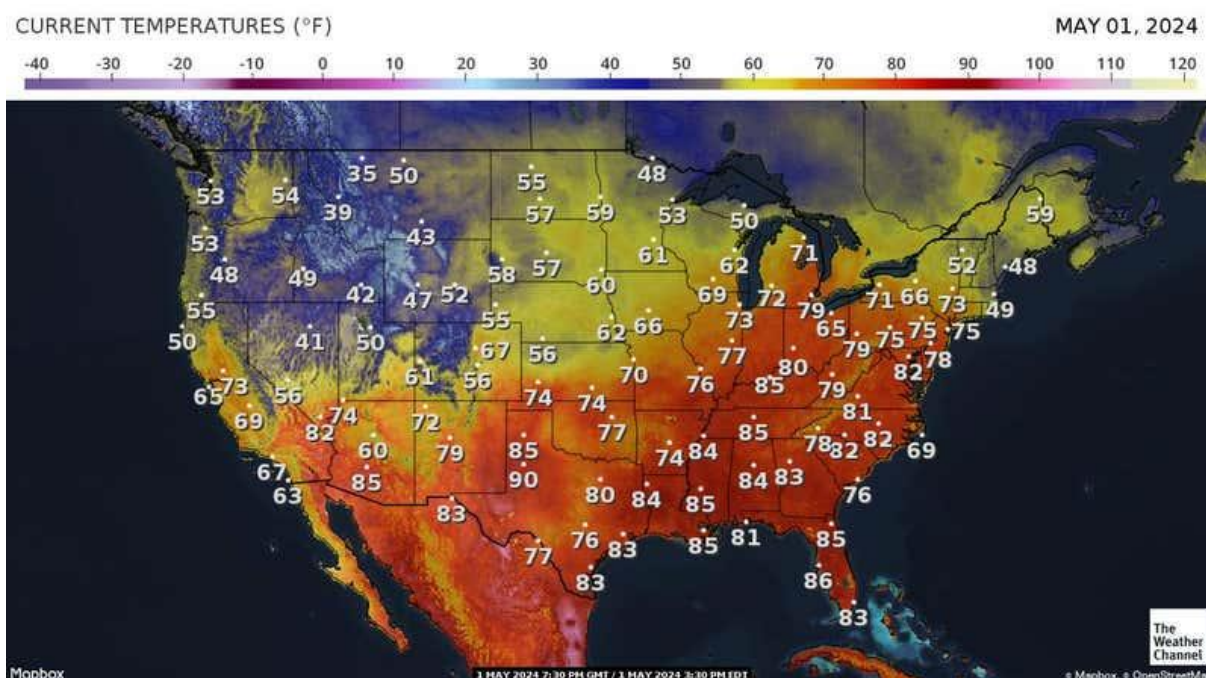
Az effektív vizualizációk elkészítéséhez elengedhetetlen egy erre fejlesztett program alapvető ismerete és használata. Egyszerűbb, táblázatos vizualizációkat bármilyen táblázatkezelő szoftverben is el lehet végezni, viszont annak érdekében hogy komplexebb, átfogóbb, és hatásosabb vizualizációkat legyünk képesek létrehozni, más, komplexebb eszközök elsajátítására lesz szükség. Sok ilyen rendszer létezik, amelyek közül több ingyenes, és rengeteg internetes oktatói tartalom áll rendelkezésünkre.

3. A megfelelő geometriai alakzatok kiválasztása:

Egy adat vagy adathalmaz szemléltetésekor sok esetben több geometriai alakzat, azaz vizualizációs stílus jöhet szóba. Ilyenkor megfontoltan kell kiválasztani az adathalmaz természetéhez és a szemléltetés céljához legmegfelelőbb stílust.

Egyes adatok értékeit, egymáshoz hasonlított arányát, eloszlását, kapcsolatait mind különböző alakzatokkal lehet a legcélrátörőbben szemléltetni. Például egy választás eredményét egy kördiagrammal lehet a legegyszerűbben szemléltetni, mivel egyértelműen tükrözi az egyes pártokra leadott szavazatok arányát.

Egy időjárás előrejelzést lehet egy olyan táblázattal szemléltetni, amelynek minden sora egy adott település hőmérsékletét, páratartalmát és csapadék valószínűségét tartalmazza, viszont nem biztos hogy ez a legcélravezetőbb megoldás. Ezeket az adatokat sokkal látványosabban lehet ábrázolni például egy színkódolt hőtérképpel, vagy páratartalom térképpel.



2.7-es ábra: A The Weather Channel weboldal hőmérsékleti előrejelzése

4. A színek fontossága:

Az adatvizualizáció terén a színek hatalmas jelentőséggel bírnak. [12] Széleskörű kutatások szerint a vizualizációk emlékezetességének legfontosabb eleme a színpaletta. A színek magukban információt hordoznak, mivel érzelmeket és emlékeket asszociálunk a különböző árnyalatokhoz. A jól olvashatóság és értelmezhetőség érdekében érdemes sok, élénk színt használni, viszont annak ellenére, hogy csaknem végtelen színskála áll rendelkezésünkre, arra is oda kell figyelnünk, hogy minden színt tudatosan alkalmazzunk, ne zsúfoljuk túl a színpalettát.

5. Részletes magyarázat és egyértelműsítés

Annak ellenére, hogy a vizuális elemek (színek, alakzatok) könnyebben kivehetőek, értelmezhetőbbek, mint a szövegrészek és feliratok, fontos hogy minden szín és alakzat részletesen és egyértelműen meg legyen magyarázva. Az egyértelmű információközlés érdekében a szemléltetés minden részét érthetővé kell tenni. Például egy emberi kapcsolatokat ábrázoló gráf esetében fel kell tüntetni, hogy a csúcsok embereket, az élek pedig a köztük fennálló kapcsolatot jelentik.

2.3.10. A web scraping jogszerűsége

[14] [15] Az adatgyűjtés jogszerűségének kérdésében több forrásnak is utána néztem, és a következőt találtam:

A nyilvánosan elérhető webes információk automatizált módon történő lekérése (scraping) jogszerű, amennyiben az adatok felhasználása:

- Nem okoz közvetlen kárt az adott weboldal üzleti tevékenységeiben vagy működésében
- Nem tartalmaz személyes azonosításra alkalmas információkat

Tehát amíg nem személyes adatokat gyűjtünk be, és nem adjuk tovább üzleti célokra a gyűjtött információkat, a folyamat teljes mértékben legális.

3. KÖVETELMÉNY SPECIFIKÁCIÓ

3.1. Adatgyűjtési folyamat

Az adatok begyűjtése Web Scraperrel és Octoparse-al fog történni. Ennek az adatgyűjtési folyamatnak képesnek kell lennie egyszerre több tartalomgyártó adatait begyűjteni, automatizáltan. Kell hogy gyűjtsön adatot a tartalomgyártókról, és az általuk feltöltött összes rövid formátumú videóról. A folyamat végén a program a begyűjtött adatokat egy CSV vagy Excel formátumú fájlként exportálja. Ezek az adatok utána betöltésre kerülnek az adatbázisba.

3.2. Adattisztítási folyamat

Az adatok átalakítását egy általam fejlesztett program fogja végzi, amely képes az adatbázis összes táblájának összes oszlopát megfelelő módon transzformálni. Ezeknek a transzformációknak a lényege, hogy előkészítsék az adatokat az elemzésre. Az oszlopok adattípusának beállítására is itt kerül sor. A program átlátható és továbbfejleszthető kell hogy legyen, valamint verziókövetéssel is kell hogy rendelkezzen.

3.3. Tesztelés

A transzformáló programnak képesnek kell lennie a hiányos, vagy üres mezők kezelésére. Emellett észlelnie kell, ha egy mezőn már megtörtént az átalakítás, hogy többszörös futtatások esetén ne történjen meg többszörösen a transzformáció. A különböző eshetőségek lefedettségét tesztek segítségével ellenőrzöm.

3.4. Elemzés

Az adatok megvizsgálását követően kijelölöm az elemzési témákat, és utánajárok a felvetéseimnek (Pl. számít-e egy videó címének a hossza, vagy a hashtagek száma). Miután végeztem az elemzésekkel, összeírom a felismert összefüggéseket, majd ezekhez hozzárendelem a legmegfelelőbb vizualizációs eszközöket (Pl. oszlopdiagram, kördiagram, trendvonal).

3.5. Vizualizáció

A Power BI segítségével az elemzéseim eredményét vizualizációkkal illusztrálom.

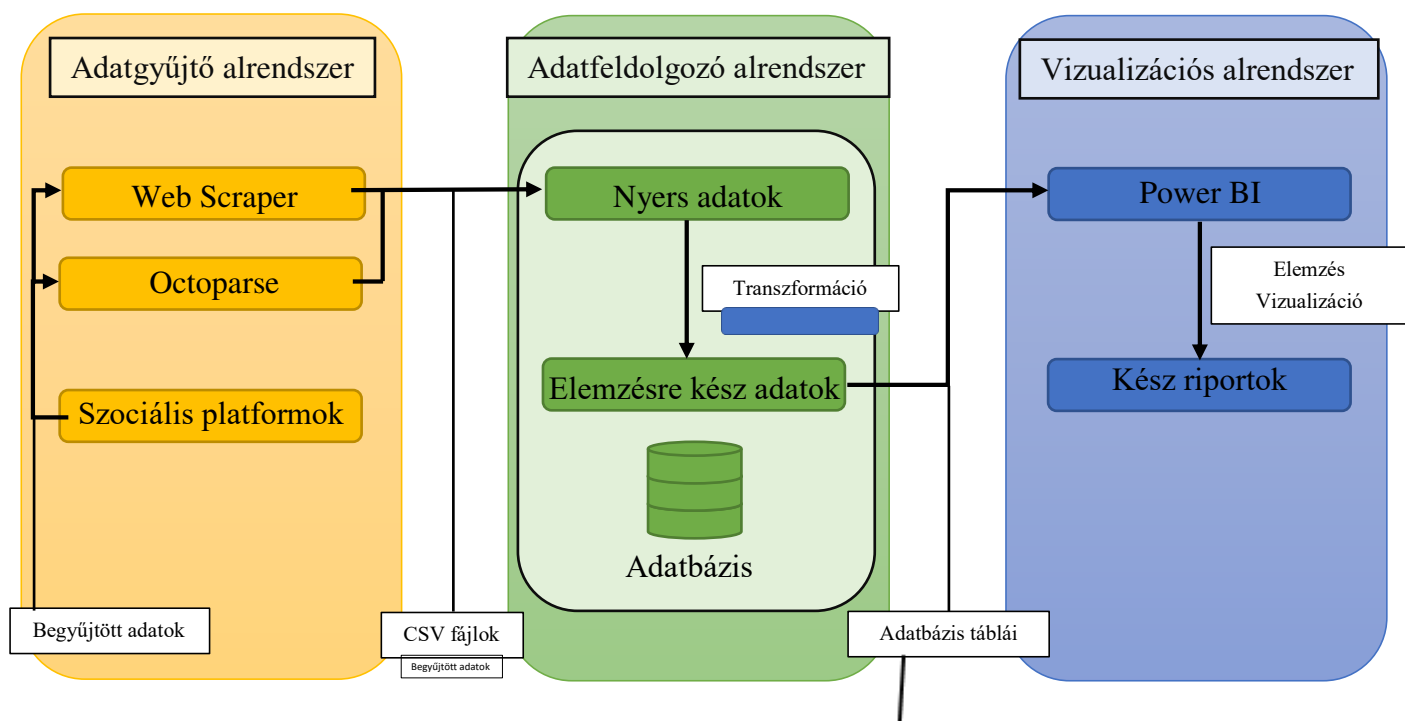
4. TERVEZÉS

A tervezés fejezetben ismertetem a rendszerem fejlesztési tervét. Először a „rendszerterv” részben mindhárom alrendszer bemutatásra kerül, majd ismertetem az ott történő folyamatokat, valamint az adatbázisom táblarajzát. Ezután, a „funkciólista” részben a rendszer főbb funkcióit egy folyamatábrával és egy rövid magyarázattal mutatom be. Végül, a „fejlesztés tervezett menete” részben felsorolom a fejlesztés lépéseit, és egy Gantt diagrammal ábrázolom az egyes részek becsült időráfordítását.

4.1. Rendszerterv

4.1.1. A rendszerterv ábra

A lenti képen látható a teljes rendszer rendszerterv ábrája. Az ábrát követő alfejezetekben az itt látható komponenseket fogom részletesen ismertetni. A képen látható nyilak az adat útját szemléltetik az alrendszereken keresztül, a szociális platformoktól az elemzések vizualizációjáig.



4.1-es ábra: Rendszertev ábra

4.1.2. Az adatgyűjtő alrendszer ismertetése

Az adatok begyűjtése Web Scraperrel (vagy szükség szerint Ocroparse-szal) fog történni. A scraper felkonfigurálása sitemap-ek létrehozásával kezdődik. A sitemap-ek, azaz „oldal térképek” célja, hogy előírják a scraper számára, hogy a kiindulási oldalon mely linkeket kell megnyitnia, és azokon belül mely mezőket kell begyűjtenie, és ezeket milyen néven kell letárolnia. Az én esetemben ez egy külön sitemap létrehozását jelenti minden platformon, (TikTok, Instagram, YouTube, Facebook). A kiinduló link mindig egy tartalomgyártó profilja lesz, mivel innen érhetőek el a fentebb tárgyalt begyűjtendő adatok (2.3.2. és 2.3.3. alfejezetek), és az adott felhasználó bejegyzései. A konfigurálás és a kiinduló linkek megadása után a scraper lefut, és összegyűjti a meghatározott adatokat.

A folyamat végén a begyűjtött adatok egy CSV, vagy egyéb táblázatos formátumú fájlként lesznek elérhetőek. A következő alrendszerbe való integráció egyszerűen történik, mivel az adatbázis rengeteg típusú adatforrást képes fogadni, köztük CSV fájlokat is.

4.1.3. Az adatfeldolgozó alrendszer ismertetése

Ez az alrendszer lényegében az adatbázist, és saját fejlesztésű adattranszformáló programot jelenti.

