

Budapesti Műszaki és Gazdaságtudományi Egyetem

Villamosmérnöki és Informatikai Kar

Távközlési és Médiainformatikai Tanszék (TMIT)

László Gábor

**Közösségi média aktivitásfigyelő szolgáltatás**

Konzulens

Nagy-Rácz István

BUDAPEST, 2018

Hallgatói nyilatkozat

Alulírott **László Gábor**, szigorló hallgató kijelentem, hogy ezt a diplomatervet meg nem engedett segítség nélkül, saját magam készítettem, csak a megadott forrásokat (szakirodalom, eszközök stb.) használtam fel. Minden olyan részt, melyet szó szerint, vagy azonos értelemben, de átfogalmazva más forrásból átvettem, egyértelműen, a forrás megadásával megjelöltem.

Hozzájárulok, hogy a jelen munkám alapadatait (szerző(k), cím, angol és magyar nyelvű tartalmi kivonat, készítés éve, konzulens(ek) neve) a BME VIK nyilvánosan hozzáférhető elektronikus formában, a munka teljes szövegét pedig az egyetem belső hálózatán keresztül (vagy hitelesített felhasználók számára) közzétegye. Kijelentem, hogy a benyújtott munka és annak elektronikus verziója megegyezik. Dékáni engedéllyel titkosított diplomatervek esetén a dolgozat szövege csak 3 év eltelte után válik hozzáférhetővé.

Kelt: Budapest, 2018. 05. 16.

...…………………………………………….

László Gábor

Tartalomjegyzék

[1 Absztrakt 6](#_Toc514243915)

[2 Abstract 8](#_Toc514243916)

[3 Bevezetés 9](#_Toc514243917)

[4 Adatgyűjtés 11](#_Toc514243918)

[4.1 A gyűjtőmodulok közös jellemzői 11](#_Toc514243919)

[4.1.1 A bemeneten kapott keresőkifejezésen elvégzett műveletek 12](#_Toc514243920)

[4.2 Facebook crawler modul 13](#_Toc514243921)

[4.2.1 Facebook API 14](#_Toc514243922)

[4.2.2 A keresés megvalósítása 14](#_Toc514243923)

[4.2.3 A Facebook-ról gyűjtött attribútumok 15](#_Toc514243924)

[4.2.4 Alternatív módszer a historikus adatok gyűjtésére 16](#_Toc514243925)

[4.2.5 A Cambridge Analytica-botrány hatása az adatgyűjtése 16](#_Toc514243926)

[4.2.6 A Facebook crawler értékelése 16](#_Toc514243927)

[4.3 Twitter crawler modul 17](#_Toc514243928)

[4.3.1 Twitter API 17](#_Toc514243929)

[4.3.2 A keresés megvalósítása 18](#_Toc514243930)

[4.3.3 A Twitter-ről gyűjtött attribútumok 18](#_Toc514243931)

[4.3.4 A Twitter crawler értékelése 20](#_Toc514243932)

[4.4 Instagram crawler modul 20](#_Toc514243933)

[4.4.1 Instagram API 20](#_Toc514243934)

[4.4.2 A keresés megvalósítása 21](#_Toc514243935)

[4.4.3 Az Instagram-ról gyűjtött attribútumok 21](#_Toc514243936)

[4.4.4 Az Instagram crawler értékelése 22](#_Toc514243937)

[4.5 Az adatok tárolása 23](#_Toc514243938)

[4.5.1 Az SQLite rendszer 23](#_Toc514243939)

[4.5.2 Statisztikák 23](#_Toc514243940)

[4.5.3 Megjelenítési módokhoz szükséges adathalmazok kialakítása 28](#_Toc514243941)

[4.5.4 Az adatgyűjtés ütemezése 29](#_Toc514243942)

[4.5.5 Az adatgyűjtés során felmerült nehézségek és megoldásaik 29](#_Toc514243943)

[4.6 Az adatgyűjtő modul értékelése 30](#_Toc514243944)

[5 A kapott adatok validálása, a keresett profilok pontos beazonosítása 31](#_Toc514243945)

[5.1 Ötletek, fejlesztési irányok 31](#_Toc514243946)

[5.1.1 Legpontosabb egyezés 31](#_Toc514243947)

[5.1.2 Legtöbb követővel rendelkező 31](#_Toc514243948)

[5.1.3 Cache-ben tárolt, kézzel validált halmaz alapján 32](#_Toc514243949)

[5.2 Validáláskor felmerült problémák 32](#_Toc514243950)

[5.3 A validálás értékelése 33](#_Toc514243951)

[6 Megjelenítés 34](#_Toc514243952)

[6.1 Igények és célcsoportok definiálása 34](#_Toc514243953)

[6.1.1 Milyen adatokból dolgozom? 35](#_Toc514243954)

[6.1.2 Kik lehetnek a felhasználók? 35](#_Toc514243955)

[6.1.3 Milyen méretet és formát válasszak? 36](#_Toc514243956)

[6.2 Megjelenítési tervek 36](#_Toc514243957)

[6.2.1 Általános áttekintő felület 37](#_Toc514243958)

[6.2.2 Összehasonlító felület 38](#_Toc514243959)

[6.2.3 Időtartományban kereső felület 39](#_Toc514243960)

[6.3 Power BI 40](#_Toc514243961)

[6.3.1 Power BI Desktop 40](#_Toc514243962)

[6.4 Felhasználható felületi elemek 41](#_Toc514243963)

[6.4.1 Diagramok 41](#_Toc514243964)

[6.4.2 Szűrők 41](#_Toc514243965)

[6.4.3 Kártyák, táblázatok 42](#_Toc514243966)

[6.4.4 Térképes megjelenítés 42](#_Toc514243967)

[6.4.5 Statikus elemek 42](#_Toc514243968)

[6.4.6 Számított oszlopok 43](#_Toc514243969)

[6.4.7 Egyéni vizualizáció importálása 43](#_Toc514243970)

[6.5 Az elkészült dashboard-ok 43](#_Toc514243971)

[6.5.1 Általános áttekintő felület 43](#_Toc514243972)

[6.5.2 Összehasonlító felület 45](#_Toc514243973)

[6.5.3 Időtartományban kereső felület 47](#_Toc514243974)

[6.6 A megjelenítés értékelése 49](#_Toc514243975)

[7 Továbbfejlesztési lehetőségek 51](#_Toc514243976)

[7.1 Több adatforrás bevonása 51](#_Toc514243977)

[7.2 Több attribútumtípus gyűjtése, más elemzési szempontok 51](#_Toc514243978)

[7.3 Más megjelenítési formák és eszközök bevonása 52](#_Toc514243979)

[8 Összegzés 53](#_Toc514243980)

[9 Függelék 54](#_Toc514243981)

[9.1 A cache-ben tárolt csapatok és főbb mutatóik 54](#_Toc514243982)

[9.2 Az elkészült dashboard-ok 59](#_Toc514243983)

[9.2.1 Általános áttekintő felület 59](#_Toc514243984)

[9.2.2 Összehasonlító felület 60](#_Toc514243985)

[9.2.3 Időtartományban kereső felület 61](#_Toc514243986)

[Irodalomjegyzék 62](#_Toc514243987)

# Absztrakt

TODO aktualizálni

A XXI. században egyre nagyobb teret nyernek azok az elemzési megoldások, amelyek a publikusan elérhető közösségi oldalak adatait dolgozzák fel. Az elérhető információk birtokában képesek lehetünk jól leírni egy közösségi profil működését és megbecsülhetjük a benne rejlő potenciált is.

Diplomamunkám célja adatgyűjtés közösségi oldalakról abból a célból, hogy a gyűjtött adatokkal biztosítsam közösségi oldalakon lévő profilok aktivitásának elemzését. Az adatok gyűjtését és feldolgozását angol nyelvű, futballcsapatokhoz kapcsolható profilokra specializálom. A feladatom része az adatok legyűjtése, feldolgozása, valamint vizualizálása, illetve a lehetséges továbbfejlesztési irányok ismertetése.

Az adatgyűjtéshez három közösségi oldal jelentett forrást. Munkám során megismertem a Facebook, a Twitter és az Instagram közösségi oldalakat, azok felhasználói profiljainak felépítését, továbbá az adatok hozzáférésének módját. Megterveztem és implementáltam egy olyan szoftvert, amely képes erről a három oldalról adatokat gyűjteni és rendezetten tárolni. Ezután a gyűjtött adatok minőségi javulása érdekében adatelkészítő lépéseket hajtottam végre.

Vizualizálási megoldásokat dolgoztam ki annak érdekében, hogy a kapott adatok jól megjeleníthetők legyenek. Áttekintettem a megjelenítésben rejlő lehetőségeket.

Megjelenítési terveket dolgoztam ki, hogy a legyűjtött adatokat minél jobban értelmezhető formában tudjam ábrázolni. Megismertem a Power BI elemző eszközt, annak képességeit és dashboard-okat valósítottam meg a segítségével a megjelenítési tervek alapján.

Ezt követően valós adatokkal teszteltem a szoftverem profilokat felismerő moduljának hatékonyságát. Feltártam az esetleges gyenge pontokat és szűk keresztmetszeteket.

Végül áttekintettem az elkészített szoftver továbbfejlesztési lehetőségeit a gyűjtő modultól az adatelőkészítésen át a tényleges elemzésig. Javaslatot tettem az így feltárt hiányosságok és a prototípus helyenkénti egyszerűségéből fakadó pontatlanságok kijavítására, valamint a hatékonyság növelésére.

# Abstract

TODO AKTUALIZÁLNI!!!

In the 21st century, those data mining solutions which are based on getting information from social networks, have gained more ground recently. With the information available, we can describe the operation of a community profile and estimate its potential.

The goal of my thesis is gathering data from social networks in order to provide the collected data to analyze the activity of profiles on social networking sites.

I specialize in collecting and processing data in English-language profiles linked to football teams. My task was to gather, and analyse and visualize information, and furthermore to describe potential development directions.

During the data mining progress, three social network sites were used. The working project was a good chance to get acquainted with Facebook, Twitter and Instagram and structure of their user’s profiles. I also managed to know the method of accessing information. A so-called data mining software was architected and implemented, which is able to gather and sort information from these three social network sites. Then, in order to improve the quality of gathered information, I made some data preparing.

I have developed visualization solutions so that the data can be displayed well. I reviewed the potential of the display.

Next, I tested the efficiency of my software with real information. Incidental weak points and cross-sections were located. I succeeded in finding the most suitable solution to finding the best match for the given query.

Finally, I reviewed the possible development directions of the finalized software, including the gathering module, data preparing and the actual analysis. I made a proposal for repairing the uncovered deficiencies, inaccuracies of the prototype and proposal for increasing effectiveness.

# Bevezetés

TODO aktualizálni

Egy magára valamit is adó futballklubnak nem elég csupán a pályán helytállnia. A marketing, az imázs éppen olyan fontos, mint a rúgott gólok száma, vagy a nézőszámok alakulása. Ahhoz, hogy egy csapat ne csak a tabellán foglaljon el jó helyezést, hanem a közösségi médiák népszerűségi listáin is, aktív részvétel kell az egyes platformokon.

Az elmúlt évtized a közösségi oldalak térnyeréséről szólt. Kezdetben a magánszemélyek kapcsolattartását segítették elő, később azonban a cégek is felismerték a benne rejlő lehetőségeket. Egy felhasználó átlagosan közel két órát tölt el valamelyik közösségi oldal böngészésével [1].

Nem kivételek ez alól a sportegyesületek sem, amelyek jelentős rajongói bázist tudtak kiépíteni ezeken az oldalakon, egyfajta mérhető népszerűségi versenyt generálva. Ezen adatok egy része, például a kedvelők vagy követők száma publikusan is elérhető, látható az egyszeri látogató számára. Vannak azonban olyan metrikák, amik nem ennyire triviálisak, illetve nem feltétlenül érhetők el akárkinek, ugyanakkor hasznosak lehetnek egy-egy klub közösségi aktivitásának leírásához. A meglévő beépített statisztikai mutatók mellett jó volna egy olyan lehetőség is, amelyik több közösségi portál adatait is képes összefésülni és ezzel átfogó képet adni az aktivitásról. Ezeken felül vizualizációval is lehetne támogatni az elemzéseket.

Feladatomban három fő közösségi oldalon, a Facebookon, a Twitter-en és az Instagramon vizsgálom a nagyobb európai futballcsapatok aktivitását. A feladatnak részét képezi a tényleges hivatalos oldalak automatikus felismerése és validálása, valamint a lehető legtöbb releváns jellemző paraméter legyűjtése. Az elkészült fő modul alkalmas lesz az egy vagy több kulcsszó szerinti keresésre, emellett pedig a keresés gyorsítása érdekében a nevesebb csapatok statikus attribútumait, például azonosító fixen tárolja. A naprakészen gyűjtött adatokból historikus adathalmaz állítható össze, ami elősegíti az elemzéseket, predikciókat, valamint a vizualizálást is. Az egyes modulok képesek önállóan is működni.

Az adatgyűjtéshez alaposan áttanulmányozom a Facebook, a Twitter és az Instagram közösségi oldalak szerkezetét, felhasználói profiljainak felépítését. Az egyes oldalak publikus API-jait felhasználva olyan szoftvert fejlesztek, amely képes egy névre keresve profiladatokat gyűjteni a három forrásból.

A találatok gyűjtésekor jól behatárolom a ténylegesen szükséges attribútumok és metaadatok halmazát és különböző adatelőkészítő lépésekkel hozom őket könnyen kezelhető formára. A gyűjtés során különösen nagy hangsúlyt kap a kedvelők és követők számának vizsgálata, emellett pedig más fontos attribútumok is legyűjtésre kerülnek. Ezután kidolgozok egy olyan eljárást, amellyel kiválaszhatók a pontos találatok, biztosítva, hogy ténylegesen a keresett csapat adatait kapjuk vissza.

Áttekintem a megjelenítési lehetőségeket, hogy a kapott adatokat érthető, elemzésre alkalmas formára tudjam hozni. Megjelenítési terveket dolgozok ki és implementálom is azokat. Megismerem a vizualizálásban rejlő lehetőségeket, az ahhoz felhasználható eszközöket és módszereket.

<Power BI>

Végül áttekintem a szoftver lehetséges továbbfejlesztési irányait és javaslatot teszek az egyes részek későbbi javítására.

# Adatgyűjtés

Az közösségi média aktivitás megfigyelését végző rendszer egyik legfontosabb része az adatok gyűjtéséért felelős részegység. Ennek feladata, hogy egy megadott inputra minél több potenciális kapcsolódó adatot gyűjtsön össze a megfelelő forrásokból.

Az interneten fellelhető közösségi oldalak kimeríthetetlen publikus adatforrást jelentenek. Az ilyen oldalak jellemzően elérhetőek a fejlesztők számára valamilyen API-n keresztül, így bányászatuk alapvetően nem ütközik különösebb nehézségbe. Az egyetlen valódi gondot az okozza, hogy az utóbbi időszakban különböző szigorítások léptek életbe az egyes oldalakon elérhető adatok láthatóságával kapcsolatban, így azok olykor csak korlátozottan érhetők el. Ilyen eset például a Cambridge Analytica-botrány miatti Facebook-ot érintő korlátozások, melyekre bővebben is kitérek a fejezetben.

A feladatomhoz három olyan közösségi oldalt választottam ki (Facebook, Twitter, Instagram), amelyek a regisztrált felhasználóik számában nagyjából azonos nagyságrendet képviselnek, stílusuk és funkciójuk viszont különböző. Fontos szempont volt még, hogy jelentős felhasználói bázissal rendelkezzenek, illetve figyelembe vettem azt is, hogy kellő mennyiségű gyűjthető adatot biztosítsanak.

A következő fejezetben ismertetem az adatgyűjtő modul működését, részletesen kitérve az egyes források sajátosságaira, az API-k használatára és az adatok tárolásának módjára.

## A gyűjtőmodulok közös jellemzői

Az általam elkészített gyűjtőmodulok moduláris szerkezetűek. Önállóan is alkalmasak az adott forrás tematikus gyűjtésére, együtt pedig megadják a kulcsszóhoz tartozó összes elért paramétert. Mindhárom modult Python nyelven implementáltam.

Minden blokk visszaadja az összes kiválasztott attribútumot egy kulcs-érték párokat tartalmazó adatszerkezetben. Modulonként átlagosan tíz jellemző paraméter érhető el, de ez a lista igény szerint könnyen tovább bővíthető. Az összesített szótár típusú eredményen felül minden attribútumra készítettem tagfüggvényeket is arra az esetre, ha csak egy-egy érték frissítésére van szükség, például a követők számának naprakész lekérdezésére. Ez a megfontolás a skálázhatóságot segíti elő.

A lekérdezések eredményeit lehetőség van egy általam definiált függvénnyel CSV formátumba is kimenteni a további elemzések és statisztikák készítésének elősegítésére. Ehhez segítségül hívtam Python Pandas [2] nevű, adatelemzést segítő modulját. Ezen felül létrehoztam egy adatbázis alapú tárolási megoldást is azokra a paraméterekre, amelyeket tartósan is érdemes tárolni. Ilyenek például a név, egyedi azonosító (ID), like-ok/követők száma, valamint egy weboldal linkje.

Vannak továbbá olyan értékek is, amik csak az adott forrást jellemezték. Ilyen például a fiók létrehozásának ideje, kategória, illetve a közzétett képek vagy bejegyzések száma, valamint néhány olyan historikus mérőszám, mint az utolsó tíz bejegyzés átlagos kedveltségi száma vagy átlagos „kommenteltsége”. A modulok attribútumaira a későbbiekben bővebben kitérek.

Az egyes adatgyűjtő modulokban közös még, hogy a szükséges kulcsokat és tokeneket külön config fájlokba rendezve adtam meg, így függetlenítve az alkalmazást a személyre szabott azonosítóktól való függéstől. Az egyes feladatrészeket az alábbiakban fejtem ki részletesen.

### A bemeneten kapott keresőkifejezésen elvégzett műveletek

Miután a cél egy olyan komplex gyűjtőmodul elkészítése, amely összefogja mindhárom adatforrást, fontos a bemenetet egységes formára hozni, amit mindegyik API fel tud dolgozni. Erre több lépést is implementáltam.

Az egyes adatforrásoknál a keresés során már egy karakter eltérés is különbségként jelenik meg. Különösen zavaró, ha az eltérés csupán abból fakad, hogy az egyik helyen ugyanaz nagy kezdőbetűvel szerepel, míg a másikon kicsivel. Ezt elkerülendő, az összes modulnál a bemeneten kapott keresőkifejezést kisbetűs formára hoztam.

Egyes csapatok nevei tartalmazhatnak speciális karaktereket, amelyeket nem minden API tud feldolgozni. A hibaüzenetek elkerülése végett készítettem egy karaktereket vizsgáló és cserélő modult, amely minden gyűjtőmodulban felhasználásra került. A futballcsapatok nevében előforduló speciális karakterek vizsgálatához generáltam egy listát az európai első osztályú csapatok neveiből, és ebből a listából kivettem az angol ABC-nek megfelelőket. Ez alapján a hagyományosabb, ékezetes kis-és nagybetűk mellett az alábbi listában szereplő karaktereket cseréltem ki az alakban hozzájuk legközelebb álló valid, angol ABC-nek megfelelő karakterre:

|  |
| --- |
| ř, ş, â, ä , ø , ñ, å, æ, č, ğ, à, Ł, ł, ń, ž, Š, Ž, š, ć, î |

A több forrásból való gyűjtés miatt időnként belefutottam karakterkódolási anomáliákba. Ezt feloldandó a Django modul[3] smart\_str függvényét hívtam segítségül, amely hatékonyan megszünteti a lehetséges hibákat.

Mivel minden forrás tartalmazhat webes linket, ezért fontos ezek elérhetőségének és helyességének ellenőrzése. Ehhez egy néhány sorból álló Python-os megoldást választottam, amely ellenőrzi, hogy az adott link elérhető-e és onnan értelmes adat visszaérkezik-e. Ha igen, akkor tárolja a linket, ellenkező esetben nincs értelme a tárolásnak, értéket nem képvisel a cím.

A CSV formátum potenciális használata miatt indokolttá vált, hogy minden gyűjtött attribútumból eltávolítsam a vesszőket. A vesszők használata nem befolyásolja az adatok értelmezését, nincs megkülönböztető erejük. Helyettük pontosvesszőket szúrtam be.

## Facebook crawler modul

A Facebook[[1]](#footnote-1) korunk egyik legelterjedtebb közösségi oldala. Csaknem minden negyedik ember használja világszerte, míg a hazai viszonyra igaz, hogy már jóval ötmillió fölött van a Facebook felhasználók száma. A későbbiekben tárgyalt két portállal ellentétben ez megkülönbözteti a természetes személyeket és az egyes publikus oldalakat. Előbbinek az automatizált vizsgálata meglehetősen komplikált a biztonsági beállítások miatt, míg utóbbinak az elemzésére számos jól konfigurálható lehetőség áll az informatikusok rendelkezésére. Az alábbiakban a Facebook crawler modulját ismeretem részletesen.

### Facebook API

A Facebook esetében a Graph Search [4] biztosít lehetőséget a keresésre, elemzésre. Használatához egy ’access token’-re van szükség, amit az admin felületen generálhatunk, ugyanakkor nincs szükség külön Python modul letöltésére és alkalmazására a Facebook eredményes bányászatához. Elég a generált tokent megadni minden lekérdezés végző URL végén.

Az így elérhető attribútumok bővebb listája elérhető a szolgáltatás online dokumentációjának oldalán[5]. Az alábbiakban csak az általam használt mérőszámokra térek ki részletesebben.

### A keresés megvalósítása

A keresés ebben a modulban három fő részből áll össze. A három lépésre azért van szükség, mert előbb meg kell tudni egy oldal ID értékét ahhoz, hogy a bővebb attribútumhalmaz hozzáférhetővég váljon.

Az első lépés a szabadszavas keresés, melynek bemenete egy csapat teljes neve. A keresőkifejezésben a szóközök használata megengedett. Ennek lényege, hogy a beírt kifejezésre visszaadja a keresés nyers eredményét egy listában (’id’, kedvelők száma). A megadott kereső kifejezést a modul mentesíti a szóközöktől. A speciális karakterek használata nem okoz problémát a kereséskor.