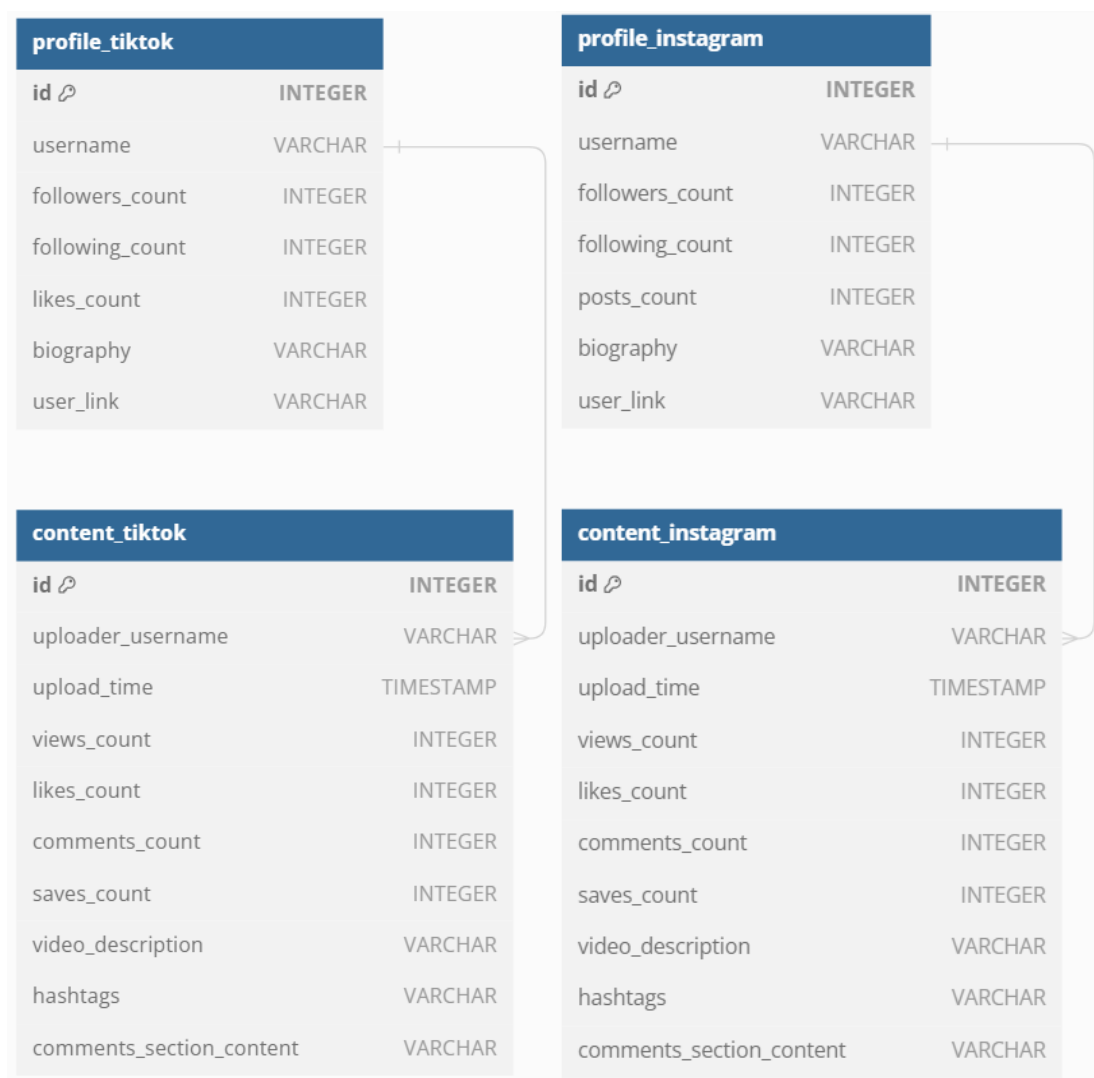
Az első lépés a nyers adatok importálása. Ilyenkor még nem történik semmilyen adattisztítás, vagy transzformáció, csupán letárolásra kerülnek az adatok.

Ezután következik a transzformációk elvégzése. Itt C# metódusokkal és SQL lekérdezésekkel történik az egyes oszlopok hozzáadása és törlése, átnevezése, a hiányos mezők kezelése, és a megfelelő adattípusok oszlopokhoz való hozzárendelése. Emellett strukturált, kapcsolatokkal rendelkező táblákat hozok létre.

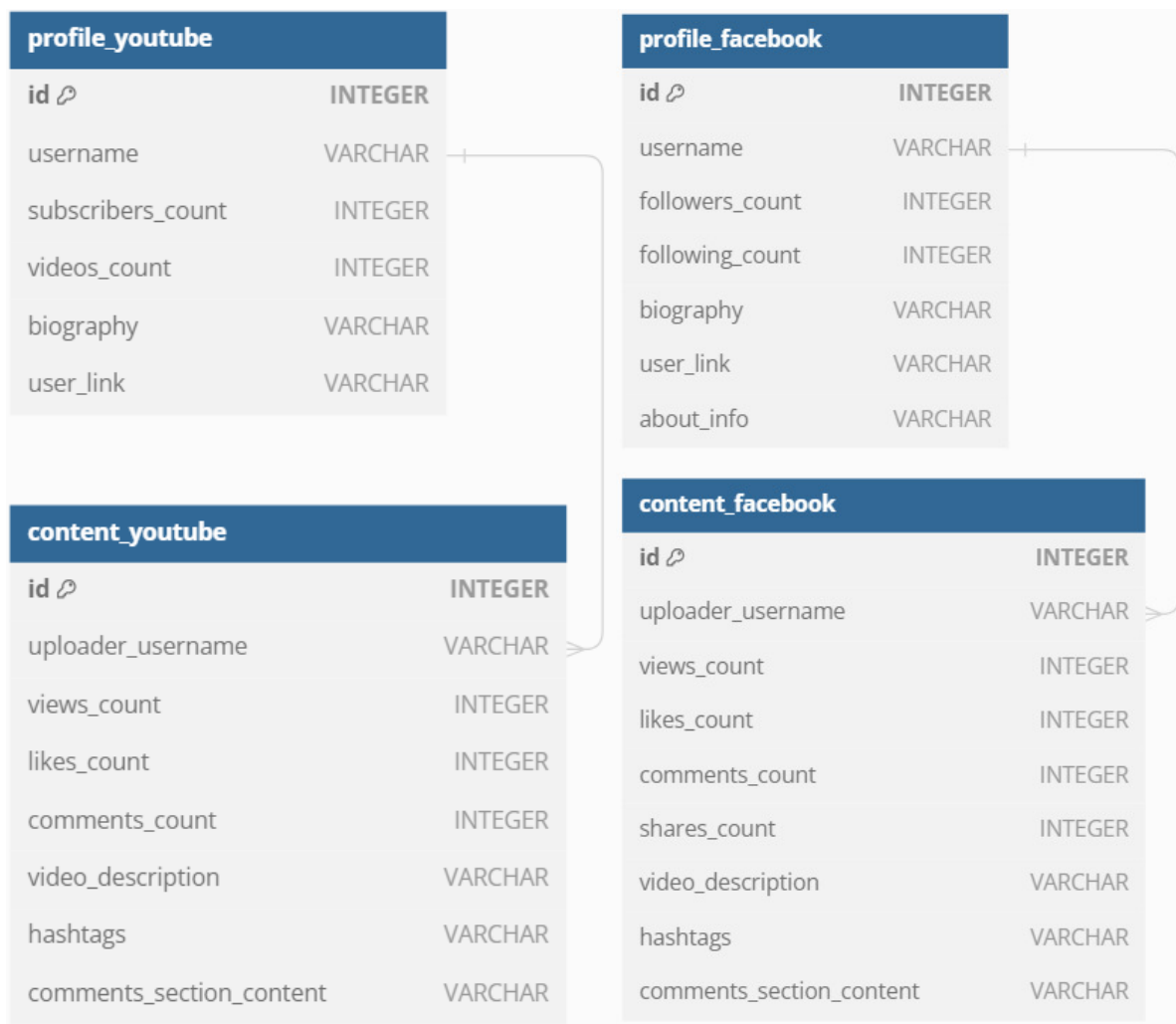
Ilyenkor már rendezett, lekérdezésre kész állapotban szerepelnek az adatok. Ezekből a táblákból készítem a riportokat, elemzéseket. Ezeknek az elemzéseknek az eredménye lesz az, amit a teljes rendszer harmadik nagy része, a vizualizációs alrendszer lát. A két alrendszer közötti kapcsolat létrehozása egyszerű, mivel az SQL Server Management Studio és a Power BI mindkettő Microsoft által fejlesztett program. Az adatbázis kapcsolat felállítása néhány perc alatt elvégezhető.

4.1.4. Az elemzési alrendszer adatbázisának felépítése

Alapvetően két típusú tábla lesz az adatbázisban, egy „profile_” előtaggal rendelkező profiladatokat tartalmazó tábla, és egy „content_” előtaggal rendelkező, feltöltött tartalmak adatait tartalmazó tábla. Mivel négy platformot vizsgálunk, ez összesen 8 táblát eredményez. A táblák végleges állapota így fog kinézni:



4.2-es ábra: A TikTok-ról és Instagram-ról gyűjtött adatok táblarajza



4.3-as ábra: A YouTube-ról és Facebook-ról gyűjtött adatok táblarajza

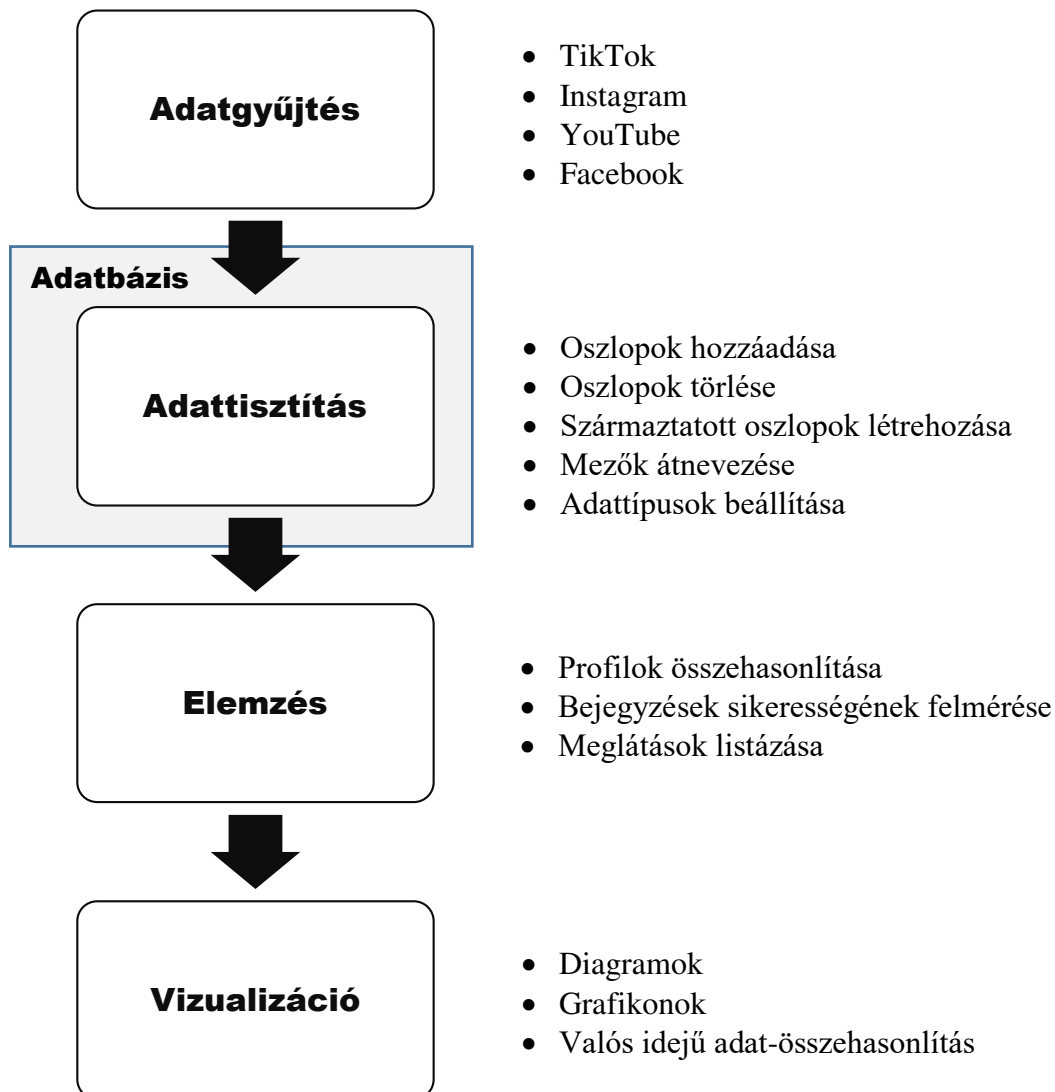
4.1.5. A vizualizációs alrendszer ismertetése

A vizualizációs alrendszer lényegében egy Power BI irányítópult (dashboard) lesz, amelynek hozzáférése lesz az adatbázis tábláihoz. A transzformáló program által létrehozott oszlopok mellett nagy valószínűséggel Power BI-ban is létre fogok hozni származtatott oszlopokat a vizualizációk elősegítésének érdekében.

Az elemzések eredményeit szemléltető vizualizációkat a 2.3.9-es pontban ismertetett alapelvek alapján fogom elkészíteni.

4.2. Funkciólista

- **Adatgyűjtés:** Web Scraper és Octoparse segítségével begyűjtésre kerülnek az adatok a négy szociális platformról.
- **Adattisztítás:** Az adatbázisban az adatok táblákba rendeződnek, majd a transzformáló program lefuttatását követően elemzésre alkalmas állapotba kerülnek.
- **Elemzés:** Különböző elemzéseket végzek a szociális platformok profiljain és bejegyzésein, annak érdekében, hogy kiderítsem milyen tényezők játszanak közre a rövid formátumú tartalmak sikerességében.
- **Vizualizáció:** Az elemzések eredményeit szemléltetem Power BI segítségével.



4.4-es ábra: A rendszer funkcióinak ábrázolása folyamatábrával

4.3. A fejlesztés tervezett menete

A rendszer lefejlesztése az alábbi lépésekből fog állni:

- Sitemap-ek létrehozása a négy szociális platformon Web Scraper-ben vagy Octoparse-ban
- A vizsgálni kívánt profilok összegyűjtése és megadása a scrapereknek, majd adatgyűjtés
- Az adatbázis kialakítása
- Az adatok betöltése az adatbázisba, majd adattisztítás
- Az elemzési logika felépítése, majd az elemzés elvégzése
- Az eredmények vizualizálása

A lépések becsült időráfordítása Gantt diagrammon ábrázolva:

Feladat neve	1.hét	2.hét	3.hét	4.hét	5.hét	6.hét	7.hét	8.hét	9.hét	10.hét	11.hét	12.hét	13.hét
Sitemapek létrehozása													
Adatgyűjtés													
Az adatbázis kialakítása													
Adatbetöltés és tisztítás													
Elemzés													
Vizualizáció													
Jelölések													

4.5-ös ábra: Tervezett haladás Gantt diagrammon ábrázolva

4.3.1. A fejlesztés időbeosztásának magyarázata

A hat meghatározott lépés időben szigorúan egymás után következik, a rendszer természetéből kifolyólag. Amíg nincsenek sitemap-ek, addig nem lehet adatot gyűjteni, amíg nincsenek adatok, addig nem lehet adatot tisztítani, és így tovább.