Szabadszavas keresésre az alábbi url mutat példát.

|  |
| --- |
| [https://graph.facebook.com/search?q="<team\_name>"&type=page&access\_token= <access\_token>](https://graph.facebook.com/search?q=%22%3cteam_name%3e%22&type=page&access_token=%20%20%3caccess_token%3e) |

A második lépés a validálás a kedvelések számának maximuma alapján. A validálás során a kedvelők számát tekintettem perdöntőnek, azaz a legnagyobb követői bázissal rendelkező oldalt választja ki az alkalmazás a keresési eredményeket tartalmazó listából. A végleges, cache-ből kinyert validált lista használatakor ez az első két lépés eldobásra került.

A validálás eredményeként kapott oldal részletes attribútum listájának lekérdezése ID alapján a harmadik lépésben történik meg. Egy ID alapú, érvényes Graph Search lekérdezésre az alábbi kifejezés ad példát. Az ID helyére a lekérdezni kívánt oldal azonosítója kerül, míg az ’access token’ értelemszerűen a fentebb említett token értékkel azonos.

|  |
| --- |
| [https://graph.facebook.com/v2.6/<id>?fields=id,name,likes,link,talking\_about\_count,category&access\_token=<access\_token>](https://graph.facebook.com/v2.6/%3cid%3e?fields=id,name,likes,link,talking_about_count,category&access_token=%3caccess_token%3e) |

Jóval kevesebb lekérdezést (és ezáltal időt) igényel, ha csak a validált fiókról kérjük le a részletes attribútum listát. Erre később a validálással foglalkozó fejezetben részben térek ki bővebben. Hibakezelésre akkor volt szükség a modulban, amikor a keresés első fázisa üres listát adott vissza. Ebben az esetben értelmezhetetlen a validálás maga, modul által reprezentált eredmény minden értéke a ’NaN’ (not a number) kifejezést kapta, jelezve, hogy a kért keresés nem járt sikerrel.

### A Facebook-ról gyűjtött attribútumok

A gyűjtött attribútumok az alábbiak:

* name  
  Az oldal neve
* id  
  Az oldal egyértelmű azonosítója
* likes  
  Az oldalra leadott kedvelések száma
* talking\_about\_count  
  Az oldal megemlítéseinek száma
* category  
  Az oldal kategória besorolása (pl. Sport Club)
* url  
  Az oldalhoz kapcsolódó hivatkozás
* rating\_count  
  Az oldalra leadott értékelések száma
* were\_here\_count  
  Az oldalhoz kapcsolódó lokációval kapcsolatos hivatkozások száma

### Alternatív módszer a historikus adatok gyűjtésére

A Facebook egyik lekérdezési típusa lehetőséget nyújt a követők számának alakulását visszamenőlegesen is megtekinteni. Erre egy három hónapos (3 x 31 = 93 napot lefedő), offsettel további idősávokra kiterjeszthető időablakot biztosít a múltban. Sajnos ez a lehetőség nem adaptálható tökéletesen a projektemben, ugyanis a tényleges követőszámot csak az általunk birtokolt oldal esetén van mód lekérdezni. Publikus oldalaknál egy trükk szükséges ahhoz, hogy a kért adathoz hozzájussunk. A lekérdezés ugyanis megadja a követők számát a múltban egy-egy adott napon, de mindezt országok szerinti bontásban teszi, így a kapott eredményt még össze kell fésülni. Az így kapott adat több visszamérés után sok helyen drasztikusan eltért a valóságtól, így a használatát végül elvetettem.

### A Cambridge Analytica-botrány hatása az adatgyűjtése

A gyanú szerint a Cambridge Analytica nevű brit tanácsadócég mintegy 87 millió ember Facebookon tárolt személyes adatait szerezte meg, és használta fel politikai kampányokhoz az érintettek tudta és hozzájárulása nélkül 2016-ban. Közülük 71 millióan amerikaiak voltak.

A cég Facebook-felhasználók adatainak feldolgozásával igyekezett képet alkotni a célba vett amerikai választók politikai beállítottságáról, és ennek alapján a 2016-os amerikai elnökválasztási kampányban személyre szabott üzenetekkel próbálta befolyásolni őket.

A botrány kirobbanását követő hónapokban, így a diplomatervem készítésének második felében a Facebook-ról történő adatgyűjtés folyamatosan akadályokba ütközött. Ezért nem sikerült összefüggő és releváns mennyiségű adathalmazt gyűjtenem a forrásból. Az elkészített crawler modul a szigorítások várható feloldása után vélhetően alkalmas lesz ismét az adatgyűjtésre.

### A Facebook crawler értékelése

A Facebook crawler a szigorítások és korlátozások előtti időszakban, 2018 márciusa előtt kielégítően működött. A választott klubok szinte kivétel nélkül megtalálhatók ezen a közösségi oldalon, bőséges adatforrást biztosítva az elemzéshez. A Cambridge Analytica-botrány elülése után várhatóan ismét a várt működést fogja biztosítani a modul, addig azonban átmenetileg felfüggesztettem a használatát. A korábban gyűjtött alapadatok szerepeltetése egy egyszerűbb dashboard-on megoldható, historikus elemzést viszont nem tesz lehetővé.

## Twitter crawler modul

A Twitter[[2]](#footnote-2) nevű közösségi oldal számos, közösségi médián alapuló szövegbányászati és kutatási feladat alapanyagául szolgál. Világszerte több mint félmilliárd regisztrált felhasználója van, melyek közül körülbelül 300 millió tekinthető aktívnak. Működésének alapjai a felhasználók által írt, maximum 280 karakter hosszúságú formázatlan üzenetek, az úgynevezett Tweet-ek. Ezek a bejegyzések 2017 szeptemberéig 140 karakterben voltak limitálva [5], így a megnövelés lehetőséget biztosít több hasznos információ közlésére, ami az egyes adatgyűjtéseknél is hasznos lehet.

Ugyan a profilok létrehozásakor nem szükséges sok személyes adatot megadni, mégis egy-egy precízen kitöltött adatlap megfelelő forrásként tud szolgálni az adott felhasználó vagy oldal rövid, felületes feltérképezéséhez. Emellett pedig a Tweet-ek hosszkorlátozása is segítségünkre lehet, ugyanis ezek a státuszüzenetek garantáltan csupán néhány szóból állnak, amelyek jó eséllyel a közlendő üzenet kulcsszavai. Az alábbiakban a Twitter crawler modulját ismeretem részletesen.

### Twitter API

A Twitter-en történő adatgyűjtéshez legelőször egy érvényes fiókra van szükség. Ezután nyílik lehetőség saját alkalmazást regisztrálni. A Twitter sok más, hasonló közösségi oldalhoz hasonlóan OAuth[7] azonosítást használ. Az OAuth egy nyílt azonosítási szabvány, amely tokenek segítségével biztosítja a biztonságos hozzáférést az egyes szolgáltatásokhoz. Az adatgyűjtőhöz a Python nyelvhez készített Twitter modult[8] használtam fel.

Az azonosításhoz szükséges kulcsok az alábbiak:

* consumer\_key
* consumer\_secret
* access\_token
* access\_token\_secret

Az első kettő állandó, a tokeneket pedig értelemszerűen időnként újra kell generálni. Mivel az összes felhasznált script esetén szükséges az azonosítás, ezért érdemes ezt a műveletet a külön fájlban (config) megtenni.

### A keresés megvalósítása

Összesen több mint ötvenféle kéréstípus és attribútum érhető el a fejlesztők számára, melyek között vannak keresésre, kapcsolati hálók felkutatására és önálló Tweet-ek kezelésére vonatkozó kérések egyaránt. Ezek közül én az User entitást [9] és a hozzá tartozó kéréseket választottam ki.

Amint azt a Twitter dokumentációja is írja, User bárki és bármi lehet az oldalon. Ennek nyomán az általam vizsgált futballklubok is User-ként jelennek meg a Twitter világában, tehát a hozzá kapcsolódó attribútumok is így érhetők el.

A Twitter fejlesztői dokumentációja részletezi, hogy melyik felhasználói adathoz milyen kérésekkel lehet hozzáférni. Első közelítésben egy adott kulcsszóhoz, kulcsszavakhoz tartozó profilokat kellett összegyűjtenem. Ehhez megvizsgáltam a lehetőségeket a Twitter dokumentációjában.

A számomra ideális kérés pont ezt a funkciót látja el, ugyanis egy vagy több kulcsszóra adja vissza a kapcsolódó profilokat. Ha több kulcsszót adunk meg, akkor azokat a profilokat adja vissza, amelyekben a keresett szavak mindegyike megtalálható.

A keresés során sem időarányos, sem darabszám szerinti limitet nem léptem át. Ez a dokumentáció szerint elvileg száz találatot jelentene. A korlátoknak főleg a Tweet-ek keresésénél van jelentősége, ám ez nem képezte részét a feladatomnak.

### A Twitter-ről gyűjtött attribútumok

A gyűjtött attribútumok az alábbiak:

* screen\_name  
  A fióknál megjelenő név
* location  
  A fiókhoz kapcsolódó földrajzi hely. Nem mindig egyértelmű, néha csak a város, néha konkrét cím.
* id  
  A fiók egyértelmű azonosítója
* followers\_count  
  A fiók követőinek száma
* friends\_count  
  A fiókhoz tartozó kapcsolati lista számossága
* favourites\_count  
  A fiókot kedvencként mentő felhasználók száma
* created\_at:  
  A fiók létrehozásának időbélyege
* statuses\_count  
  A fiók által publikált státuszok száma
* avg\_fav  
  Az utolsó tíz státusz átlagos kedvelői száma
* avg\_retweet  
  Az utolsó tíz státusz átlagos megosztási száma
* url  
  A fiókhoz tartozó hivatkozás
* hashtags  
  A fiók utolsó státuszaiban szereplő hashtag-ek gyűjteménye

Az ’expanded url’ meglétét külön kellene ellenőrizni, ugyanis nem kapcsolódik minden Tweet-hez, hiánya pedig hibaüzenetet generál. A sima url paraméter csak a Twitter által rövidített linkeket adja meg, ezért az url esetén elkészítettem egy segédfüggvényt, amely a Twitter által előírt, rövidített (t.co) linkeket[10] kibontja a teljes címmé.

A dátum formátumát szintén átalakítottam egy könnyebben kezelhető, kényelmesebb alakra. Az eredeti dátumformátumra az alábbi sor mutat egy példát.

|  |
| --- |
| Wed Okt 05 21:18:15 +0000 2016 |

Ebből az alábbi időbélyeget generáltam le.

|  |
| --- |
| 2016-10-05 |

A lekérdezések eredményét JSON formátumban kaptam meg, amely már nem tartalmazza az óra és perc adatait.

Miután a Twitter keresője követőszám szerinti csökkenő sorrendben adja vissza a találatokat, a Facebooknál alkalmazott validálási eljárás egyszerűbben alkalmazható, ugyanis elegendő a lista első elemét venni és a későbbiekben ezt kézzel ellenőrizni.

### A Twitter crawler értékelése

A tesztelés során hasonló problémákba és akadályokba futottam bele, mint az előbbiekben tárgyalt esetben. A lefedettség valamivel kisebb, mint a Facebook esetén, de még itt is 90% fölötti.

## Instagram crawler modul

Az Instagram[[3]](#footnote-3) egy fényképek és rövid videók megosztásán alapuló közösségi hálózat, mely kifejezetten okostelefonról történő használatra készült. A szolgáltatás iOS-ről, Androidról és Windows Phone-ról egyaránt elérhető.

Felhasználóinak száma 2017 szeptemberében lépte át a 800 milliót[11]. Az alábbiakban részletesen ismertetem az Instragram crawler modulját.

### Instagram API

Az Instagram API-ja hasonló elven működik, mint a korábban már bemutatott API-k. Egy bejegyzett fiók birtokában lehetőség nyílik alkalmazást regisztrálni, amihez egy kliensazonosító tartozik. Ezt az azonosítót felhasználva kérhetünk tokent az alkalmazásunkhoz. Az alkalmazás kódjában már csak a kapott tokent kell megadni, ami a Twitter-nél látott, négy különböző API key-t és egyéb tokeneket kérő megoldáshoz képest egy lényegesen egyszerűbb forma.

### A keresés megvalósítása

Az Instagram bányászatakor az InstagramAPI[12] nevet viselő Python modult használtam. Ez lehetőséged biztosít felhasználónévre keresni. A kereső a speciális karakterek jelenlétére érzékenyen reagált ezért ebben az esetben kiemelten szükség van a szűrésre. A keresés eredményét követők száma szerinti csökkenő sorrendben adja vissza, ez pedig megfelel az eddig alkalmazott alap validálási mechanizmusnak. Az így kiválasztott profil bővebb adatait egy következő lekérdezés adja meg, amely a profil egyedi azonosítóját kapja paraméterként.

Az eredményt JSON formátumban adja vissza a program. Az ezzel az eljárással gyűjthető adatok számossága messze nem olyan nagy, mint a korábbiakban ismertetett két modulé, de ez főképp abból fakad, hogy az Instagram lényegesen lazább, kötetlenebb, mint az eddig ismertetett oldalak, tehát jóval kevesebb profiladatot kér. Ezt a problémát néhány származtatott mérőszám létrehozásával orvosoltam. Az Instagram-os aktivitást nem csupán a követők vagy követettek száma írja le jól, hanem a közzétett képek és videók mennyisége is, ezért ezt is beválogattam a gyűjtendők közé. Ezeken kívül elérhetőek még a közzétett képek részletes adatai is, amiket részben felhasználtam. Így az Instagram crawler gyűjti a legtöbb adatot, szám szerint tizenhármat.

### Az Instagram-ról gyűjtött attribútumok

A gyűjtött attribútumok az alábbiak:

* username  
  A fiókhoz tartozó felhasználónév
* id  
  A fiók egyedi azonosítója
* profile\_pic\_url  
  A fiókhoz tartozó profilképre mutató hivatkozás
* instagram\_media\_count  
  A fiók által közzétett képek és videók száma
* followers  
  A fiók követőinek száma
* follows  
  A fiók által követett profilok
* external\_url  
  A fiókhoz tartozó hivatkozás
* instagram\_last\_like  
  Az utolsó kép kedvelőinek száma
* last\_comment  
  Az utolsó kép kommentjeinek szma
* avg\_like  
  Az utolsó tíz kép átlagos kedvelői száma
* instagram\_avg\_comment  
  Az utolsó tíz kép átlagos kommentszáma
* last\_img\_url  
  Az utoljára publikált kép hivatkozása
* hashtags  
  Az utolsó tíz képhez tartozó hashtag-ből képzett lista

### Az Instagram crawler értékelése

A három vizsgált közösségi oldal közül az Instagram mutatta a legkisebb lefedettséget. Ez hozzávetőlegesen 80%-ot jelent. Ennek oka szintén a laza formátumban és elnevezési szokásokban keresendő.

Bár az API által visszaadott attribútumok számossága alacsony, néhány származtatott adattal, például az utolsó tíz közzétett kép mérőszámaival már értékelhető mennyiségű leíró érhető el.

## Az adatok tárolása

### Az SQLite rendszer

Az adatok tárolását a hagyományos CSV fájlok mellett SQLite3 nevet viselő relációs adatbáziskezelő rendszerben valósítottam meg. Az SQLite számos programozási nyelvből és fejlesztői környezetből használható kényelmesen, köztük Pythonból is, amely nyelvet a gyűjtőmodulok (és később a megjelenítő modul) is használnak.

Ez a megvalósítás platformfüggetlen és egyszerű felépítésű, ezért könnyen hordozható. Mivel a tárolásra szánt adataim között nincsenek bonyolult kapcsolati rendszerek és maguk a tárolandó adatok sem képviselnek jelentős méretet, ez a pehelysúlyú tárolási megfontolás éppen elegendő a feladatok elvégzésre.

Ezen kívül a Python nyelven jól használható Pandas nevű, adatelemzést segítő modul nem csak a leggyakrabban használt CSV formátummal boldogul, hanem az SQLite-tal is, így nincs szükség más eszköz bevonására az elemzések során.

### Statisztikák

Ahhoz, hogy előzetesen visszamérjem és megismerjem az adatgyűjtő modul viselkedését, készítettem néhány statisztikát a vizsgált csapatokról. Ebből most néhány konkrét értéket ismertetek. Országok szerinti eloszlás tekintetében az 1. táblázat nyújt felvilágosítást.

|  |  |
| --- | --- |
| Ország | Darabszám |
| Spain | 20 |
| France | 20 |
| United Kingdom | 20 |
| Italy | 20 |
| Germany | 18 |
| Turkey | 3 |
| Argentina | 2 |
| United States | 2 |
| Brasil | 1 |
| Egypt | 1 |

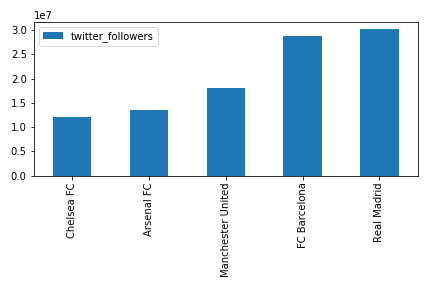
1. táblázat - a vizsgált csapatok országok szerinti eloszlása

Az európai topligákból, tehát Spanyolországból, Franciaországból, Angliából és Olaszországból egyaránt 20-20 csapatot választottam ki elemzésre. Az országukat tekintve kevésbé érdekes, azonban csapatszinten jelentős tömegeket vonzó együttesek a táblázat alján láthatók. Ilyen országok Argentína, az USA, Törökország, Brazília és Egyiptom. Ezeket a csapatokat azért választottam ki, hogy némileg árnyaljam a sztárcsapatokról kialakult képet azzal, hogy megmutatom, hogy néhány kevésbé felkapott országban is található milliós rajongói bázist bevonzó egyesület.

A legáltalánosabb és legmeghatározóbb jellemző a közösségi oldalak publikus profiljainál a kedvelők és a követők számának alakulása. Ennek alapján a teszteléshez generált futballklub lista minden elemére lekérdeztem ezeket az értékeket. Ezen felül jól leírja a csapatok népszerűségét a közzétett képek és bejegyzések száma, illetve az is, hogy ők mennyire aktívan követnek más oldalakat.

Mivel a Facebook használata a feladat második félévében lényegében meghiúsult, ezért az alábbi statisztikák csak a másik két adatforrás alapján készültek. Az itt szereplő konkrét számokhoz a gyűjtést 2018. április 29-én végeztem. Minden elemzett adattípusnál az adott érték szerinti első öt klubot hasonlítottam össze.

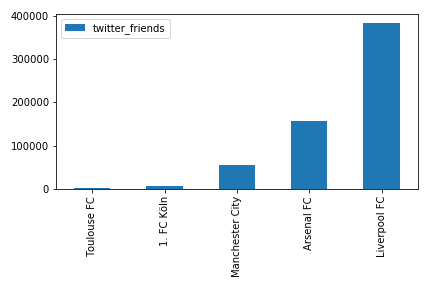
A Twitter vizsgálata során gyűjtöttem a követők, a barátok és a közzétett állapotok, szöveges üzenetek mennyiségét. Az 1. ábra a Twitter-es kedvelésszám szerint kapott Top 5 csapatot tartalmazza, követőszám szerinti növekvő sorrendben.



1. ábra - Az öt legtöbb Twitter követővel rendelkező csapat

Az első helyen a Real Madrid található mintegy kétmillió kedvelővel megelőzve a Barcelona oldalát. A dobogó legalsó fokára felférő Manchester United is közel húszmillió rajongóval büszkélkedhet, utánuk viszont már egy nagyobb szakadék van. Az Arsenal és a Chelsea hasonló népszerűségnek örvend.

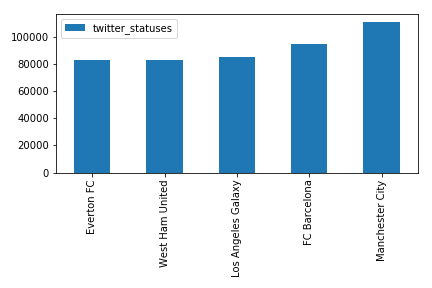
Kapcsolatépítés terén meglepő módon három brit klub sajátította ki a dobogót. Ezeket a számadatokat a 2. ábra rejti.



2. ábra - Az öt legaktívabb kapcsolatépítő csapat

A Liverpool a legaktívabb ezen a téren. Őket követi jelentős lemaradással a londoni Arsenal, és a Manchester City. A 4. helytől lefelé drasztikusan lecsökken az ilyen irányú aktivitás mértéke.

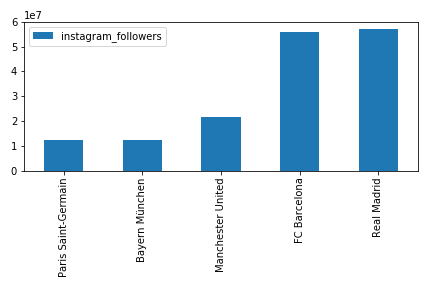
A Twitter lényege a szöveges üzenet közzététele, nem véletlen tehát, hogy ezek mennyisége is kiemelten fontos leíró adat. A 3. ábra ezekről szól.



3. ábra - Az ö legtöbb Tweet-et kiposztoló csapat

Az első helyen ezúttal a Manchester City hivatalos oldala szerepel százezernél is több Tweet-tel. Utánuk a Barcelona és az USA-ból beválasztott L.A. Galaxy következik. Őket a West Ham United és az Everton követi fej-fej mellett közel nyolcvanháromezer státusszal.

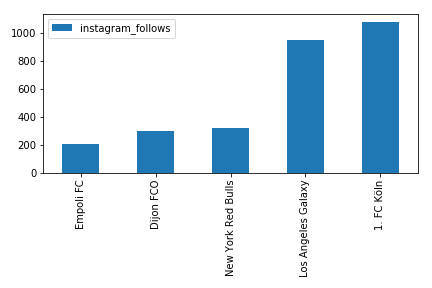
Ami az Instagram-ot illeti, a követők számának ranglistáján ugyanaz a három klub jeleskedik, mint a Twitter esetén, ráadásul a sorrendjük is változatlan. A konkrét számok az 4. ábrán láthatók.