A sitemap-ek létrehozása becsléseim szerint mindössze néhány napot fog igénybe venni, ezért erre mindössze egy hetet szánok.

Ezután viszont a vizsgálandó profilok kézzel történő kiválogatása és összegyűjtése időigényes lesz. Ezeket utána be kell tölteni a scraperekbe, majd várni, amíg végigfut az adatgyűjtés. Erre a feladatra egy hét időtartaleket terveztem, mivel előfordulhat, hogy valami miatt az egyik scraper hibásan fogja gyűjteni az adatokat, ezért lehetséges, hogy további konfigurációra lesz szükség.

Az adatbázis és az adattisztító program kialakítása viszonylag hosszabb folyamat lesz. A transzformáló metódusok megírása és tesztelése sok időt fog igénybe venni. Két hét valószínűleg elegendő lesz erre a lépésre.

Feltéve, hogy az előző lépésben felállított adatbázis és tisztítóprogram rendeltetésszerűen működnek, megkezdődhet az adatok betöltése és transzformálása.

A következő lépés lesz a legtöbb időt igénylő lépés, az elemzés. Itt kell az elemzés logikáját felépíteni, azaz meghatározni, hogy pontosan milyen jellegű elemzéseket fogok elvégezni az adatokon. Mivel ez a rész sok kísérletezéssel és lekérdezés írással jár, három hetet, és egy hétnyi időtartalékot rendeltem ehhez a lépéshez.

Miután megvannak a végleges adatok, már csak a vizualizálás maradt hátra, amit Power BI-al fogok végezni. Ennek időigénye 1-2 hét, a maradék időt a rendszer összes többi komponensének finomítására fogom fordítani.

5. FEJLESZTÉS

Az alábbi részben a Web Scraperrel történő adatgyűjtést, a nyers adatokat transzformáló program lefejlesztését, és az elemzés előkészületeit fogom bemutatni. Emellett a fejlesztés során felmerülő problémákra, és az ezekre kialakított megoldásokra is kitérek.

5.1. Adatgyűjtés

A videók tényleges tartalmát (kép, hang) nem elemzem a szakdolgozat keretein belül, mivel ehhez egy sokkal összetettebb rendszerre lenne szükségem. (A továbbfejlesztési lehetőségekre egy későbbi fejezetben kitérek). Ettől függetlenül a rendelkezésre álló mutatószámokból is számos következtetést le tudtam vonni.

Az eredeti terveim szerint négy szociális platformról szerettem volna adatokat gyűjteni, viszont ezzel kapcsolatban több akadállyal szembesültem. Először az Instagram felületére konfiguráltam fel a Web Scrapert, és el is tudtam indítani az adatgyűjtést, viszont néhány profil begyűjtése után az Instagram letiltotta a fiókomat, mivel automatizált tevékenységeket észlelt. Ezt követően a TikTokot próbáltam meg, viszont itt minden alkalommal amikor elindítottam a scraper-t, csak egy CAPTCHA teszt után engedett át a kívánt TikTok profilra. Ezt a tesztet a Web Scraper nem tudja kitölteni, mivel minden alkalommal más kódot kell beírnia, tehát a TikTok-os adatgyűjtés is lehetetlenné vált. Ezen a ponton a YouTube és a Facebook maradtak. Facebook-on viszont azzal a problémával találkoztam, hogy a Facebook Reels formátum nem tartalmaz dátum mezőt. Ez azt jelentette volna, hogy a begyűjtött videóadatok nem tartalmazznak feltöltési dátumot, ami nélkül csak nagyon limitált és felületes elemzéseket tudtam volna végezni, ezért a Facebook-ot is ki kellett zárnom az adatgyűjtésből. Végül a YouTube tartalmazott minden szükséges adatot, működött rajta a Web Scraper, és még CAPTCHA által sem volt védve. A fentebb taglalt okok miatt csak erről a platformról gyűjtöttem adatokat (80 csatorna, és kb. 20 ezer videó adatát).

Az adatgyűjtést szándékosan nem véletlenszerűen, hanem témákra bontva végeztem (Pl. főzés/ételek, technológia, videojátékok). Minden témában összeválogattam 20-50 tartalomgyártót, akik a saját szegmensükben viszonylag sikeresek, és egymáshoz képest hasonló tartalmakat készítenek. Ez a gyűjtési folyamatot leegyszerűsítette, valamint az elemzésnél is lehetővé tette azt, hogy az általános elemzések mellett téma-specifikus vizsgálatokat is végezzek.

5.1.1. A Web Scraper felkonfigurálásakor felmerülő problémák

A Web Scraper kezelőfelülete a böngészőn belül található, a forráskód vizsgálata opción belül. Itt állítottam be, hogy az adott weboldalakon mely linkeket nyissa meg, illetve mely HTML elemeket gyűjtse be a program. A YouTube felületén összeszedtem egy listába az 50 darab,

gyűjtésre kijelölt csatornát, és úgy állítottam be a programot, hogy az egyenként mindet megnyissa. Ezután a csatornán belül átnavigáljon a „YouTube Shorts” menüpontra, és ott tekerjen le az oldal legaljára, hogy minden videó betöltődjön. Ezt követően a Web Scraper egyenként 1-1 másodpercre megnyitotta az összes videót, majd amikor az összes HTML mező megjelent és ezeket beolvasta, ment tovább a következő videóra, majd a következő csatornára, amíg végig nem ért a listán.

Ez a felkonfigurálási folyamat viszont nem volt problémamentes, mivel a program sok esetben hibásan, vagy egyáltalán nem olvasott be egyes mezőket, vagy előfordult hogy egy 1700 videóval rendelkező csatornának csupán 100 videóján ment végig, majd leállt. Az ilyen és ehhez hasonló problémákat a konfiguráció változtatásával tudtam kiküszöbölni. A hiányos gyűjtés például azért fordult elő, mivel túl alacsony kérelmezési intervallumot (request interval) állítottam be a Web Scrapernek, azaz a program hamarabb próbálta beolvasni az egyes elemeket, mint ahogy azok betöltődtek. Amikor a 2 másodperces intervallumot 4 másodpercre emeltem, ez a probléma megoldódott.

5.2. Az adatok transzformálása

5.2.1. A projektek létrehozása és előkészületek

Az adatbázist és a táblákat Microsoft SQL Server Management Studio-ban hoztam létre, majd Visual Studio-ban készítettem egy konzol applikációt. Ezt követően a projektet egy kapcsolati karakterlánc (connection string) segítségével hozzákapcsoltam az adatbázishoz. [16] A fejlesztés folyamatát GitHubon is dokumentáltam.

A programot VS 2019-ben, .NET 5.0-ás verzióval fejlesztettem. Az alap konzol applikáció funkcióin kívül a Microsoft.Data.SqlClient és az NUnit NuGet csomagokat vettem igénybe, mivel ezekre szükségem volt az adatbázis szerkesztéséhez és a teszteléshez. Power BI-ban is létrehoztam egy üres projektet, amit hozzákapcsoltam az adatbázishoz.

5.2.2. Single-responsibility principle

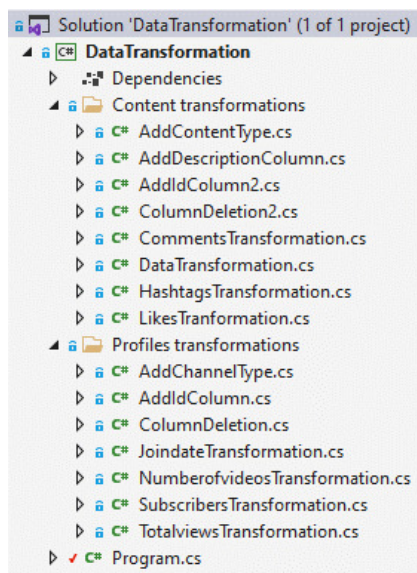
Az adatok tisztítását végző programot az egyetemen is tanult „Single-responsibility principle” (SRP) szerint fejlesztettem. Ennek lényege, hogy a program egyes egységei egy-egy kis feladatot látnak el. Ez a programozási elv azért volt célravezető az én esetemben, mivel a program sok egyszerű és könnyen elválasztható feladatot kellett hogy elvégezzen (egy-egy táblázat egy-egy oszlopán elvégzendő egyszerű transzformációk).

5.2.3. A program felépítése

Az SRP-nek megfelelően, sok kis osztályt hoztam létre, amelyek csupán 1-2 metódust tartalmaznak, és 1-1 oszlop transzformálásáért felelősek. Ezeket a metódusokat a program

futtatásakor egymás után sorban hívtam meg a Program.cs-ben, így ezek egyenként transzformálták az összes oszlopot a táblázatban. A metódusok működése alapvetően hasonló volt. Mindegyik kiolvasta az adott oszlop adatait, elvégezte az előírt transzformációt, majd visszaírta őket az adatbázisba.

Az alábbi képeken látható a kész program felépítése a Solution explorerben, valamint a transzformáló függvények meghívása a Program.cs-ben:



5.1-es ábra: A transzformáló program felépítése a Solution explorerben

```
//-----A PROFILES TÁBLA TRANSZFORMÁLÁSA-----//-----A CONTENT TÁBLA TRANSZFORMÁLÁSA-----
//-----A PROFILES TÁBLA TRANSZFORMÁLÁSA-----//-----A CONTENT TÁBLA TRANSZFORMÁLÁSA-----

// A profiles tábla felesleges oszlopainak törlése
ColumnDeletion columnDeletion = new ColumnDeletion(connectionString);
columnDeletion.DeleteColumns();
Console.WriteLine();

// Az Id oszlop létrehozása és feltöltése
AddIdColumn addIdColumn = new AddIdColumn(connectionString);
addIdColumn.AddIdentityColumn();
Console.WriteLine();

// A channeltype oszlop létrehozása és kitöltése
AddChannelType addChannelType = new AddChannelType(connectionString);
addChannelType.AddChannelTypeColumn();
Console.WriteLine();

// A subscribers oszlop formátumának transzformálása
var transformSubscribers = new TransformSubscribers(connectionString);
transformSubscribers.TransformSubscribersColumn();
Console.WriteLine();

// A numberofvideos oszlop formátumának transzformálása
NumberofvideosTransformation numberofvideosTransformation =
new NumberofvideosTransformation(connectionString);
numberofvideosTransformation.TransformNumberOfVideosColumn();
Console.WriteLine();

// A totalviews oszlop formátumának transzformálása
var transformTotalViews = new TransformTotalViews(connectionString);
transformTotalViews.TransformTotalViewsColumn();
Console.WriteLine();

// A joindate oszlop formátumának transzformálása
JoindateTransformation transformJoindate =
new JoindateTransformation(connectionString);
transformJoindate.TransformJoinDateColumn();
Console.WriteLine();

// A content tábla felesleges oszlopainak törlése
ColumnDeletion2 columnDeleter = new ColumnDeletion2(connectionString);
columnDeleter.DeleteColumns();
Console.WriteLine();

// Az Id oszlop hozzáadása és kitöltése
AddIdColumn2 addIdColumn2 = new AddIdColumn2(connectionString);
addIdColumn2.AddIdColumn();
Console.WriteLine();

// Az contenttype oszlop létrehozása és kitöltése
var contentTypeAdder = new AddContentType(connectionString);
contentTypeAdder.AddAndPopulateContentTypeColumn();
Console.WriteLine();

// A likes oszlop formátumának transzformálása
var likesTransformation = new LikesTransformation(connectionString);
likesTransformation.TransformLikesColumn();
Console.WriteLine();

// A comments oszlop formátumának transzformálása
var commentsTransformation =
new CommentsTransformation(connectionString);
commentsTransformation.TransformCommentsColumn();
Console.WriteLine();

// A hashtags oszlop kitöltése
var hashtagCounter = new HashtagsTransformation(connectionString);
hashtagCounter.CountHashtags();
Console.WriteLine();

// A description oszlop létrehozása
var addDescriptionColumn = new AddDescriptionColumn(connectionString);
addDescriptionColumn.AddColumn();
Console.WriteLine();

// A date oszlop transzformálása
var dateTransformation = new DateTransformation(connectionString);
dateTransformation.TransformDateColumn();
Console.WriteLine();
```

5.2-es ábra: A táblákat transzformáló függvények meghívása

5.2.4. A transzformációk ismertetése

A program 16 osztályt, és ennél még több metódust tartalmaz. A metódusok által végzett transzformációkat az alábbi felsorolásban ismertetem, majd ezt követően két képpel szemléltetem a nyers és transzformált adatok közötti különbségeket.

A Web Scraper-ből kiexportált adatok a „yt_profiles_raw” és „yt_content_raw” táblákba kerültek betöltésre. Az előbbi a tartalomgyártók profil adatait tartalmazza, az utóbbi pedig az általuk feltöltött videók adatait.

A yt_profiles_raw transzformációinak ismertetése:

- Az első metódus a Web Scraper által feleslegesen begyűjtött adatok törlését végzi. Ezek a „web_scraper_order”, „web_scraper_start_url”, és „bio_href” oszlopok, melyek a scraper által megnyitott linkeket tartalmazzák (Az elemzéshez ezekre nincs szükség).
- A második metódus létrehoz egy „channeltype” nevű oszlopot, amely az adott tartalomgyártó témáját, avagy szegmensét tartalmazza, amin belül készíti a videóit. Ezt a létrehozás után ki is tölti a metódus (Cooking, Gaming, vagy Tech értékekkel).
- A harmadik metódus a „totalviews” oszlop formátumát transzformálja. Ez az oszlop az adott tartalomgyártó összes megtekintését tartalmazza, amelyet a YouTube felülete elválasztó karakterekkel tagolt formátumban ábrázolt. A metódus kitörli a számok közötti üres karaktereket, hogy később szöveg helyett számként lehessen felhasználni az adatokat.
- A negyedik és ötödik metódusok a „subscribers” és a „numberofvideos” oszlopok átalakítását végzik. Ezek az oszlopok a feliratkozók és a videók számát tartalmazzák, szintén a YouTube sajátos formázásában (Például 10 E, vagy 10 M). Itt a program az E és M betűket ezres és milliós szorzókra cseréli, majd egy szövegből törtszámmá konvertálást követően elvégzi a szorzást. Ez az átalakítás megint azért fontos, mert így szöveg helyett számként kezelhetők az adatok.
- A „joindate” oszlopban a csatorna létrehozásának dátuma szerepel. A program itt csupán a „jan, febr, márc” formázást cseréli le „01. 02. 03.”-ra, hogy a Power BI dátumként tudja kezelni a mezőket
- Végül az AddIdColumn.cs létrehoz egy Id oszlopot, majd egyesével inkrementáló Id értékekkel látja el az összes rekordot

A programot úgy alakítottam ki, hogy ugyanilyen adatok jövőbeli betöltése ugyanolyan eredményre vezessen, tehát a folyamat skálázható legyen (csak az átalakítandó tábla nevét kell átírni).