4. ábra - Az öt legtöbb Instagram követővel rendelkező csapat

A Real Madrid 57 millió követői bázisa szinte a duplája a Twitter-es értéknek, de ugyanilyen arány figyelhető meg a másik spanyol egyesületnél, a Barcelonánál is. A Manchester United esetén messze nem ilyen jelentős az eltérés a két platform között.

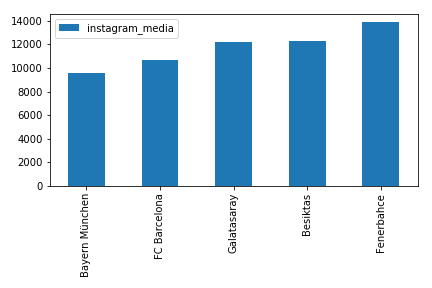
A felhasználók vagy más csapatok, profilok irányába az Instagramnál már nem olyan számottevő az aktivitás. Az 5. ábrán látható, hogy a leginkább kapcsolatépítő csapat, az FC Köln is alig több, mint ezer barátot, visszakövetést tudhat magáénak.



5. ábra - Az öt legtöbb visszakövetéssel rendelkező csapat

A Köln csapatát két amerikai klub, az LA Galaxy és a New York Red Bulls követi jelentősen lemaradva. A Twitter-rel összemérve sokkal alacsonyabb a kétirányú kapcsolat, de ennek az is az oka, hogy ezen a platformon nincs klasszikus értelemben vett baráti kapcsolat definiálva.

Ami a közzétett képek és videók számát illeti, a 6. ábra adatai alapján egyhangú török fölény figyelhető meg.



6. ábra - Az öt legtöbb Instagram képet kiposztoló csapat

A gyűjtésre és elemzésre kiválasztott három török csapat foglalja el az első három helyet, egyaránt tizenkétezernél is több publikált médiatartalommal. A negyedik és ötödik helyen már kevésbé okoz meglepetést a Barcelona és a német Bayern München jelenléte egyaránt tízezer körüli megosztott tartalommal.

### Megjelenítési módokhoz szükséges adathalmazok kialakítása

Egy ilyen, adatokkal foglalkozó feladat során, mint amilyen a közösségi média aktivitás megfigyelése, szükség van valamilyen jól konfigurálható adattárolási megoldásra. Ahhoz, hogy ezt a számunkra legmegfelelőbb módon tudjuk eszközölni, szükség van arra, hogy átgondoljuk, pontosan mik azok az adatok, amelyeket érdemes tárolni, és mik azok, amiknek a tárolása irreleváns.

A későbbiekben tárgyalt validálási megoldás hibái és hiányosságai miatt fontosnak tartottam kiépíteni egy cache-t, amiben kezdetben százas nagyságrendben tároltam el népszerű futballcsapatok neveit aszerint, hogy milyen formában szerepelnek az egyes platformokon. Emellett szintén eltároltam az ID értékeket, amiket a három közösségi oldalon használnak. Ezeket elég egyszer gyűjteni és tárolni, mivel nem változnak az idők során.

Szükséges lehet még egy vagy több földrajzi információ tárolása is a csapatokról, főleg akkor, amikor az átlagos, futballt legfeljebb felszínesen ismerő felhasználóra gondolunk. Ehhez tároltam a Twitter lokációs adatát, melyek kézzel korrigáltam. Ez általában várost jelölt, így ezt a tudást bővítettem még az ország megadásával is. Ezeket a cache-ben rögzítettem a többi statikus adattal együtt.

Ezen kívül a historikus adatok tárolása is hasznos lehet, mivel egyik oldal sem támogatja a korábbi dátumokhoz tartozó támogatószámok visszakeresését. Itt szintén a népszerű csapatoknál van értelme rögzíteni naponta egyszer az aktuális követőszámot. Ezen felül tárolom még dinamikusan, naponta lekérdezve és kiszámítva az utolsó tíz közzétett tartalom átlagos kedveltségét és a rájuk adott hozzászólások átlagos számát.

### Az adatgyűjtés ütemezése

Az adatok legyűjtése az alapműködéshez használt 108 csapat esetén megközelítőleg egy órát vesz igénybe. Ezért az adatgyűjtés ütemezésekor a napi egy lekérdezést állítottam be automatikusan, mindig ugyanabban az időpontban. Ez még bőven elegendő egy hosszútávú, több hónapnyi adatot tartalmazó halmaz elkészítésére, és az ez idő alatt végbemenő változások lekövetésére.

### Az adatgyűjtés során felmerült nehézségek és megoldásaik

Az egyes forrásokból elérhető adatok összegyűjtése során az alábbi főbb problémákkal kellett szembenéznem:

* Downtime  
  Hálózati hiba vagy más, a gyűjtés folytonosságát meggátoló probléma esetén a hibakezelő ág egy óra elteltével újra kísérletet tesz a lekérdezés megismétlésére. A tapasztalataim szerint ez a megvalósítás kielégítően orvosolta a felmerült hibalehetőségeket.
* Kimaradó értékek, inaktív fiókok  
  Amennyiben az adatok folytonossága is veszélybe került, például legalább egy napig nincs elérhető adat, úgy a hiányzó napokat a meglévő környező napok átlagával képeztem egy erre a célra létrehozott függvénnyel. Erre a gyakorlatban egyetlen alkalommal volt szükség, illetve a Facebook-ról gyűjtött adatok hiányának mértéke már meghaladja ennek az eljárásnak a hatáskörét.
* A környezetükből kiugró, irreális értékek  
  Előfordult olyan eset is, amikor 1-1 érték irreális mértékben eltért az előző és a következő nap során mért értékről. Ezen értékek detektálására definiáltam egy függvényt, ami paraméterezhető százalékos érték fölötti növekedés esetén végrehajtja a visszasimító lépést. Az outlier adatokat visszasimítottam a többi közé átlagképzéssel. Nem volt számottevő outlier a teljes adathalmazban a vizsgált időszakban.

## Az adatgyűjtő modul értékelése

A dolgozatomban bemutatott szolgáltatás első és egyik legfontosabb része az adatgyűjtés. Ennek során közösségi oldalakról gyűjtöttem le a megadott input névre illeszkedő oldalak adatait. A három közösségi oldal vizsgálata során kiválogattam a számomra lényeges attribútumokat.

Ahhoz, hogy ezt el tudjam végezni, megismertem három közösségi oldal API-ját. A gyűjtött adatok egy részét egy jól kezelhető és könnyen feldolgozható formátumban eltároltam. Az adatgyűjtő modult alkalmassá tettem a későbbi bővíthetőségre mind gyűjtött attribútumok, mind további közösségi oldalak tekintetében.

# A kapott adatok validálása, a keresett profilok pontos beazonosítása

Ahhoz, hogy az alkalmazás megbízhatóan és pontosan működjön, fontos biztosítani azt a tulajdonságot, hogy a visszaadott profiladatok pontosak és relevánsak legyenek. Ennek megoldására több lehetőséget is véggondoltam, míg több iteráció után kialakítottam egy kellően pontos formulát. A lehetséges alternatívákat a következő rész ismerteti.

## Ötletek, fejlesztési irányok

### Legpontosabb egyezés

Kézenfekvő megoldás az, ha a megadott kulcsszóra, kulcsszavakra a legpontosabban illeszkedő találatot tekintjük a megfelelőnek. Sok esetben valóban működik ez a szisztéma, de a tesztelés során egyre többször futottam bele abba a jelenségbe, hogy a tökéletes egyezés sem elégséges. Például, ha egy csapatnak a rövidített alakja szerepel a hivatalos oldalnál és nem a kézenfekvő, ám hosszú forma, akkor a gyűjtőmodul rendszerint pontatlanul működött.

Official szó jelenléte egy oldal esetén árulkodó lehet, de nem elég általános jelenség. Különösen ritkán fordul elő a kis rajongói bázissal rendelkező klubok esetén.

### Legtöbb követővel rendelkező

A másik, valamivel eredményesebb működi mód volt az, amikor a találati listát követők száma szerint csökkenő sorrendbe rendeztem és a lista első elemét tekintettem a keresett oldalnak. Ezt mindhárom forrásra el tudtam végezni, ugyanis a kedvelők és követők száma egy olyan érték, amit az összes forrásnál el tudtam érni.

Általában elégséges is lehetne ez a gondolkodásmód, de viszonylag hamar találkoztam is egy súlyos tévedéssel. A Panathinaikos keresőkifejezés ugyanis a görög kosárlabdacsapatot adja vissza, ugyanis hiába ismert az azonos nevű futballegyesület, a kosár szakosztály európai, sőt, világszinten is lényegesen magasabb szinten jegyzett. Ez a tulajdonság pedig a követők számában is megmutatkozik.

### Cache-ben tárolt, kézzel validált halmaz alapján

A végleges megoldásnak a cache alapú, validált halmazból történő keresést választottam. Ez ugyan időigényes, de csak egyszer kell elvégezni. Ennek oka abban keresendő, hogy egy jól felfuttatott, milliós támogatói bázissal rendelkező oldal kis eséllyel fog nevet változtatni, az ID érték pedig szinte teljesen biztosan megmarad.

A cache-t a legbiztosabban kézi ellenőrzéssel lehet pontossá tenni. A befektetett munka eredménye egy 100%-ban pontos és ellenőrzött halmaz, ami lefedi az összes szükséges attribútumot. Ezt érdemes minden népszerű csapatra elvégezni.

A cache használatával akár az input field is kiiktatható, legördülő vagy ikonos, esetleg hierarchikus menü elég. Emellett fenntartható egy szabadszavas keresési lehetőség is, amiből a gyakorta előforduló keresőkifejezéseket naplózni lehet és idővel beépíteni a cache-be.

Tárolni elegendő a pontos neveket és a hozzájuk kapcsolódó ID értékeket. Az elkészített crawler-ek ezek alapján már el tudják végezni a keresési feladatokat.

## Validáláskor felmerült problémák

A megtalált oldalak validálása az egyik kulcsfeladata az elkészült alkalmazásnak, ezért kiemelten fontos, hogy ez a modul a legjobban működjön. A vizsgálat során tapasztalt érdekes hibaforrásokról az alábbiakban írok.

A szabadszavas keresés elgépelés ellen nem védett. Ezt külön kellene beleépíteni, ami megoldható, de a korábban említett, pontatlanságot okozó nehézségeket nem oldaná meg.

Nehézséget okozhatnak még az egyes névalternatívák, becenevek. Elég csak a Ferencváros-Fradi, esetleg a Tottenham-Spurs név-becenév párokra gondolni, de például a Manchester Unitedre használt „Red Devils” is elég jellegzetes hivatkozás. Ezeket külön kezelni alapos ismereteket igényelne és jelentős többletmunkát okozna. Ugyancsak problémás, hogy szinte az összes csapat nevének van mozaikszavas megfelelője, vagy más rövidítésre. A legkirívóbb ilyen példa az Olympique Lyon nevű francia csapatra használt „OL” profilnév, amely egy hagyományos kereső számára akár túl rövid is lehet.