A yt_content_raw transzformációinak ismertetése:

Sok transzformáció yt_profiles_raw táblához hasonlóan történt, ezeket nem részletezem még egyszer, csak az új átalakítások kerülnek felsorolásra:

- A „hashtags” oszlopot úgy tölti fel a program, hogy megszámolja a caption (videó címe) oszlopban található „#” karaktereket.
- A Web Scraper a feltöltés dátuma HTML elemet nem tudta konzisztensen begyűjteni, viszont ennek a parent elemét igen. Ez viszont a következő formátumban jelent meg: "4,8 EKedvelések111 039Megtekintések2021.jan. 11.". Ebből csak a végén lévő dátumra volt szükségem, ezért úgy írtam meg a metódust, hogy ezen kívül mindent töröljön. Ezután a joindate oszlophoz hasonlóan a metódus kicseréli a betűkkel kiírt hónapokat a szám megfelelőjükre.
- Voltak olyan videók, ahol a kommentek le voltak tiltva, ezeket a sorokat a program törölte, mivel az elemzésem egyik témája az „engagement rate”, amely egy olyan mutatószám, amit a like-ok és kommentek összegének a megtekintések számával való leosztásával kapunk meg. Ez az érték pontatlan lett volna ha a kommentek számát nem tudom beolvasni.

A alábbi képek az adatok transzformáció előtti és utáni állapotát szemléltetik:

web-scraper-start-url	reel opener	reel opener-href	uploader	caption	date	views	likes	comments
https://www.youtube.com/England - Spain,Wembley,Lallan	https://www.youtube.com/England-Spain Wembley stadium,Alvaro Morata	https://www.youtube.com/England-Spain Wembley stadium,Alvaro Morata	https://www.youtube.com/England-Spain Wembley stadium,Alvaro Morata	https://www.youtube.com/England-Spain Wembley stadium,Alvaro Morata	2016.nov.16.	28 619	1,6 E	113
https://www.youtube.com/Have you tried this life hack ? #lifehacks #cheese #shorts	https://www.youtube.com/Have you tried this life hack ? #lifehacks #cheese #shorts	https://www.youtube.com/Have you tried this life hack ? #lifehacks #cheese #shorts	https://www.youtube.com/Have you tried this life hack ? #lifehacks #cheese #shorts	https://www.youtube.com/Have you tried this life hack ? #lifehacks #cheese #shorts	2021.dec.6.	25 701	1,3 E	52
https://www.youtube.com/#skeleton #funny #comedy #howtoplease subscribe	https://www.youtube.com/#skeleton #funny #comedy #howtoplease subscribe	https://www.youtube.com/#skeleton #funny #comedy #howtoplease subscribe	https://www.youtube.com/#skeleton #funny #comedy #howtoplease subscribe	https://www.youtube.com/#skeleton #funny #comedy #howtoplease subscribe	2021.dec.7.	25 135	1 E	23
https://www.youtube.com/Eggciting#egg	https://www.youtube.com/Eggciting#egg	https://www.youtube.com/Eggciting#egg	https://www.youtube.com/Eggciting#egg	https://www.youtube.com/Eggciting#egg	2021.dec.7.	23 960	942	25
https://www.youtube.com/The golden steak #steak #shorts #adrianghervan #viral	https://www.youtube.com/The golden steak #steak #shorts #adrianghervan #viral	https://www.youtube.com/The golden steak #steak #shorts #adrianghervan #viral	https://www.youtube.com/The golden steak #steak #shorts #adrianghervan #viral	https://www.youtube.com/The golden steak #steak #shorts #adrianghervan #viral	2021.dec.8.	20 419	1,4 E	49
https://www.youtube.com/Digestive biscuits + chocolate + marshmallow =	https://www.youtube.com/Digestive biscuits + chocolate + marshmallow =	https://www.youtube.com/Digestive biscuits + chocolate + marshmallow =	https://www.youtube.com/Digestive biscuits + chocolate + marshmallow =	https://www.youtube.com/Digestive biscuits + chocolate + marshmallow =	2021.dec.8.	16 138	820	18
https://www.youtube.com/Steak	https://www.youtube.com/Steak	https://www.youtube.com/Steak	https://www.youtube.com/Steak	https://www.youtube.com/Steak	2021.dec.9.	6161	509	10
https://www.youtube.com/10 December 2021	https://www.youtube.com/10 December 2021	https://www.youtube.com/10 December 2021	https://www.youtube.com/10 December 2021	https://www.youtube.com/10 December 2021	2021.dec.10.	10 487	557	15
https://www.youtube.com/10 December 2021	https://www.youtube.com/10 December 2021	https://www.youtube.com/10 December 2021	https://www.youtube.com/10 December 2021	https://www.youtube.com/10 December 2021	2021.dec.10.	12 520	389	9
https://www.youtube.com/12 December 2021	https://www.youtube.com/12 December 2021	https://www.youtube.com/12 December 2021	https://www.youtube.com/12 December 2021	https://www.youtube.com/12 December 2021	2021.dec.12.	5553	289	19
https://www.youtube.com/14 December 2021	https://www.youtube.com/14 December 2021	https://www.youtube.com/14 December 2021	https://www.youtube.com/14 December 2021	https://www.youtube.com/14 December 2021	2021.dec.14.	73 813	1,5 E	11
https://www.youtube.com/Sugar crush	https://www.youtube.com/Sugar crush	https://www.youtube.com/Sugar crush	https://www.youtube.com/Sugar crush	https://www.youtube.com/Sugar crush	2021.dec.16.	8822	331	14
https://www.youtube.com/Oreo chocolate toastie	https://www.youtube.com/Oreo chocolate toastie	https://www.youtube.com/Oreo chocolate toastie	https://www.youtube.com/Oreo chocolate toastie	https://www.youtube.com/Oreo chocolate toastie	2021.dec.16.	8516	354	4
https://www.youtube.com/Squidgame	https://www.youtube.com/Squidgame	https://www.youtube.com/Squidgame	https://www.youtube.com/Squidgame	https://www.youtube.com/Squidgame	2021.dec.16.	30 564	739	17
https://www.youtube.com/I know what you was thinking	https://www.youtube.com/I know what you was thinking	https://www.youtube.com/I know what you was thinking	https://www.youtube.com/I know what you was thinking	https://www.youtube.com/I know what you was thinking	2021.dec.16.	10 551	254	22
https://www.youtube.com/Tiny hands toastie	https://www.youtube.com/Tiny hands toastie	https://www.youtube.com/Tiny hands toastie	https://www.youtube.com/Tiny hands toastie	https://www.youtube.com/Tiny hands toastie	2021.dec.16.	6962	284	2

5.3-as ábra: A yt_content tábla tartalma a transzformáció előtt

Id	contentty	uploader	caption	date	views	likes	comments	description	hashtags
1	Cooking	@adrianghervan	England - Spain,Wembley,Lallana scores penalty	2016.nov.16.	28619	1600	113	1	0
2	Cooking	@adrianghervan	England-Spain Wembley stadium,Alvaro Morata	2016.nov.16.	8152	641	18	1	0
3	Cooking	@adrianghervan	Have you tried this life hack ? #lifehacks #cheese #shorts	2021.dec.6.	25701	1300	52	0	3
4	Cooking	@adrianghervan	#skeleton #funny #comedy #howtoplease subscribe	2021.dec.7.	25135	1000	23	0	4
5	Cooking	@adrianghervan	Eggciting#egg	2021.dec.7.	23960	942	25	0	1
6	Cooking	@adrianghervan	The golden steak #steak #shorts #adrianghervan #viral	2021.dec.8.	20419	1400	49	0	4
7	Cooking	@adrianghervan	Digestive biscuits + chocolate + marshmallow =	2021.dec.8.	16138	820	18	0	0
8	Cooking	@adrianghervan	Steak	2021.dec.9.	6161	509	10	0	0
9	Cooking	@adrianghervan	10 December 2021	2021.dec.10.	10487	557	15	0	0
10	Cooking	@adrianghervan	10 December 2021	2021.dec.10.	12520	389	9	0	0
11	Cooking	@adrianghervan	12 December 2021	2021.dec.12.	5553	289	19	0	0
12	Cooking	@adrianghervan	14 December 2021	2021.dec.14.	73813	1500	11	0	0
13	Cooking	@adrianghervan	Sugar crush	2021.dec.16.	8822	331	14	0	0
14	Cooking	@adrianghervan	Oreo chocolate toastie	2021.dec.16.	8516	354	4	0	0
15	Cooking	@adrianghervan	Squidgame	2021.dec.16.	30564	739	17	0	0
16	Cooking	@adrianghervan	I know what you was thinking	2021.dec.16.	10551	254	22	0	0
17	Cooking	@adrianghervan	Tiny hands toastie	2021.dec.16.	6962	284	2	0	0

5.4-es ábra: A yt_content tábla tartalma a transzformáció után

6. TESZTELÉS

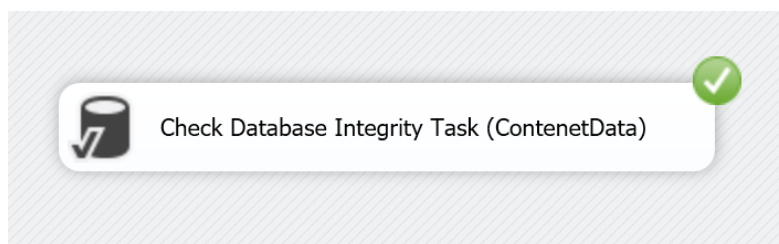
A tesztelési folyamatot a tesztek típusának meghatározásával kezdtem. Mivel a rendszerem három nagyobb részből áll (adatgyűjtés, adattisztítás, adatvizualizáció), és ebből az elsőt és a harmadikat nem igazán tudom tesztelni, főleg az adattisztító programhoz írtam unit teszteket. Ezeknek a teszteknek a lényege az volt, hogy a metódusok működését egyenként, modulárisan ellenőrizzék.

6.1. Az adatok integritásának ellenőrzése

A unit tesztek előtt végeztem egy adatbázis integritás ellenőrzést a Visual Studio Integrated Services Project projekttypusával.

[17] A „Check Database Integrity Task” célja, hogy ellenőrizze egy kijelölt adatbázis összes objektumának (Pl. a rendszer vagy felhasználó által létrehozott táblák, nézetek) strukturális integritását. Hibák vagy hiányosságok esetén készít egy riportot, hogy ezeket ki lehessen javítani.

Ennek az ellenőrzésnek az elvégzéséhez egyszerűen létrehoztam egy Integrated Services Project-et, majd a „kapcsolatok kezelése” résznél hozzáadtam az adatbázisomat, majd lefuttattam a task-ot. Az alábbi képen látható, hogy az adatbázisomon futtatott ellenőrzés sikeres volt, tehát az adatbázis integritása fennáll.



6.1-es ábra: Az adatbázis integritásának ellenőrzése

6.2. Unit tesztelés

A unit tesztelés azért egy megfelelő tesztelési típus a szoftveremhez, mivel rendeltetésszerű használat esetén mindig egy specifikus formátumban érkeznek az adatok, amelyeken mindig ugyanazokat a transzformációkat kell elvégezniük a metódusoknak. Ez azt jelenti, hogy ha a unit tesztek során a metódusok a nyers adatokat megfelelően alakítják át, akkor azzal bizonyítható, hogy azok megfelelően működnek.

6.2.1. A unit tesztek működése

A unit teszteknek alapvetően három fő lépése van. Az első lépés (arrange) lényege a teszt előkészítése, azaz a szükséges objektumok létrehozása, állapotaiknak beállítása, valamint a bemeneti adatok előkészítése. Mivel az én szoftverem szöveget transzformál, ezért ebben a lépésben mindig létrehoztam egy string típusú változót, amelynek beállítottam egy értéket.

A második lépés (act) a metódus meghívása az előző lépésben megadott bemeneti adatokkal. Ez lényegében a teszt futtatása. A teszt eredménye lementésre kerül.

A harmadik lépésben (assert) a program összehasonlítja a teszt eredményét az általunk előre megadott elvárt eredménnyel. Az én esetemben megadtam hogy hogy kéne kinéznie az átalakított szövegnek, és a program ezzel hasonlította össze metódus meghívása után keletkezett eredményt. Ha ezek egyeztek, akkor a teszt sikeres volt.

Az alábbi bekezdésekben néhány tesztet részletesen elmagyarázok, majd a fejezet végén egy táblázattal szemléltetem a tesztek eredményét.

6.2.2. A joindate oszlopot átalakító metódus tesztelése

Ez a függvény lényegében annyit csinál, hogy a bemeneti string első 21 karakterét törli, illetve ha a string ennél rövidebb, akkor kivételt dob, majd kiírja hogy a mező már át lett alakítva.

A tesztek előtt létrehoztam egy „SetUp” függvényt, amely létrehozott egy példányt a transzformáló függvényből. Ennek a függvénynek null-t állítottam be argumentumnak, mivel a teszteléshez nincs szükség a kapcsolati karakterláncra.

Az első teszt a függvény rendeltetésszerű használatának működését ellenőrzi. Itt a bemenet a következő: "Csatlakozás dátuma: 2015. ápr. 16.". Erre a bemenetre az elvárt eredmény pedig: „2015. ápr. 16.”

A második teszt azt az esetet szimulálja, amikor egy érték egy korábbi futtatás alatt már transzformálva lett (azaz a bemeneti string rövidebb mint 21 karakter). A bemenet a következő: „Csatlakozás dátuma:”. Erre az elvárt eredmény egy kivétel dobása, és a „joindate already transformed” üzenet kiírása a konzolra.

6.2.3. A totalviews oszlopot átalakító metódus tesztelése

A TransformTotalViews függvény szöveg formátumú, szóközökkel tagolt számokat alakít át úgy, hogy eltávolítja a tagolást, majd számmá konvertálja őket.

Az első teszt a függvény konvertáló képességét teszteli. Bemenetként egy string típusú számot kap: „12345”. Az elvárt eredmény pedig ugyanez a karakterlánc, csak long formátumban (12345).

A második teszt a függvény másodlagos funkcióját teszteli, ami az összes nem numerikus karakter törlése. Ez a lépés a konverzió előtt történik, és azért van rá szükség, mert ha a karakterláncok vesszőket vagy szóközőket tartalmaznak, akkor a számmá alakítás sikertelen lesz. Itt a következő a bemenet: „12, 345 megtekintés”. Ez a bemenet tartalmaz szóközőket, betűket és speciális karaktereket is. Az elvárt eredmény az előző teszthez hasonlóan ismét „12345”.

A metódus harmadik és negyedik tesztje az üres stringek és null értékek kezelését teszteli. A bemenet az előbbinél „” (üres string), az utóbbinál null. Az elvárt eredmény az előbbinél egy formátum kivétel dobása, az utóbbinál pedig egy null-argumentum kivétel dobása.

6.2.4. A subscribers oszlopot átalakító metódus tesztelése

A TransformSubscribers metódus dolga, hogy a YouTube sajátos formázását (10 E feliratkozó, 24 M feliratkozó” átalakítsa általános szám formátummá (10000, 24000000). A metódus működését az 5.2.4.-es alfejezetben mutatom be részletesen.

Az első teszt egy egyszerű konverzió teszt. A bemenet egy string típusú szám: „12345”. Az elvárt eredmény pedig ugyanez a szám, csak int típusban.

A második teszt a milliós szorzók helyes átalakítását és a szorzás elvégzését teszteli. A bemenet a következő: „1,23 M”. Itt a függvény helyes működés szerint észreveszi az „M” karaktert, és a szorzó változót 1000000-ra állítja. Ezután eltünteti minden karaktert a számokon és vesszőkön kívül, majd átváltja a számot törtszámmá, és összeszorozza a szorzó változóval. Abban az esetben amikor nincs „E”, vagy „M” az eredeti szövegben, akkor a szorzó változó 1 lesz, így nem történik szorzás. A teszt elvárt eredménye 1230000.

A harmadik teszt ugyanez a folyamat, csak az „M” karakterek helyett az „E” karaktereket, azaz az 1000-es szorzókat vizsgálja.

A negyedik teszt azt az esetet szimulálja, amikor nincs semmilyen szorzó a bemenetben, viszont törlendő karakterek igen. Ilyen például a következő: „123 feliratkozó”. Itt nincs szükség semmilyen szorzás elvégzésére, csupán a „ feliratkozó” szöveget kell törölnie a programnak. Az elvárt kimenet 123.

Az ötödik teszt az üres bemeneti stringeket teszteli, az elvárt eredménye egy kivétel dobása.

A két tábla további oszlopait az eddigiekhez hasonló módon teszteltem, így ezeket nem részletezem tovább. A következő oldalon látható táblázat szemlélteti a teszteredményeket:

6.3. A tesztek eredményeinek összefoglalása

#	Tesztelendő funkció	Elvárt eredmény	Teszt sikeressége
1.	Az adatbázis integritásának ellenőrzése	A „Check Database Integrity Task” sikeresen lefut az adatbázison	✓
2.	A date oszlopot transzformáló metódus működése üres string bemenettel	Üres string bemenetre kivételt kell hogy dobjon	✓
3.	A date oszlopot transzformáló metódus működése helyes bemenettel	Helyes bemenetre helyesen végzi el az átalakítást	✓
4.	A date oszlopot transzformáló metódus működése, amikor másodszor futtatjuk ugyanazon a mezőn	Felismeri hogy a transzformáció már megtörtént, nem változtat semmit a mezőn	✓
5.	A date oszlopot transzformáló metódus működése, amikor a bemenet felesleges szóközöket tartalmaz	A metódus eltávolítja a felesleges szóközöket, majd elvégzi az átalakítást	✓
6.	A joindate oszlopot transzformáló metódus működése 21 karakternél hosszabb bemenettel	A metódus eltávolítja az első 21 karaktert	✓
7.	A joindate oszlopot transzformáló metódus működése 21 karakternél rövidebb bemenettel	A metódus felismeri hogy a transzformáció már megtörtént, ezért nem változtat semmit az mezőn	✓
8.	A subscribers oszlopot transzformáló metódus működése üres string bemenettel	Üres string bemenetre kivételt kell hogy dobjon	✓
9.	A subscribers oszlopot transzformáló metódus konvertálási képessége	Sikeresen elvégzi a szöveg-szám konverziót	✓
10.	A subscribers oszlopot transzformáló metódus működése ezres szorzó esetén	Elmenti a szorzót, kitörli az összes karaktert ami nem szám vagy vessző, átkonvertálja a szöveget törtszámmá, majd elvégzi az ezres szorzást	✓
11.	A subscribers oszlopot transzformáló metódus működése milliós szorzó esetén	Elmenti a szorzót, kitörli az összes karaktert ami nem szám vagy vessző, átkonvertálja a szöveget törtszámmá, majd elvégzi a milliós szorzást	✓
12.	A subscribers oszlopot transzformáló metódus működése nem numerikus karaktereket tartalmazó bemenet esetén	A metódus eltávolítja a nem numerikus karaktereket, majd elvégzi a konvertálást és szorzást	✓
13.	A totalviews oszlopot transzformáló metódus működése üres string bemenettel	Üres string bemenetre kivételt kell hogy dobjon	✓
14.	A totalviews oszlopot transzformáló metódus működése null bemenettel	Null bemenetre kivételt kell hogy dobjon	✓
15.	A totalviews oszlopot transzformáló metódus konvertálási képessége	Sikeresen elvégzi a szöveg-szám konverziót	✓
16.	A totalviews oszlopot transzformáló metódus működése vegyes (numerikus és nem numerikus) bemenet esetén	A metódus törli a nem numerikus karaktereket, majd elvégzi az átváltást	✓

6.1-es táblázat: A programon végzett tesztek eredményei

A teszteket a program solution-jén belül hoztam létre egy külön teszt projektben. A tesztek fejlesztését szintén dokumentáltam GitHub-on is. Az alábbi ábra egy képernyőkép a Visual Studio Test Explorer felületéről, az összes teszt a futtatását követően:

Test	Duration	Traits	Error Message
▲ ✓ DataTransformation.Tests (15)	37 ms		
▲ ✓ DataTransformation.Tests (15)	37 ms		
▲ ✓ DateTransformationTests (4)	12 ms		
✓ EmptyInput_ReturnsEmptyString	9 ms		
✓ InputContainsMegtekintések_ReturnsRemainingDate	3 ms		
✓ InputDoesNotContainMegtekintések_ReturnsOriginalString	< 1 ms		
✓ InputHasSpacesAroundMegtekintések_RemovesPrefixAndTrimSpaces	< 1 ms		
▲ ✓ JoindateTransformationTests (2)	12 ms		
✓ InputTooShort_ThrowsArgumentException	12 ms		
✓ ValidInput_RemovesFirst21Characters	< 1 ms		
▲ ✓ TransformSubscribersTests (5)	13 ms		
✓ EmptyString_ThrowsFormatException	9 ms		
✓ StringWithInvalidCharacters_RemovesNonNumericCharacters	4 ms		
✓ StringWithMultiplierE_ReturnsParsedInteger	< 1 ms		
✓ StringWithMultiplierM_ReturnsParsedInteger	< 1 ms		
✓ ValidNumericString_ReturnsParsedInteger	< 1 ms		
▲ ✓ TransformTotalViewsTests (4)	< 1 ms		
✓ EmptyString_ThrowsFormatException	< 1 ms		
✓ NullInput_ThrowsArgumentNullException	< 1 ms		
✓ NumericAndTextString_ReturnsParsedLong	< 1 ms		
✓ ValidNumericString_ReturnsParsedLong	< 1 ms		

6.2-es ábra: A tesztek eredménye a Test Explorer-ben

7. ELEMZÉS

Miután a táblázatokon lefuttattam a transzformáló programot, létrehoztam a kapcsolatot a Microsoft Power BI projekt és az SSMS között, majd megkezdtem az elemzést.

Az elemzésem célja az volt, hogy a az adatok vizsgálatával kiderítsem hogy milyen szempontok befolyásolják a rövid formátumú videók sikerességét, illetve mik azok a „jó gyakorlatok”, melyeket a legsikeresebb tartalomgyártók alkalmaznak.

Az elemzések sorrendjében nincs különösebb szabályszerűség, csupán végigmentem az összes begyűjtött adaton az adatbázisomban, és megpróbáltam a legtöbb logikus következtetést levonni belőlük.

Az elemzések általában mindig egy kérdés feltevésével kezdődtek. Számít hogy milyen hosszú és mit tartalmaz egy videó címe? Számít a hashtagek száma? Számít hogy milyen hosszú a biográfia? Van-e különbség a különböző szociális platformok átlagos demográfiája között? Milyen alapelveket érdemes követni annak érdekében, hogy minél sikeresebb legyen egy rövid formátumú videó?

Ezeknek a kérdéseknek és felvetéseknek jártam utána az elemzéseim során, majd a meglátásaimat különféle grafikonokkal illusztráltam.

Megjegyzés: Az elemzéseket legjobb tudásom szerint és minél több forrásanyag ismeretében végeztem. Megpróbáltam minden felvetésemnek utána nézni, hogy valaki más már végzett-e alaposabb elemzést a témában, és ezeket is számításba vettem. Nem állítom hogy minden következtetés amit levontam minden esetben tévedhetetlen és helyes. Az olyan következtetéseket amikben nem voltam teljesen biztos hogy megalapozottak, nem írtam bele az elemzésbe.

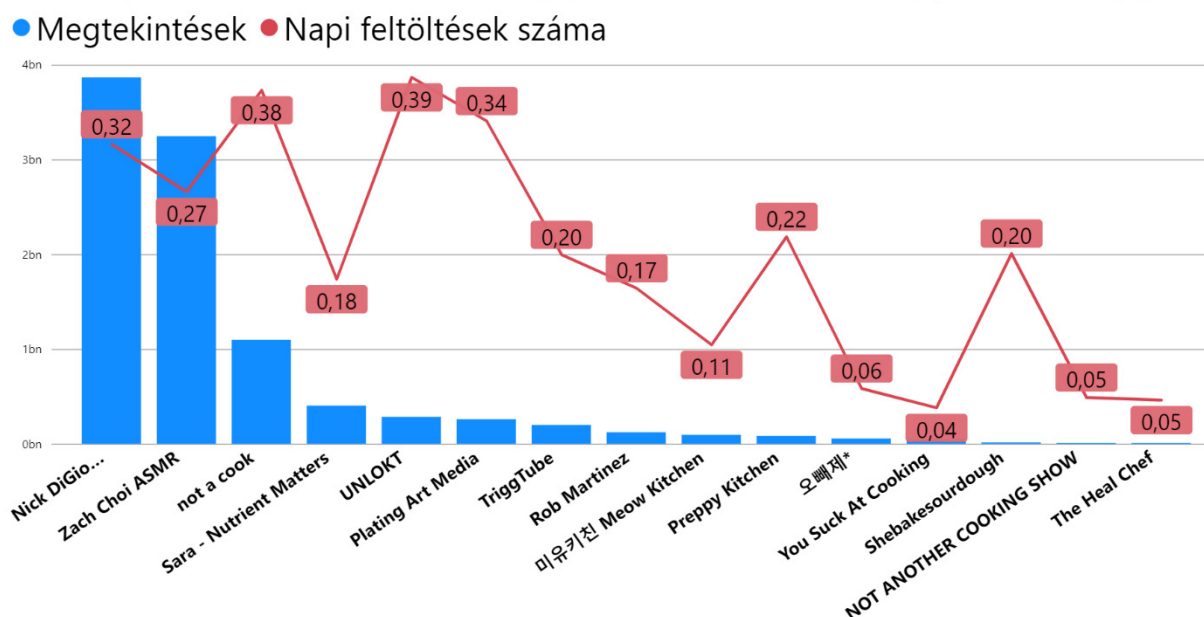
7.1. Az elemzések ismertetése

Az alábbi elemzések eredményei olyan módszertanokat és stratégiákat fedtek fel, amelyek tartalomtípustól függetlenül, minden platformon, minden témában alkalmazhatóak. Ilyen például a feltöltési gyakoriság kérdése, a cím hossza és tartalma, a hashtagek helyes használata és a naprakészség jelentősége.

7.1.1. Milyen gyakran érdemes feltölteni?

Magától értetődő, hogy egy tartalomgyártó nézettsége csak akkor tud igazán jelentősen növekedni, ha gyakran és nagy mennyiségben tölt fel új tartalmakat. A lenti grafikon azt szemlélteti, hogy a napi feltöltések száma összefüggésben áll a megtekintések számával.

A megtekintések összefüggése a feltöltési gyakorisággal



7.1-es ábra: A megtekintések összefüggése a feltöltési gyakorisággal

A kék oszlopok jelzik a 15 legnézettebb főzéssel/étellel kapcsolatos YouTube csatorna megtekintéseinek számát az elmúlt két évben, csökkenő sorrendben. A piros vonal és a pirossal kiemelt számok pedig az ugyanebben az időszakban naponta feltöltött videók számát szemléltetik. **Leolvasható, hogy (néhány érték kivételével) a feltöltések rendszerességének növelése a megtekintések növekedését vonja maga után, azaz fontos a konzisztencia.**

A közönség szemszögéből nézve is érthető, hogy minél gyakrabban látnak egy alkotót az adott platformon, annál nagyobb eséllyel fogják bekövetni, illetve like-olni a videóit.

7.1.2. Számít egy videó címének hossza?

A videó címének optimális hossza egy elég összetett kérdés, amire nincs egyszerű válasz. Sok forrásnak utána néztem, viszont majdnem mindenhol más intervallumot adtak meg „optimális”-nak, és nem nagyon találtam adatokkal alátámasztott elemzéseket.

Egy fontos aspektus volt ami többször megjelent, ez a telefonok kijelzői által megjeleníthető maximális karakterszám volt. [18] A YouTube Shorts formátum telefonokra lett optimalizálva, ezt a statisztikák is alátámasztják, mivel a YouTube forgalmának 70%-a telefonon történik.