Vannak olyan kifejezések, amelyek számos egyesület nevében szerepelnek. A legáltalánosabb „Club” szó mellett ilyen még angolszász területeken például a „United”. Konkrét példákat az 5. táblázat tartalmaz.

|  |
| --- |
| Gyakori névelemek |
| FC, SC, FK |
| City |
| Club |
| Athletic |
| United |
| Town |
| Atletico |
| Real |
| Dinamo, Torpedo |
| Lokomotiv |

2. táblázat - Gyakori névelemek futballcsapatok neveiben

Ezeken felül nem jelentenek jelentős megkülönböztető erőt a városok, kiváltképp a fővárosok nevei, mivel azokban jellemzően több csapat is működik. Fontos kezelni ezeket a névelemeket, ugyanis jelenlétük, vagy éppen hiányuk érzékenyen érintheti a modulok működését.

## A validálás értékelése

Összességében megállapítható, hogy bár sokféle fejlesztési irány elképzelhető a gyűjtött adatok validálása során, a fix cache használata tűnik a legpontosabb és legmegbízhatóbb választásnak. Előnye, hogy gyors, pontos, illetve mivel a futballcsapatok száma véges, előbb-utóbb teljes lefedettséget is biztosíthat. Cserébe viszont korlátozott, csak az adott témakörre illik rá, ezért a szabadszavas keresés lehetősége sem elvetendő. Az alkalmazás felajánlható alternatívákkal félig automatizálható marad, ezért egy félig cache-re épülő, félig szabadszavas keresőmodul nyújthatja a legoptimálisabb megoldást.

# Megjelenítés

Az eddig tárgyalt modulokban ismertettem, hogyan működik az általam implementált közösségi média aktivitást figyelő alkalmazásom adatgyűjtő és profilokat ellenőrző modulja. Az így kapott adathalmaz ugyan már pontos és releváns, de egy átlagos felhasználó számára nem több néhány számnál. A soron következő lépés a gyűjtött adatok érthető és látványos reprezentálása.

A megjelenítő modul során több alapfeltételt is kitűztem célul. Az egyik, hogy az ábrák legyenek érthetőek lényegében bárki számára. Fontos, hogy nem feltételezhetünk semmilyen előzetes ismeretet a témában a felhasználótól, ezért szükség lehet az egyes elemek, esetünkben csapatok kontextusba helyezésére. Mivel döntően numerikus adatokat kell vizuálisan ábrázolni, ezért különféle grafikon- és diagramtípusok használatában gondolkodtam. Szintén lényeges, hogy a megjelenítés ne legyen túl tömény, hanem még emészthető mennyiségű információt mutassanak meg a felületek.

A másik szempont, amit választottam, az interaktivitás biztosítása. Ne csupán statikus ábrákat tárjak a felhasználók elé, hanem módjuk legyen szűkíteni-bővíteni a látni kívánt attribútumokat, oszlopokat és egyéb elemeket. Az egyes lekérdezések személyre szabhatósága is elengedhetetlen, ezt viszonylag könnyen biztosítani lehet egy bizonyos szabadsági fokig.

A megjelenítő modul tervezése és implementálása során átgondoltam, hogy melyek lehetnek azok a metrikák, amelyek a legjobban jellemezhetik a közösségi oldalon mutatott aktivitást. Számításba vettem, hogy milyen kiinduló halmazból dolgozhatok. Ezen felül áttekintettem a további bevonható forrásokat, például más adathalmazokat, amelyek hatással lehetnek a csapatok közösségi média oldalainak sikeréhez vagy bukásához. Az alábbiakban a megjelenítő modult ismertetem részletesen.

## Igények és célcsoportok definiálása

Ahhoz, hogy valóban hasznos és informatív vizualizációt tudjak a potenciális felhasználók elé tárni, szükséges áttekinteni, hogy milyen rendelkezésre álló forrásokból és kiknek akarok információt átadni. Ezért a megjelenítéssel foglalkozó modul első állomásaként kérdéseket fogalmaztam meg, amelyek válaszaiból már el lehet indulni a vázlatos felülettervek, így a későbbi implementációs fázis felé.

### Milyen adatokból dolgozom?

Amint azt az adatgyűjtő modulról szóló szakaszban már ismertettem, futballcsapatok közösségi média aktivitásáról gyűjtök le adatokat. Ebben beletartoznak a klubok olyan alapadatai, amelyek nem, vagy csak nagyon ritkán változnak az idők során, de beletartoznak időben dinamikusan változó mérőszámok is, például a rajongók és a közétett tartalmak száma. Az adatok egy része tehát folyamatosan változik, így szükséges egy olyan eszközt választani, amely támogatja, hogy a megjelenítésre szánt metrikák megváltozása mellett is folyamatosan naprakész legyen a vizualizáció.

Egy futballcsapat népszerűségét nagyban befolyásolják az elért eredmények, ezért importáltam még olyan adatokat is, amelyek a pályán nyújtott teljesítményre vonatkoznak. Ezeket részben egy már meglévő adathalmazomból tudtam lekérdezni.

### Kik lehetnek a felhasználók?

A kiindulási feltételezésem az volt, hogy úgy tekintek a riportokra, hogy azok bárki számára értelmezhetők legyenek. Alapelvem szerint nem követelek meg előismereteket sem a futball, illetve a konkrét klubok, sem a közösségi médiák terén.

Mivel a futball hazánkban és Európában a legnépszerűbb csapatsportok között van, ezért szinte bárki meg tud nevezni egy-egy klubot. Ha a klub nemzetiség szerinti hovatartozása a kérdés, a többség még viszonylag stabil válaszokat tud adni egy szűkebb, az eredményesebb csapatokra vonatkozó szűrletre, de például egy jellemzően alsóházi egyesületet már nehezebben helyeznek el a térképen. Ha a várost vesszük alapul, ott már az ismertebb gárdákat is körülményesebben tájolja be az egyszeri felhasználó. Ha pedig Európán kívüli az adott egylet, akkor már nagyon ritkák a helyes válaszok. Éppen ezért fontosnak tartom, hogy a földrajzi elhelyezkedést valamilyen módon megjelenítsem egy globálisabb és egy országon belüli nézettel. A Twitter-ről gyűjtött lokációs adat néhány transzformációs lépés után alkalmas lehet a feladatra.

Szintén fontos jellemzők egy csapatnál a klubszínek, melyek a mezen és általában a címeren hangsúlyosan megjelennek. Ezeket a színeket is szerettem volna szerepeltetni legalább az egyik felületen. Erre alkalmas lehet egy kép, címer vagy a színek tényleges alkalmazása a kinézet meghatározásánál.

A mérőszámok mellett, amelyek jellemzik egy csapat közösségi média aktivitását, érdekes lehet néhány szöveges adat is. A hashtag-ek tökéletesen megfelelnek a célra, kinyerésük már megoldott az adatgyűjtés során. A bennük szereplő tömör információ segíthet betájolni az aktuális trendeket a vizsgált egyesület háza táján, például az elmúlt, vagy a soron következő mérkőzést, netán az éppen közelgő rangadó kiemelt szerepét (pl. #derbytime).

### Milyen méretet és formát válasszak?

Amikor az ember megjelenítési tervet készít, az egyik alapkérdés, hogy milyen platformokon lehet megtekinteni az eredményt. Ez még digitális eszközöknél maradva sem egyértelmű, tekintettel kell lenni a megjelenítő eszközök sokféleségére mind méretben, mind teljesítményben. Arra is gondolni kell, hogy gyakran adódik olyan helyzet, amikor egy-egy kimutatásra offline van szükség. A nyomtatásban rejlő potenciál pedig legalább annyira sokféle lehet, mint a képernyőméret.

A felmerült eshetőségeket összefogva arra a meggondolásra jutottam, hogy a hagyományosnak mondható A4-es lapméretre, illetve annak arányaira való optimalizálás jó kiindulási alap lehet.

A méreten túl természetesen az is lényeges kérdés, hogy milyen és mennyi adat kerüljön a felületekre. A kevesebb jelen esetben több, hiszen a túl sok információ között könnyen elvesznek a lényeges részek. Ügyelni kell egyfajta egészséges egyensúly kialakítására. Azokat a mérőszámokat és elemeket érdemes felvenni a felületekre, amik tényleg hordoznak többlettudást a felhasználó számára. A fölösleges, funkció nélküli látványelemek csak a figyelem elvonására alkalmasak, ezért mellőztem a használatukat.

## Megjelenítési tervek

Az implementálás megkezdése előtt szükség van olyan felületi tervekre, amelyeken előzetesen össze lehet állítani a szükséges összetételt és elrendezést. Ezeknek az elkészítésére a Moqups[[4]](#footnote-4) nevű webes alkalmazást használtam fel. A Moqups ingyenesen hozzáférhető bizonyos korlátozásokkal, például csak 300 objektumot lehet egyszerre elhelyezni, de ez a célnak éppen megfelelt. Az előzetes megjelenítési tervek lényege, hogy a tényleges céleszköz ismerete nélkül, szabadon lehet megtervezni az elvárt működést. Ez persze a későbbiekben a választott eszköz lehetőségeinek függvényében változhat.

A korábbi alapgondolatok és igények alapján három terv kidolgozását tűztem ki célul. Az első felületen egy-egy csapat aktuálisan érvényes, tehát egy napnál nem régebbi alapadatai láthatók majd. A második egy összehasonlító dashboard, amin két csapat ugyanazon adatai szerepelnek. A harmadik terv egy csapat időben értelmezhető, historikusan gyűjtött adatait jeleníti meg.

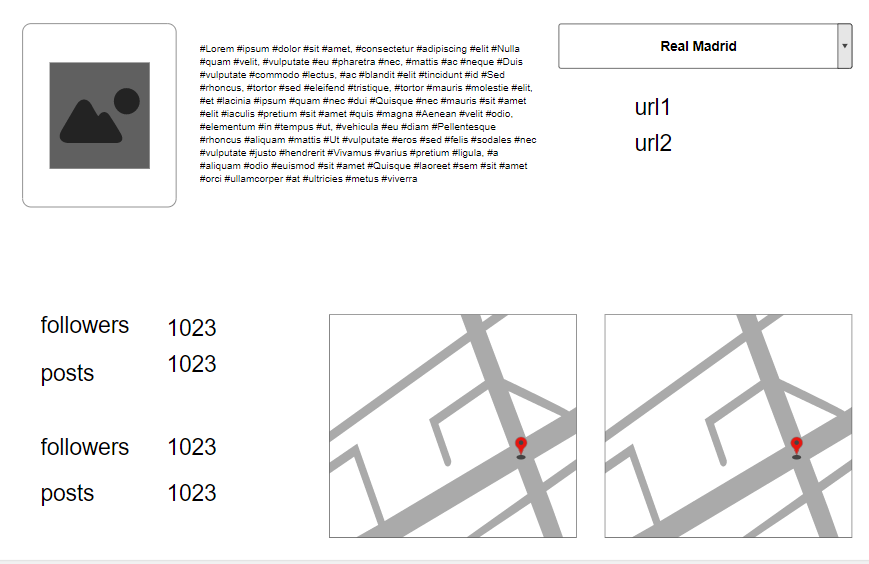
A megjelenítési tervek még nem a végleges és feltétlenül megvalósítható vizualizációt jelentik, de a végleges implementációk elrendezésben nem térnek el jelentősen ezektől. Az alábbiakban az általam választott három felület mockup-jait ismertetem.

### Általános áttekintő felület

Ahhoz, hogy egy futballcsapatot nagyjából be tudjunk tájolni a sportág kontextusán belül, szükséges ismernünk néhány alapvető információt róluk. Ilyen például a földrajzi elhelyezkedés, tehát az ország és a város, a klubszínek és a címer, a főbb mutatók, valamint a hozzáférhető weblapok.

Az első dashboardra ezeket a mértékeket helyeztem az alábbi elrendezés szerint. A jobboldali szűrőn kiválasztható a vizsgálandó klub, középen a legyűjtött hashtag-ekből képzett wordcloud látható, míg baloldalon a klubcímert helyeztem el. Jobboldalon lentebb a két térképes nézet van, míg a bal szélen a közösségi oldalaik főbb mutatói, úgy, mint a követők száma, illetve a közzétett tartalmak mennyisége. Utóbbi metrikák mindig naprakészek a naponkénti adatgyűjtésnek köszönhetően.

Az általános áttekintő felület megjelenítési tervék a 7. ábra tartalmazza.



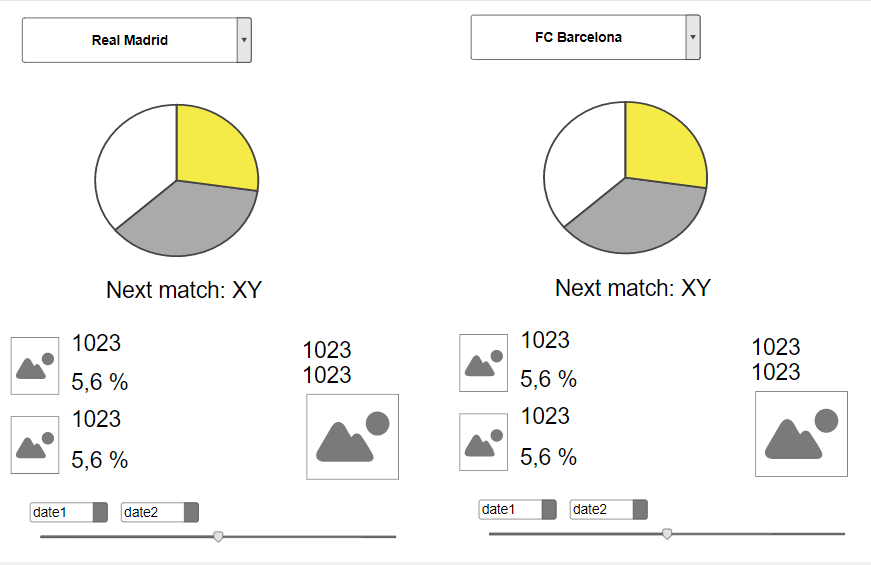
7. ábra - Az általános áttekintő felület megjelenítési terve

### Összehasonlító felület

Az összehasonlító felületen két csapat ugyanazon metrikáit lehet megjeleníteni. A csapatok kiválasztásáról egy-egy dropdown szeletelő gondoskodik. A felülethez csatoltan elérhető a csapat éves, a pályán nyújtott teljesítménye és az ebből képzett kördiagram, valamint a soron következő ellenfél. A kördiagram a szezon korábbi részében lejátszott mérkőzések eredményeit mutatva győzelem, döntetlen és vereség elosztásban, numerikusan és százalékosan is megjelenítve az egyes kimenetelek eloszlását. Ezeket az adatokat egy általam korábban elkészített hobbiprojektben implementált alkalmazás gyűjti naprakészen össze.

A nézet alsó felében a legutóbb közzétett Instagram poszt, továbbá az időcsúszkával beállítható tartományban mutatott közösségi média aktivitás mérőszámai láthatók. A dátum szűrése megadható kézzel, illetve beállítható csúszkával is. Lehetséges a két csapatnál eltérő időintervallumot vizsgálni.

Az összehasonlító felület megjelenítési terve a 8. ábrán látható.



8. ábra - Az összehasonlító felület megjelenítési terve

### Időtartományban kereső felület

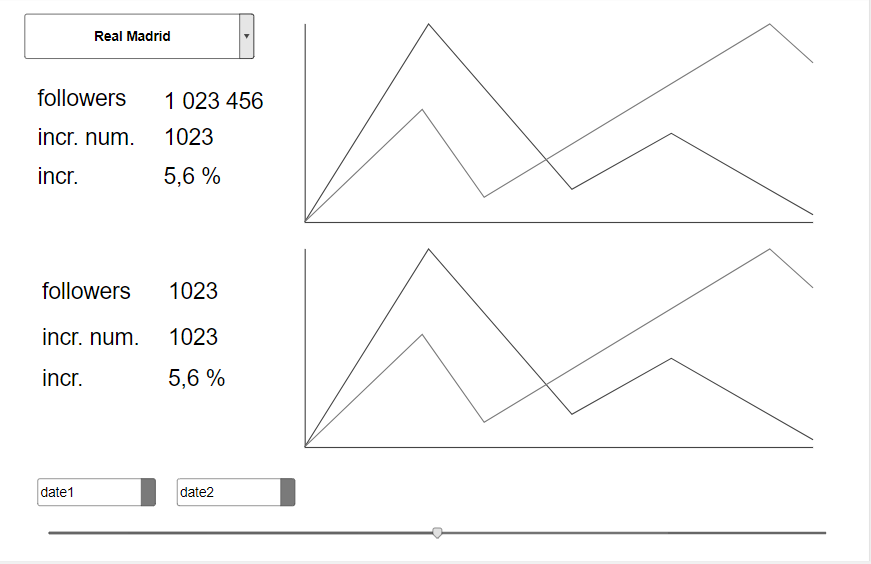
Az időtartományban kereső felületen egy csapathoz kaphatók meg a választott időintervallumra szűkített adatok.

A dropdown szűrőn beállítható a vizsgálandó csapat, melynek az időtartományra értelmezett főbb mérőszámai a szűrő alatt kaptak helyet platformonként csoportosítva. A mérőszámok egy része lehet számított érték, amelyet tehát nem a tárolt adathalmazból nyerünk ki nyersen, hanem a Power BI számít ki a megadott képletek alapján.

Jobboldalt a vonaldiagramok megjelenítik a választott időszakra az egyes platformokon mutatott követői számok alakulását a kijelölt csapatra vonatkozóan. A nézet alján lévő csúszkával tudja a felhasználó paraméterezni a vizsgálandó tartományt.

Ezeken kívül helyt kaphat még egy vagy több mérőszám, amelyek a következő rövidebb időszakra (nap, hét) megbecsülik a várható követőszámot a korábbi trendet alapul véve.

Az időtartományban kereső felület megjelenítési terve a 9. ábrán látható.



9. ábra -Az időtartományban kereső felület

## Power BI

A vizualizálást megvalósító eszköznek a Power BI online, valamint Desktop változatát választottam. Az alábbiakban ismertetem a hozzá kapcsolódó új tapasztalatokat, a felhasználásában rejlő lehetőségeket.

### Power BI Desktop

A Power BI Desktop a Power BI asztali társalkalmazása. Segítségével sokféle adatforrásból importálhatunk adatokat. Az adatforráshoz való kapcsolódás után az elemzési és jelentéskészítési igényeknek megfelelően formázhatjuk az adatokat. A táblák közti kapcsolatokat a szoftver automatikusan is tudja detektálni, de a felhasználó is létrehozhat kapcsolatokat. A jelentésnézetben kiválaszthatók a használni kívánt mezők, szűrők vehetők fel. A Desktop-ban létrehozott és mentett jelentéseket feltölthetjük Power BI-webhelyére és megoszthatjuk másokkal. Ugyanitt a Desktop-ban használt eszköztár használatával módosíthatók is a riportok. A szoftver jól kezeli a háttéradatok időbeli változásait. A frissítés ütemezhető, de végrehajtható manuálisan is.

## Felhasználható felületi elemek

A Power BI Desktop bőséges választékot kínál különböző felületi elemekből. Ezek között vannak kifejezetten a megjelenítést szolgálók, és vannak vezérlők, amelyek a láttatni kívánt adatok pontosítására, leszűkítésére valók. Az alapcsomagban szereplő típusokon felül lehetőség van száznál is több online elérhető kiegészítő beépülő modul használatára is. Az alábbiakban a Power BI vizualizációs palettáját tekintem át.

### Diagramok

Az adatok közti kapcsolatok, illetve időbeli viselkedésük megmutatására a legkézenfekvőbb ábrázolási forma valamilyen diagramtípus alkalmazása. Az általam választott eszközben számos diagramfajta megtalálható, úgy, mint oszlop, sáv, vonal, terület, pont, esetleg kör vagy fánk. Ezek közül vannak, amelyeket érdemes felhasználni, például idősoros megjelenítésre, vagy az aktuális metrikák egymáshoz való viszonyának szemléltetésére.

Az utolsó kettőt ugyanakkor nem minden esetben javasolt, ugyanis a rész-egész viszony, illetve a nagyságrendek érzékeltetésén túl nem sok mindenre alkalmasak, mégis túl sokszor találkozni velük.

### Szűrők

Az adatok egy általunk érdekelt részhalmazának kiválasztására más eszközökhöz hasonlóan a Power BI is szűrőtípusokat biztosít a felhasználó számára. Ezek a szűrők, vagy más néven szeletelők (slicer) többféle formában alkalmazhatók, melyeket az esetek többségében automatikusan is jól detektál a Power BI. Lehet egyszerű és lehet kombinált is a szeletelő aszerint, hogy egy vagy több táblán alapul a riport. A táblák közti kapcsolatot szintén képes kitalálni a szoftver, de lehetőség van kézzel is definiálni azokat.

A legáltalánosabb használatuk a dropdown mód, amikor a szűrő alapesetben egy sort tölt ki, a választáshoz pedig le kell nyitni azt. Ez helytakarékossági szempontokat nézve kifejezetten előnyös viselkedés. Ezek a szűrők módot adhatnak a hagyományos egyes kijelölés mellett többes kijelölésre, vagy akár az összes entitás egyszerre való szelektálására is.

Ha idősoros adatokat szeretnénk lekorlátozni, akkor a csúszka alapú formátum lehet a legszerencsésebb választás, melyhez szöveges dátum mezők is társulhatnak a még pontosabb kiválasztáshoz.

A felsoroltakon túl a különböző diagramok és más megjelenítési módok is viselkedhetnek szeletelőként, ha rákattintunk egy-egy oszlopára vagy más vizuális egységére. Például a később tárgyalt térképes nézet is képes lehet a szelektálás támogatására.

### Kártyák, táblázatok

Néha csupán egy-egy szám a legfontosabb, amit nyomon szeretnénk követni jelentésben. Az ilyen típusú vizualizációkat kártyáknak nevezzük a Power BI-ban. Egy kártya lehet egy vagy több soros.

Ezeken kívül elhelyezhetünk fő teljesítménymutatót (KPI) is. A KPI olyan vizuális jel, amely egy mérhető cél terén elért előrehaladás mértékét jelzi. A KPI jó választás az előrehaladás mérésére, vagy egy céltól való távolság mérésére.