A Shorts címek 0 és 100 karakter között lehetnek, a telefonok kijelzői viszont legtöbb esetben maximálisan 40 karaktert képesek megjeleníteni. Ez a határ felett a YouTube lerövidíti a címet. Ez azt jelenti, hogy ennél hosszabb címeket nem érdemes használni, mivel a felhasználók nagyrésze nem fogja külön lenyitni és elolvasni a teljes címet.

Ennek az információnak az ismeretében megnéztem hogy az én 20 ezer videós adathalmazomon mennyire igaz ez az állítás. A videókat szétválogattam címhossz alapján, 0 és 100 között, tízesével, majd megnéztem hogy az egyes hossz kategóriák videói hány megtekintést és like-ot szereztek.

A videók összmegtekintésének 67%-át a 40 karakternél rövidebb címhosszal rendelkező videók szereztek (0-10, 11-20, 21-30, 31-40 címhossz kategóriák). A videók összeadott like számának pedig 71%-át tették ki a 40 karakternél rövidebb címhosszal rendelkező videók. Nagyon hasonló arány jelent meg amikor a kategóriák megtekintéseinek és like-jainak összege helyett helyett a kategóriák megtekintéseinek és like-jainak az átlagát vizsgáltam.

Ezek szerint tényleg érdemes 40 karakter alatt tartani a videók címének hosszát.

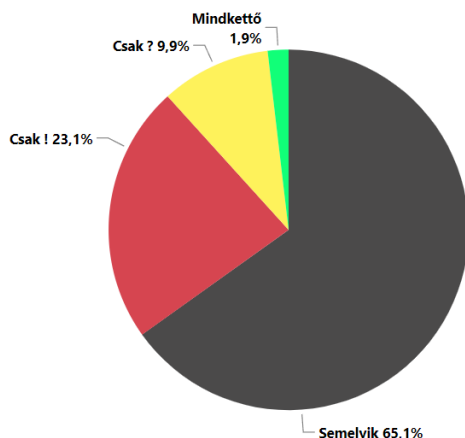
7.1.3. Számít egy videó címének tartalma?

A YouTube videók címei tartalmazhatnak kis- és nagybetűket, számokat, speciális karaktereket és még emoji-kat is. Mindegyik karaktertípus használatát hosszan lehetne elemezni, viszont ebben az elemzésben a kérdőjelek és felkiáltójelek használatára fogok fókuszálni. [19] [20] Több forrás is azt állítja, hogyha a videók címében kérdéseket helyezünk el, azzal könnyebb felkelteni az emberek érdeklődését, mint egy egyszerű kijelentéssel.

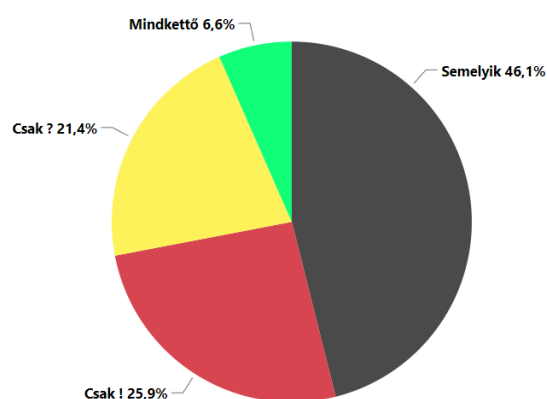
Ennek a megvizsgálására létrehoztam két új oszlopot a yt_content táblába, amelyből az egyik a címben szereplő kérdőjelek, illetve felkiáltójelek jelenlétét ellenőrzi. Ezek alapján négy csoportra bontottam a videókat. Ezek a címükben csak kérdőjelet, csak felkiáltójelet, semelyiket, vagy mindkettőt tartalmazó videók voltak.

A vagy kérdőjelet, vagy felkiáltójelet, vagy mindkettőt tartalmazó videók száma csupán az összes videó 35%-át tették ki, viszont ugyanezek a videók az össznézettség 54%-át szerezték meg.

Az címükben mondatvégi írásjeleket tartalmazó videók számának aránya



Az címükben mondatvégi írásjeleket tartalmazó videók nézettségének aránya



7.2-es ábra: A mondatvégi írásjeleket tartalmazó videók elemzése

A diagramokról leolvasható, hogy a videócímekben viszonylag kevészer használtak kérdőjeleket és felkiáltójeleket, viszont ezek a videók több megtekintést generáltak.

Ez arra utalhat, hogy tényleg érdemes kérdéseket (vagy felkiáltójeleket) rakni a Shorts videók címébe, mivel ezzel növelni lehet a nézettséget.

7.1.4. Mi az az engagement rate, és miért jelentős?

Az engagement rate egy olyan arányszám, amely azt mutatja, hogy egy bejegyzés vagy videó megtekintéseinek mekkora aránya eredményezett interakciót (like, komment, repost, megosztás). Többféle engagement rate létezik, én az elemzésemben a „basic engagement rate”-re fogok kitérni, melynek kiszámítása a következő: a megtekintések számát leosztjuk a like-ok és kommentek számának összegével, majd megszorozzuk 100-al. Ez egy százalékot fog eredményezni.

Az engagement rate azért nagyon fontos az internetes tartalmakban, mivel az egyes platformok algoritmusai részben ez a mérőszám alapján döntenek el, hogy mely tartalmakat promótáljanak szélesebb közönségek előtt. Ez azt jelenti hogy az engagement rate egyenesen arányos az eléréssel.

Volt egy olyan felvetésem, hogy a címükben kérdőjelet vagy felkiáltójelet tartalmazó videók engagement rate-je magasabb lesz, mint a többi videóé. Ennek kiderítésére ismét szétbontottam a videókat négy csoportra a kérdőjelek és felkiáltójelek alapján. Az első csoportba azok a videók kerültek, amelyek címe nem tartalmazott se kérdőjelet, se felkiáltójelet. Ennek a csoportnak az átlagos engagement rate-je 5,29% volt. A következő csoport a címében csak kérdőjelet tartalmazó videók voltak, ezek átlagos engagement rate-je 5,36% volt. A harmadik a címükben csak felkiáltójelet tartalmazó videók voltak, 5,86%-os engagement rate-tel. A negyedik csoport pedig a címükben kérdőjelet és felkiáltójelet is tartalmazó videók voltak, 5,93%-os engagement rate-tel.

Annak ellenére hogy csak kis mértékű különbségek vannak az értékek között, felismerhető egy trend, miszerint az „érdekesebb” (kérdőjelet vagy felkiáltójelet tartalmazó) címekkel ellátott videók több interakciót eredményeznek. Egy 1%-os engagement rate javulás nem tűnik soknak, viszont például egy 10 millió megtekintéssel rendelkező videó esetében ez 10 ezerrel több kommentet és like-ot jelent, ami már jelentős hatással van az algoritmusra.

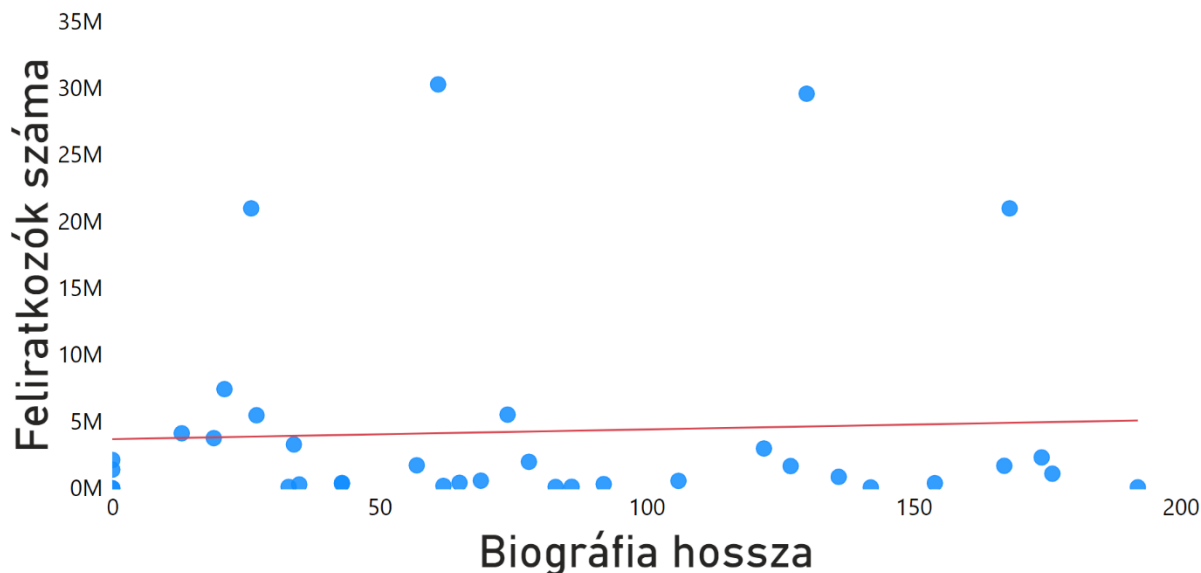
Megjegyzés: A YouTube Shorts felülete nem írja ki a megosztások vagy lementések számát, ezért ezt nem tudtam beleszámolni a számításba. Ezekkel együtt valószínűleg mind a négy kategória engagement rate-je magasabb százalékot mutatna.

A levonható következtetés az, hogy a címben (és ezzel együtt a videón belül is) érdemes kérdések feltevésével, vagy más módon bevonni a közönséget a beszélgetésbe.

7.1.5. Számít a biográfia hossza?

Minden szociális platformon meg lehet adni egy biográfiát, ami egy rövid leírásként funkcionál. Ebben az elemzésben azt derítem ki, hogy a biográfia hossza befolyással van-e a feliratkozók számára.

A biográfia hosszának összefüggése a feliratkozók számával



7.3-as ábra: A biográfia hosszának összefüggése a feliratkozók számával

A fenti grafikonon a függőleges tengely a feliratkozók számát jelzi, a vízszintes pedig a biográfia hosszát karakterekben. A piros vonal pedig egy trendvonal, ami majdhogynem lapos, ezzel indikálva, hogy nincs jelentős összefüggés a két érték között, azaz **a biográfia hossza látszólag nincs semmilyen befolyással a tartalmak sikerességére.**

Ennek az az oka, hogy a rövid formátumú videókat a felhasználók nagy része nem az egyes csatornákon keresztül éri el (ahol van esélyük látni a biográfiát), hanem a platformok külön erre kialakított szekcióin keresztül. A TikTok például az applikáció megnyitáskor egyből a „For You Page”-re visz, ahol a videók egy algoritmus által összeválogatott sorrendben jelennek meg. A többi platform is rendelkezik egy ugyanilyen szekcióval. Insragram-on „Reels Tab”-nak, Facebook-on „Reels Section”-nek és YouTube-on „Shorts Shelf”-nek hívják.

7.1.6. A „hangok” és zenék használata a videóknban

A az emberek koncentrációs képességének gyengülése jelentősen befolyásolta az internetes tartalmak gyártásának és fogyasztásának módját. A rövid formátumú videók népszerűségének növekedésével a tartalomgyártóknak alkalmazkodniuk kellett, hogy fenntartsák a nézők figyelmét. Mivel az emberek koncentrációs képessége egyre romlik, sok felhasználó azonnal továbbteker a következő videóra, ha az előző nem ragadja meg azonnal az érdeklődését.

Ennek orvoslására a tartalomgyártók egyszerre több figyelemfelkeltő és megtartó stratégiát alkalmaznak a videóikban. Ezek lehetnek vizuális-, verbális- és hang- „stratégiák”. A zenék használata is ide tartozik, mivel a széles körben ismert zeneszámok érzelmi reakciókat válthatnak ki, és ismerős érzéseket kelthetnek.

Az összes rövid formátumú platformon feltöltéskor a saját hangsáv mellett hozzá lehet adni egy meglévő „hangot” is, ami lehet egy zeneszám néhány másodperces részlete, vagy akármilyen más hangfájl. Ezek a hangok közül mindig van néhány, ami nagyon felkapott (viral), azaz több millió videóban szerepel, ezért sokan egyből felismerik.

Ebben az elemzésben annak jártam utána, hogy a hanggal ellátott videók több megtekintést szereznek-e, mint a hang nélküliek (tehát ahol nem használták a Shorts „Add Sound” funkcióját). Az én 20 ezer videós adathalmazomból mindössze a videók 20%-ánál használtak „hangokat”. A hangokkal ellátott videók átlagnézettsége 2,31 millió megtekintés volt, a hangok nélküli videóké pedig 2,41 millió megtekintés. Ezek majdhogynem egyező értékek, tehát első látásra úgy tűnhet, hogy nincs különbség a két stratégia között.

Itt viszont fontos megjegyezni, hogy attól hogy egy videónál nem használták az „Add Sound” funkciót, attól még rengeteg videónál előfordulhat, hogy az eredeti (feltöltő által szerkesztett) hangsávon szerepelnek különböző zenék vagy felkapott hangok, tehát a zenék használata a rövid formátumú videóknban minden bizonnyal gyakoribb, mint az én adathalmazomban fellelhető közel 50%-os arány. Ez azt jelenti, hogy sok olyan videó, aminek a megtekintései a korábbi elemzésemben a „hang nélküli” kategóriába estek, igazából a másik kategóriába tartoznak.

Ebben az esetben a zenéket és egyéb hangokat kihasználó videók átlagnézettsége, lehet hogy csak kis mértékben, de magasabb, mint az ezek nélkül készült videóké.

7.1.7. A hashtagek definíciója és célja

[21] A hashtagek olyan kulcsszavak, amelyek a „#” karakter után szerepelnek az internetes tartalmakban. A céljuk az említett tartalmak kategorizálása és követése a szociális platformokon. A hashtagek megjelenhetnek bejegyzések címében, leírásában, biográfiákban, vagy kommentekben.

A hashtagek összecsoportosítják a hasonló bejegyzéseket és beszélgetéseket, ezzel elősegítve azt, hogy az azonos érdekeltségű emberek könnyebben megtalálják a számukra érdekes tartalmakat. Miután egy hashtag létrejön, az adott platform felhasználói rá tudnak szűrni az általuk keresett hashtaget tartalmazó bejegyzésekre.

7.1.8. A hashtagek használatának előnyei

A tartalomgyártók és cégek szempontjából azért előnyös a hashtagek használata, mert a használatukkal képesek elérni bármilyen specifikus célközönséget.

Például ha egy technológiai érdekeltségű YouTuber készít egy kiértékelő videót az új iPhone 16-ról, akkor érdemes a videó címébe vagy leírásába elhelyeznie egy #iPhone16 hashtaget, hiszen így a YouTube algoritmus össze tudja csoportosítani a videót a többi iPhone 16-al kapcsolatos videóval, illetve könnyebben el tudja juttatni az iPhone felhasználókhoz a tartalmat.

7.1.9. A hashtagek használatának esetleges hátrányai

[22] A Sproutsocial 2024-es elemzése szerint a hashtagek akkor effektívek és előnyösek, ha megfelelő mennyiségben, és rendeltetésszerűen használják őket. Ha egy bejegyzés túl sok irreleváns, vagy túl általános hashtaget tartalmaz, akkor negatívan befolyásolja a bejegyzés elérését. A bekezdés elején megjelölt elemzés szerint platformtól függően minimum 1-2, és maximum 5 releváns hashtaget érdemes alkalmazni.

7.1.10. A hashtagek hatása a megtekintésekre és like-okra

Az adathalmazomban összehasonlítottam a hashtag nélküli és a hashtagekkel ellátott videók átlagnézettségét és átlag like számát. A hashtag nélküli videók átlagnézettsége 2,6 millió megtekintés volt, míg a hashtages videóké csupán 2 millió. A hashtag nélküli videók átlagos like száma 125 ezer volt, a hashtageseké pedig csak 113 ezer. A hashtages videók korlátozott sikere több okból kifolyólag történhet. Először is, az adathalmazomban rengeteg olyan videó volt, ahol az alkotó sokkal több mint 5 hashtaget alkalmazott (a legtöbb alkalmazott hashtag 11 volt). Ez az előző alfejezetben részletezett okok miatt negatívan hat a megtekintésekre.

Végeztem egy módosított lekérdezést is, ahol az 5-nél több hashtaggel ellátott videókat nem számoltam semelyik kategóriába, így csak a „rendeltetésszerűen hashtagezett” videók adatai

kerültek a hashtages adathalmazba. Nem meglepő módon, a hashtages videók átlagnézetsége majdnem 2 millióval, az átlagos like-ok száma pedig 9 ezerrel nőtt.

Ebből azt a következtetést tudtam levonni, hogy nem érdemes több mint 5 hashtaget alkalmazni a címekben.

7.1.11. Az naprakészség és a trendek követésének jelentősége

Minden szegmensben vannak kritikus időszakok, amikor bizonyos tartalmak feltöltése nagyobb nézettséget fog jelenteni, mint más időszakokban. Ilyen például a tech szférában az új készülékek bejelentései és kiadásai, például a legújabb iPhone, Intel/AMD/Apple processzor bejelentése. Ezekben az időszakokban megemelkedik a fogyasztói érdekltség, és az ilyen híreket feldolgozó tartalmak nagyobb nézettségre tesznek szert.

Ugyanez a jelenség más tartalomtípusokban is fellelhető. Minden nagyobb ünnephez kapcsolódik egy jellemző étel vagy ital, és az ezek elkészítésével kapcsolatos videók nézettsége ilyenkor általában megnövekedik. Azt vettem észre, hogy az olyan ünnepek előtt (és ideje alatt) mint a Húsvét, Hálaadás, Karácsony, a „Cooking” tartalomtípussal megjelölt videók nézettsége minden évben jelentősen megemelkedett.

Ebből azt a következtetést lehet levonni, hogy érdemes követni az adott szegmensben belüli fontosabb időszakokat, híreket és ilyenkor releváns tartalmakat feltölteni. Tehát fontos a naprakészség.

7.1.12. A pandémia hatása YouTube Shorts videók nézettségére

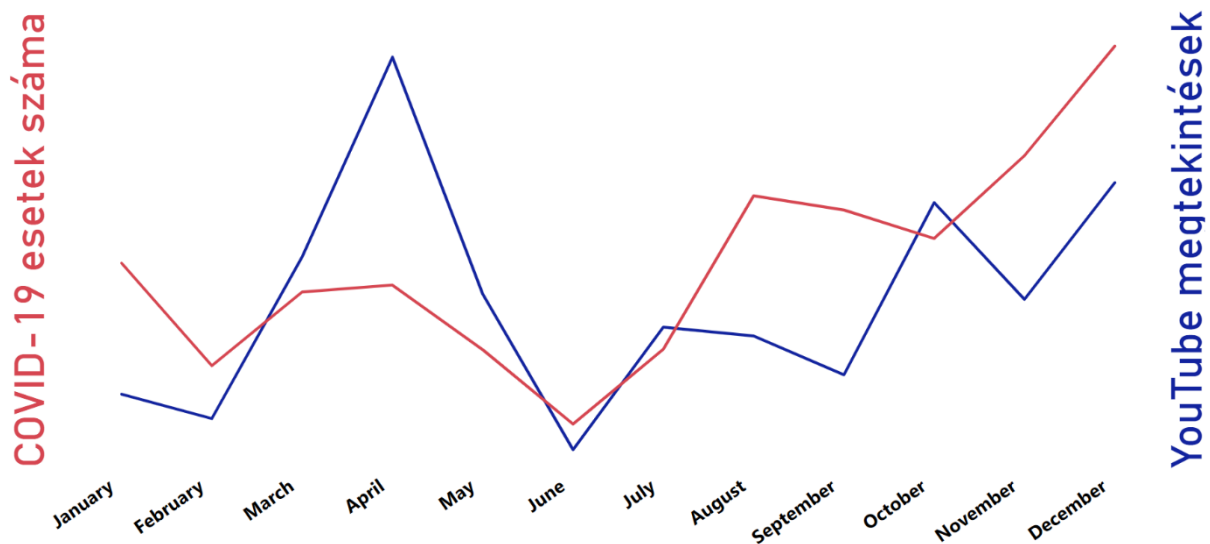
[23] A 2020-ban kialakuló világjárvány következtében a telefonok és számítógépek használata világszerte jelentősen megnőtt, és ezzel együtt a szociális platformok forgalma is.

[24] Ennek az illusztrálására letöltöttem egy 2021-es COVID-19 fertőzéseket dokumentáló táblázatot, amelyet összevetettem a 2021-es YouTube megtekintésekkel, hónapokra bontva.

A 2020 márciusában kezdődő karantén időszakról nem tudtam adatokat gyűjteni YouTube-on, mivel a YouTube Shorts mint videó formátum, csak 2020 szeptemberében került kiadásra (Ekkor tiltották be a TikTokot Indiában). A 2020-as adatok hiánya viszont nem okozott gondot, mivel a 2021-es adatok is hasonló tendenciát mutattak:

A YouTube megtekintések és COVID esetek számának összefüggése 2021-ben

● YouTube megtekintések 2021-ben ● COVID-19 esetek száma 2021-ben



7.4-es ábra: A YouTube megtekintések és COVID esetek számának összefüggése 2021-ben

A vonaldiagramról leolvasható, hogy a megtekintéseinek száma nagy vonalakban követi a COVID fertőzések számát. Ez azért volt, mivel ahogy a fertőzések emelkedtek, a világ legtöbb országában karanténokat rendeltek el, és ennek következtében sokan otthon voltak. Ez több szabadidőt eredményezett és ennek a szabadidőnek jelentős részét sokan YouTube videók megtekintésével töltötték.

Ezzel csak azt szerettem volna illusztrálni, hogy sokszor teljesen váratlan és első látásra irreleváns tényezők is befolyásolhatják az internetes tartalmak nézettségét és sikerességét.

8. ÉRTÉKELÉS

A végleges rendszerem nagyrészt lefedte a tervezett rendszert, viszont az eredetileg eltervezett és a végleges rendszer között sok módosítás történt. Voltak olyan részei a fejlesztésnek amik a tervezettnél könnyebben és gyorsabban mentek, viszont voltak olyan részek is, ahol elakadtam, és kerülő megoldásokat kellett alkalmaznom.

8.1.1. A szakdolgozat eredményeinek kiértékelése

Teljesített elvárások:

- Az irodalomkutatás során sikeresen utána jártam a hasonló rendszereknek, és kapcsolódó témáknak, majd pontosítottam a saját munkámmal szemben felállított elvárásokat
- Sikeresen begyűjtöttem nagymennyiségű analitikai adatot szociális platformokról, automatizáltan
- Sikerült az adatokat riportkész állapotba hozni (szintén automatizáltan)
- A begyűjtött adatokból sikerült új információkat kinyerni
- Az elkészült rendszert teszteltem
- A fejlesztési és egyéb folyamatokat az előírások szerint dokumentáltam

Területek ahol jobban teljesítettem az elvártnál:

- Egyszeri adattisztítás helyett kifejlesztettem egy skálázható, moduláris programot, amely képes egy adatbázis tábláit különféle módokon transzformálni, és nagyobb adatmennyiségeket kezelni

Területek ahol alulteljesítettem:

- Az elemzés és vizualizáció részeket nem sikerült automatizálnom, ezért ezeket manuálisan végeztem

8.1.2. Egyéb akadályok, melyek kerülőmegoldást igényeltek

A Web Scraper használatának hátulütője az volt, hogy nem minden adatot tudott begyűjteni a szociális platformokról. Ilyen hiányosság volt például a videók időtartamának begyűjtése. Ez az adat sok jelentős elemzés alapját képezhette volna, amiket emiatt nem tudtam elvégezni.

Az adatgyűjtés során a tervezett négy szociális platform helyett csak egyről tudtam használható adatokat begyűjteni, ez lehetetlenné tette a platformokat egymással összehasonlító elemzéseket.

Volt néhány olyan elemzés is, amiket végül nem tudtam belerakni a szakdolgozatba, mivel nem volt elegendő adatom, hogy megalapozott következtetéseket tudjak levonni.

9. TOVÁBBFEJLESZTÉSI LEHETŐSÉGEK

9.1.1. Az adathalmaz növelése

A szakdolgozat keretein belül „csak” 20 ezer videó adatát sikerült begyűjtenem, ami a big data világában nem számít hatalmas mennyiségnek. Az elemzések pontosításának és kiterjesztésének érdekében az egyik legegyszerűbb továbbfejlesztési lehetőség több adat beszerzése lenne.

9.1.2. Az elemzés és vizualizáció automatizálása

A jelenlegi rendszeremben csak az adatgyűjtés és az adattisztítás lépéseket sikerült automatizálnom. Az elemzés és vizualizáció lépések automatizálása lenne a legkézenfekvőbb továbbfejlesztési lehetőség. A programnak előre meg lehetne adni az elemzési szempontokat, majd az adott szemponthoz tartozó vizualizációs típusokat, így képes lenne magától következtetéseket levonni és tanácsokat adni.

9.1.3. A videó tartalmát elemző rendszer

A szakdolgozatomban szereplő elemzések csak a videók mutatószámait használták fel. Sokkal mélyebb és összetettebb elemzéseket tudnék végezni, ha a rendszerem képes lenne a videók konkrét tartalmát is elemzi. Ilyen lenne például a videókban megjelenő színárnyalatok elemzése, a videó hangereje, a tartalomgyártók beszédstílusa. Egy ilyen rendszer valószínűleg mesterséges intelligencián alapulna, amit több tízezer videón kellene betanítani, hogy képes legyen jelentős összefüggések felismerésére.

9.1.4. Különböző szociális platformokat összehasonlító rendszer

Mint ahogy azt korábban részleteztem, a tervezett négy szociális platform helyett csak egyről (YouTube) tudtam begyűjteni használható adatokat. Ha a Web Scraper és az Octoparse helyett például az egyes platformok API-ján keresztül próbálkoznék az adatok begyűjtésével, lehet hogy több sikerrel járnék. Több szociális platform adatainak összehasonlítása új elemzési lehetőségeket nyitna meg.

9.1.5. Tartalomgyártókat egymással összehasonlító rendszer

Az előző rendszerek kombinálva létre lehetne hozni egy olyan programot vagy applikációt, ami képes két tartalomgyártót egymással összehasonlítani különböző szempontok alapján. Ezután pedig képes lenne kiemelni a két influenszer közötti különbségeket, ezzel útmutatást adni a kevésbé sikeresnek, hogy milyen módon tudná javítani a nézettségét.

10. ÖSSZEGZÉS

Ebben a fejezetben a saját szavaimmal szeretném összefoglalni az elemzések eredményét, meglátásaimat és következtetéseimet. A szakdolgozatot azzal a céllal kezdtem el írni hogy kiderítsem mitől lesz sikeres egy rövid formátumú videó. Mik azok a dolgok amikre tartalomgyártóként oda kell figyelni annak érdekében, hogy nagyobb elérése és sikere legyen a videóknak.

A kérdéseim megválaszolására felhasználtam mind az egyetemi képzésem során megszerzett tudásomat, mind a szabadidőmben megszerzett háttértudásomat a szociális platformok terén.

Az rövid formátumú videók elemzése során az alábbiakra jöttem rá:

A szociális platformokon való siker elérésének legfontosabb eleme a konzisztencia, azaz minél gyakrabban tölt fel az ember (jó minőségű) tartalmakat, annál gyorsabban tud felépíteni egy követőbázist. Ez minden ezután következő „tanácsnál” és alapelnél fontosabb.

A videók címének hosszát érdemes 40 karakter alatt tartani, mivel a legtöbb mobiltelefon ennél többet nem képes levágás nélkül megjeleníteni. A másik érv a rövid címek mellett pedig az, hogy ha a cím a videó egész tartalmát lefedi, akkor a felhasználónak nincs indítéka rányomni a videóra, mivel úgy érzi hogy „már megkapta amiért jött”.

Ehhez szorosan kapcsolódik a következő meglátásom, miszerint érdemes kérdéseket, felkiáltójeleket és emoji-kat rakni a videók címébe, mivel ezekkel az eszközökkel sokkal effektívebben fel lehet kelteni az emberek érdeklődését. A kérdésekkel például be lehet vonni a nézőket a beszélgetésbe. Ez két okból nagyon hatékony. Először is, a videó így több interakciót fog generálni, ami az adott platform algoritmusának szemszögéből azt jelenti, hogy a videó sikeres, ezért egy egyre szélesebb közönség elé fogja kiterjeszteni a videó elérését. A másik haszna a kérdésfeltevéseknek és a követőkkel való rendszeres interakciónak a közösség építés. A videó megtekintése egy egyirányú interakció, amire a néző néhány órán belül már valószínűleg nem is fog emlékezni. Ezzel ellentétben ha egy beszélgetés keretein belül egy kétirányú interakció alakul ki, ez az élmény könnyebben megmarad a néző emlékezetében.

A figyelem felkeltésének és megtartásának fontos eszközei még a hangok és a vizuális elemek. Vegyünk például egy néhány perces videóesszét, amiből készül két változat. Az elsőben csak felolvasásra kerül a videóesszé tartalma, néhány kép kíséretében. A második változat pedig a képek mellett releváns videó szegmenseket is tartalmaz, és egész végig feliratozva van egy könnyen olvasható betűtípussal, amely egy élénk árnyalatra van színezve. A vizuális elemek mellett a háttérben halkán játszódik egy szöveg nélküli zeneszám és időnként lejátszódik egy hangfekktus, ami a videóesszé adott szegmensének a végét jelzi. A második változat egyértelműen hatékonyabb abban, hogy lekösse az embert. A videók szerkesztési stílusában valami ilyesmire érdemes törekedni.

A hashtagek rendeltetésszerű használata szintén fontos az internetes tartalmakban. Az irodalomkutatás és elemzés után tisztává vált, hogy a hashtageket akkor érdemes használni, ha az ember egy adott célközönséget akar elérni a tartalmaival. A hashtagek nem varázsszavak, amik minden esetben megemelik a videó nézettségét. A hashtageknek relevánsnak kell lenniük a videó tartalmához, és maximum 5 darabot érdemes használni belőlük.

Olyan tényezőkkel is találkoztam, amiknek látszólag semmilyen hatása nem volt a tartalmak sikerességére. Ilyen volt például a YouTube csatornák biográfiája. Semmilyen korrelációt nem fedeztem fel a biográfiák hossza és az adott alkotó videóinak sikeressége között. Ugyanilyen irreleváns a nézettség szempontjából a videók leírásának hossza és tartalma is. Ez viszont nem azt jelenti, hogy ezek haszontalan funkciók. A leírások és biográfiák hasznos funkciója, hogy lehet linkeket elhelyezni bennük, ami megkönnyíti a tartalomgyártók forgalmának átirányítását más platformokra. Legyen ez egy webáruház, vagy egy másik szociális platformon levő profilja.

Fontos még a naprakészség, ami ebben az esetben azt jelenti, hogy fel kell ismerni az adott tartalom szegmensben (Pl. technológia, gaming, ételek) a legfontosabb időszakokat, illetve eseményeket. Egy videójátékok értékelésével foglalkozó csatorna például akkor tud hosszútávon sikeressé válni, ha a legtöbb új videójátékkal kapcsolatban készít videókat. A főzős csatornák is sok nézettségre tehetnek szert, ha a nagyobb ünnepek idején releváns recept videókat készítenek.

Az időhöz vagy eseményekhez kötött trendek mellett minden szegmens rendelkezik még régóta használt, bevált formátumokkal. A tartalomgyártás világában fontos az eredetiség, viszont emellett fontos felismerni az adott szegmensben belüli bevált formátumokat is. A tech videókon belül például népszerűek az unboxing videók (elektronikus készülékek kicsomagolása és beüzemelése), vagy a különböző eszközök egymással való összehasonlítása (comparative tech review). Ezeket a bevált formátumokat érdemes kihasználni ha az ember egy specifikus szegmens célközönségét akarja megcélozni.

Ugyan ezt az elemzésemmel adathiány okából nem tudtam alátámasztani, de egy olyan meglátásom is volt, hogy a különböző platformok felhasználói más preferenciával rendelkeznek a tartalmakkal kapcsolatban. Egyes tartalomtípusok egyszerűen sikeresebbek bizonyos platformokon. Ez részben amiatt van, mert más demográfiai csoportok használják az egyes platformokat. A TikTok felhasználóinak átlagos életkora például 13 és 24 év közé, a Facebook-é pedig 30 és 45 év közé esik. Az egyes platformok analitikai felületei részletes betekintést nyújtanak a közönség demográfiai adataiba. Érdemes ezeket is kihasználni és ezek alapján alakítani a tartalomstratégiát.

10.1. Záró gondolatok

A big data eszközeinek köszönhetően ebből a limitált adathalmazból is sok új információt tudtam kinyerni. Utána tudtam járni olyan feltételezéseimnek, amelyeket eddig nem tudtam bebizonyítani.

Mint ahogy azt az értékelés fejezetben leírtam, sok helyen eltértem az eredeti fejlesztési tervtől előreláthatatlan akadályok miatt, és mert más irányba akartam vinni a projektet. Ennek ellenére elégedett vagyok a munkámmal, és úgy érzem rengeteg új tudásra tettem szert és sokat fejlődtem szakmailag. A problémamegoldó képességemet is fejlesztettem a szakdolgozat írása közben, mivel számos kisebb-nagyobb döntést kellett meghoznom annak érdekében, hogy tovább tudjam fejleszteni a rendszeremet.

A modern világban szinte mindenről tudunk adatokat gyűjteni, legyen szó emberek viselkedéséről, gazdasági folyamatokról vagy technológiai trendekről. Az adatok gyűjtése és elemzése kulcsfontosságú ahhoz, hogy megértsük a körülöttünk lévő rendszereket, és hatékonyabbá, fenntarthatóbbá tegyük mindennapi életünket.

IRODALOMJEGYZÉK

- [1] U. Shreya, P. Vishal, B. Shivansh és P. K. Mahantesh, „Articulating the construction of a web scraper for massive data extraction,” Coimbatore, India, 22-24 February 2017. [Hozzáférés dátuma: 10 04 2024]
- [2] B. Zhao, „Web Scraping,” Springer International Publishing, Oregon State University, Corvallis, OR, 2017. [Hozzáférés dátuma: 10 04 2024]
- [3] Web Scraper, [Online]. Available: <https://webscraper.io/>. [Hozzáférés dátuma: 10 04 2024].
- [4] B. Pfalzgraf, „How to Use Selenium to Web-Scrape with Example,” 03 04 2020. [Online]. Available: <https://towardsdatascience.com/how-to-use-selenium-to-web-scrape-with-example-80f9b23a843a>. [Hozzáférés dátuma: 10 04 2024].
- [5] Scrapy, „Scrapy főoldala,” [Online]. Available: <https://scrapy.org/>. [Hozzáférés dátuma: 10 04 2024].
- [6] B. Soup, „Beautiful Soup 4.12.0 Documentation,” [Online]. Available: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>. [Hozzáférés dátuma: 10 04 2024].
- [7] „11 Ways to Use Instagram's API for Your Business,” 07 08 2023. [Online]. Available: <https://blog.hubspot.com/website/instagram-api-examples>. [Hozzáférés dátuma: 10 04 2024].
- [8] TikTok, „TikTok for Developers terms and conditions,” 2024. [Online]. Available: <https://developers.tiktok.com/products/research-api/>. [Hozzáférés dátuma: 10 04 2024].
- [9] IBM, „What is ETL (extract, transform, load)?,” 2024. [Online]. Available: <https://www.ibm.com/topics/etl>. [Hozzáférés dátuma: 10 04 2024].
- [10] S. D’Aurizio, „Are Attention Spans Actually Decreasing?,” [Online]. Available: <https://brainmindsociety.org/posts/are-attention-spans-actually-decreasing>. [Hozzáférés dátuma: 21 04 2024].
- [11] T. f. Salesforce, „What Is Data Visualization? Definition, Examples, And Learning Resources,” Salesforce, [Online]. Available: <https://www.tableau.com/learn/articles/data-visualization>. [Hozzáférés dátuma: 01 05 2024].
- [12] S. R. Midway, „Principles of Effective Data Visualization,” CellPress open access, [Online]. Available: [https://www.cell.com/patterns/pdf/S2666-3899\(20\)30189-6.pdf](https://www.cell.com/patterns/pdf/S2666-3899(20)30189-6.pdf). [Hozzáférés dátuma: 01 05 2024].
- [13] A. C. Telea, Data Visualization principles and practice - second edition, CRC Press, 2007. [Hozzáférés dátuma: 10 04 2024]

- [14] C. Dilmegani, „Is Web Scraping Legal? Ethical Web Scraping Guide,” AIMultiple, 05 01 2024. [Online]. Available: <https://research.aimultiple.com/web-scraping-ethics/>. [Hozzáférés dátuma: 04 12 2024].
- [15] DataDome, „Is Web Scraping Illegal?,” DataDome, 18 06 2024. [Online]. Available: <https://datadome.co/guides/scraping/is-it-legal/>. [Hozzáférés dátuma: 04 12 2024].
- [16] F. L. András, „fuzerlevente-followeranalysis,” GitHub, 12 2024. [Online]. Available: <https://github.com/szakdolgozok-siposm/fuzerlevente-followeranalysis>. [Hozzáférés dátuma: 12 2024].
- [17] Microsoft, „Check Database Integrity Task,” Microsoft, 03 01 2023. [Online]. Available: <https://learn.microsoft.com/en-us/sql/integration-services/control-flow/check-database-integrity-task?view=sql-server-ver16>. [Hozzáférés dátuma: 30 11 2024].
- [18] StoryLab.ai, „YouTube Video Shorts Statistics You Need to Know in 2024,” StoryLab.ai, 2024. [Online]. Available: <https://storylab.ai/youtube-video-shorts-statistics/>. [Hozzáférés dátuma: 02 12 2024].
- [19] N. Gyant, „5 Best Tips for Writing YouTube Titles That Get Views (with Examples!),” LocaliQ, 03 04 2023. [Online]. Available: <https://localiq.com/blog/youtube-titles/>. [Hozzáférés dátuma: 02 12 2024].
- [20] H. V, „7 Awesome Tips to Write YouTube Titles That Generate Views,” Medium, 25 07 2022. [Online]. Available: <https://medium.com/illuminations-mirror/youtube-titles-e4f92125ded6>. [Hozzáférés dátuma: 02 12 2024].
- [21] SproutSocial, „Hashtags: What they are and how to use them effectively,” SproutSocial, 16 03 2023. [Online]. Available: https://sproutsocial-com.translate.goog/insights/what-is-hashtagging/?_x_tr_sl=en&_x_tr_tl=hu&_x_tr_hl=hu&_x_tr_pto=sc. [Hozzáférés dátuma: 12 11 2024].
- [22] J. Zote, „How hashtags on Facebook still work for businesses in 2024,” SproutSocial, 23 01 2024. [Online]. Available: https://sproutsocial-com.translate.goog/insights/hashtags-on-facebook/?_x_tr_sl=en&_x_tr_tl=hu&_x_tr_hl=hu&_x_tr_pto=sc. [Hozzáférés dátuma: 23 01 2024].
- [23] S. T. M. M. H. S. B. N. P. Abida Sultana, „Digital screen time during the COVID-19 pandemic: a public health concern,” F1000Research, Bangladesh, 2021. [Hozzáférés dátuma: 10 04 2024]
- [24] H. R. L. R.-G. C. A. D. G. C. G. J. H. B. M. S. D. D. B. E. O.-O. M. R. Edouard Mathieu, „Coronavirus (COVID-19) Cases,” Our World in Data, 2024. [Online]. Available: <https://ourworldindata.org/covid-cases>. [Hozzáférés dátuma: 15 11 2024].
- [25] D. glossary, „What is a star schema?,” 2024. [Online]. Available: <https://www.databricks.com/glossary/star-schema>. [Hozzáférés dátuma: 10 04 2024].
- [26] Z. Milek, „www.zebrabi.com,” zebrabi, 05 10 2024. [Online]. Available: <https://zebrabi.com/power-bi-dashboard-design/>. [Hozzáférés dátuma: 08 12 2024].

- [27] Reddit, „Awful retention rate resulting in low views, any tips for how to improve this metric? Or is 21% normal?,” 2022. [Online]. Available: https://www.reddit.com/r/Tiktokhelp/comments/11wajp1/awful_retention_rate_resulting_in_low_views_any/. [Hozzáférés dátuma: 02 10 2024].
- [28] C. Mahler, „7 Best Practices for Data Visualization,” TheNewStack, 30 01 2023. [Online]. Available: <https://thenewstack.io/7-best-practices-for-data-visualization/>. [Hozzáférés dátuma: 08 12 2024].
- [29] T. W. Channel, „The Weather Channel,” The Weather Channel, 01 05 2024. [Online]. Available: <https://weather.com/maps/ustemperaturemap>. [Hozzáférés dátuma: 01 05 2024].

ÁBRAJEGYZÉK

2.1-es ábra: A Web Scraper fontosabb funkciói.....	15
2.2-es ábra: Az Instagram Analytics Tool vizualizációs felülete.....	17
2.3-as ábra: Az Instagram Analytics Tool „Best Time to Post” funkciója.....	18
2.4-es ábra: PowerBI dashboard minta.....	22
2.5-ös ábra: A TikTok beépített analitikai felületének „Viewer retention rate” grafikonja.....	27
2.6-os ábra: Példa a különböző adatvizualizációs módszerekre.....	29
2.7-es ábra: A The Weather Channel weboldal hőmérsékleti előrejelzése.....	31
4.1-es ábra: Rendszertev ábra.....	34
4.2-es ábra : A TikTok-ról és Instagram-ról gyűjtött adatok táblarajza	36
4.3-as ábra : A YouTube-ról és Facebook-ról gyűjtött adatok táblarajza.....	37
4.4-es ábra : A rendszer funkcióinak ábrázolása folyamatábrával.....	38
4.5-ös ábra: Tervezett haladás Gantt diagrammon ábrázolva	39
5.1-es ábra: A transzformáló program felépítése a Solution explorerben.....	43
5.2-es ábra: A táblákat transzformáló függvények meghívása	43
5.3-as ábra: A yt_content tábla tartalma transzformáció előtt.....	45
5.4-es ábra: A yt_content tábla tartalma transzformáció után	45
6.1-es ábra: Az adatbázis integritásának ellenőrzése.....	46
6.2-es ábra: A tesztek eredménye a Test Explorer-ben.....	50
7.1-es ábra: A megtekintések összefüggése a feltöltési gyakorisággal.....	52
7.2-es ábra: A mondatvégi írásjeleket tartalmazó videók elemzése	54
7.3-as ábra: A biográfia hosszának összefüggése a feliratkozók számával.....	56
7.4-es ábra: A YouTube megtekintések és COVID esetek számának összefüggése 2021-ben	60

TÁBLAJEGYZÉK

<i>2.1.-es táblázat: Az adatgyűjtő rendszerek összehasonlítása.....</i>	<i>19</i>
<i>2.2.-es táblázat: Az adatvizualizáló rendszerek összehasonlítása.....</i>	<i>23</i>
<i>6.1.-es táblázat: A programon végzett tesztek eredményei.....</i>	<i>49</i>