Amennyiben webes url-ből akarunk képet elhelyezni a jelentésen, úgy szintén valamelyik kártya-vagy táblázattípus jöhet szóba. Mindenekelőtt azonban az adatnézetben is definiálni kell, hogy a kívánt oszlop image-url típusú legyen.

### Térképes megjelenítés

Az igények és célcsoportok meghatározásakor már említést tettem róla, hogy hasznos lenne az egyes csapatok földrajzi ismertetésére valamilyen térképes nézetet létrehozni. Erre alkalmas a Power BI, ugyanis többféle földrajzi nézetet biztosít. A térképeket a Bing biztosítja. Az adatnézetben megfelelően beállított City, Country, esetleg Latitude és Longitude típusok helyes használatával városokat, országokat, vagy tetszőlegesen földrajzi koordinátákat is elhelyezhetünk. Ábrázolhatunk pontszerűen, felskálázva egy-egy mérőszám nagyságával, illetve az egész országot vagy más területet (pl. megyét) kitöltve is.

### Statikus elemek

Szükség lehet időnként olyan látványelemekre is, amelyre nincsen hatással az adatok viselkedése. Ilyenek például a magyarázó feliratok, leírások, valamint a különböző help-ek is. Ezeket meg lehet adni szövegesen, de néhány esetben érdemes képi megoldást választani, például logók vagy piktogramok formájában. A Power BI-ban módunk van ilyen statikus összetevőket is létrehozni. Ezek a komponensek minden riporton ugyanúgy néznek ki.

### Számított oszlopok

A számított oszlopokkal új adatokat adhatunk hozzá a modellben már meglévő táblázathoz. Ezeket a Data Analysis Expressions- (DAX-) képletekkel lehet megadni, amely meghatározza az új oszlop értékeit [13].

A DAX több mint kétszáz függvényt, operátort és szerkezetet tartalmazó kódbázist foglal magába, rendkívüli rugalmasságot biztosítva a képletek létrehozása során. A DAX-képletek az Excel-ből ismertekhez hasonlók. Az Exceltől eltérően, ahol a felhasználó különböző képletet használhat egy táblázat mindegyik sorában, a DAX-képlet a táblázat minden sorának értékét kiszámolja. Az oszlopértékeket szükség szerint újrakalkulálja a rendszer, például az adatok frissítésekor vagy az értékek módosításakor.

### Egyéni vizualizáció importálása

Power BI-jelentések létrehozása vagy szerkesztése során számos különböző vizualizációtípus használatára van lehetőség. Ezeken kívül még száznál is több egyéni vizualizációs típus érhető el, ha nem találnánk az igényinknek megfelelőt a beépítettek között. Az egyéni vizualizációkat fejlesztők hozzák létre az egyéni vizualizációk SDK-val, az üzleti felhasználók így az üzleti igényekhez leginkább igazodó módon tekinthetik meg az adataikat.

Ilyen importálható látványelem lehet a WordCloud, a Gantt-diagram, esetleg valamilyen R nyelvi beépülő modul [14].

## Az elkészült dashboard-ok

Az elkészített felületterveket Power Bi-ban valósítottam meg. Az alábbi szakasz ismerteti a riportok alkotásakor felhasznált elemeket és szerepüket.

### Általános áttekintő felület

Az általános áttekintő felület lényege, hogy egy gyors képet adjon a vizsgált csapatról annak néhány alapadata segítségével. Ebbe beletartoznak a naprakész követői számok, a földrajzi elhelyezkedés és az utóbbi időben publikált médiatartalmak főbb kifejezései, a hashtag-ek, illetve a csapathoz elérhető webes linkek, esetleg YouTube-fiók címe.

Az alapvető kinézetnél a sötét háttérszín dominál, amihez sárga és fehér kiegészítő színeket választottam az adatok megjelenítéséhez.

Felhasznált adatok:

* A legutóbbi gyűjtéskor keletkezett, tehát 24 óránál nem régebbi adatok a Twitter és az Instagram aktivitásról minden cache-ben tárolt csapatra. Az időben nem változó adatok, például az ország és a város természetesen nem frissülnek minden lekérdezéskor.

Felhasznált felületi elemek és funkciójuk:

* Dropdown szeletelő  
  Segítségével megadható a vizsgált csapat neve. A többes kijelölést, beleértve az összes elem választását nem engedélyeztem, ugyanis értelmezhetetlen eredményt adna.
* Tábla  
  A tábla elem használata egy lehetőség a képekhez tartozó webes linkek képként való ábrázolására. Ilyen például a klubcímer. A maximális magasság 150 pixel lehet, ezt ki is használtam a nézeten.
* Kártya  
  Minden időben változó számadatot kártyákra helyeztem el.
* Térkép és kartogram  
  A földrajzi elhelyezkedés illusztrálására két térképes nézetet alkalmaztam. A térképet a Bing szolgáltatja. A háttérhez igazodva a sötét témát választottam. A Country oszlop egy kitöltött országot ad meg, míg a City a csapat székhelyét városként, pontszerűen ábrázolva. A térkép lehetőséget ad automatikus zoomra, amivel éltem is.
* Wordcloud   
  A Wordcloud egy egyéni vizualizációs lehetőség a Power BI honlapjáról letöltve és importálva. Segítségével a csapat Twitter és Instagram tartalmainak legyűjtött hashtag elemeit lehet megmutatni a felhasználónak. A beállításom alapján legfeljebb kétszáz névelem fér a felületre, amelyeket 50% és 100% közötti méretre alakít. A modul a speciális karaktereket (pl. #) felismeri és eldobja.
* Logók  
  Az egyes platformok logóit, továbbá néhány piktogramot a felület sablonjára helyeztem el. Ezek nem változnak a választás függvényében.

Az általános áttekintő felület elkészült dashboard-ja a 10. ábrán látható.



10. ábra - Az általános áttekintő felület dashboard-ja

### Összehasonlító felület

Az összehasonlító felületen két klub azonos mérőszámait lehet megvizsgálni. Ebbe beletartozik a csapatok teljesítménye is, ugyanis jelentősen befolyásolhatja a népszerűséget a pályán nyújtott eredmény.

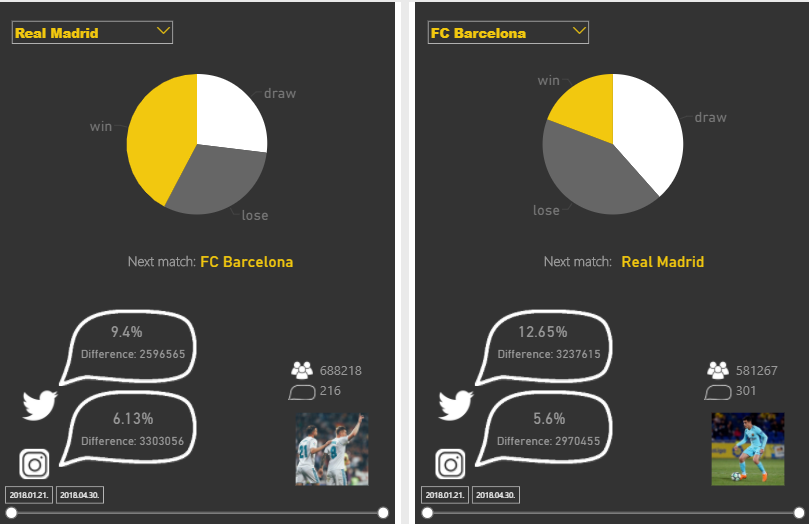
Felhasznált adatok:

* Az összehasonlításhoz felhasználtam a csapatok aktivitásáról készült historikus adathalmazt, továbbá egy külső, eredményeket tartalmazó gyűjteményt [15]. Ez az adathalmaz egymilliónál több futballmérkőzés főbb adatait foglalja magába és ebből a megfelelő csapatnév szerinti szűréssel kinyerhetők a szükséges információk.

Felhasznált felületi elemek és funkciójuk:

* Dropdown szeletelő  
  Ugyanaz igaz rá, mint az első dashboard-nál.
* Csúszka szeletelő  
  Segíti az időtartományban való szűrést. A kívánt érték megadható numerikusan is, ami egy nagyobb tartománynál hasznos lehet.
* Tábla  
  Hasonlóan működik, mint az első esetben, ám ennél a riportnál a vizsgált csapat legutolsó közzétett Instagram tartalmát mutatja meg, 150 pixel magas képként.
* Kördiagram  
  A külső, eredményeket és soron következő meccseket tartalmazó adathalmaz alapján megjeleníti a csapat szezonbeli teljesítményét győzelmek, vereségek és döntetlenek szempontjából.
* Kártya  
  Megjeleníti a kördiagram alatt a soron következő ellenfél nevét. Szintén kártyákkal oldottam meg a legutóbb Instagram poszt mérőszámait is, úgy, mint kedvelések és kommentek száma

Az összehasonlító felületet a 11. ábra mutatja.



11. ábra - Az összehasonlító felület dashboard-ja

### Időtartományban kereső felület

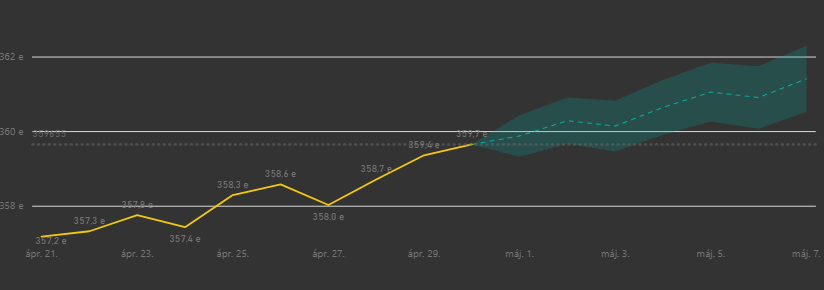
Az időtartományban kereső felület lényege, hogy visszaadja egy klub népszerűségének alakulását két dátumérték között. A változást megadtam számértéként és százalékosan egyaránt. A szemléltetésről vonaldiagramok gondolkodnak. Ezen kívül pedig a Power Bi beépített forecast funkcióját is kipróbáltam a követő számok jövőbeli alakulásának becslésére.

Felhasznált adatok:

* Felhasználtam a csapatokról készült historikus adathalmazt, amely a gyűjtött időszakra tartalmazza a követők, tartalmak stb. számának alakulását csapatonként.

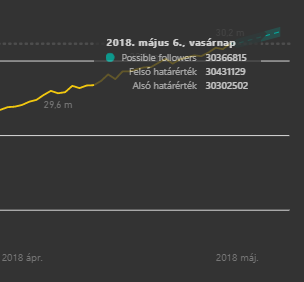
Felhasznált felületi elemek és funkciójuk:

* Dropdown szeletelő  
  Ugyanaz igaz rá, mint az első dashboard-nál.
* Vonaldiagramok  
  A követők alakulásának szemléltetésére vonaldiagramokat hoztam létre mindkét platformra. A diagramok dátumszűrő szerint mutatják a változást. Mivel az elemzett csapatoknál eléggé lineáris volt a változás mértéke, ezért a Power Bi által biztosított előrejelzés(forecast) funkció jó szolgálatot tett a jövőbeli értékek vizsgálatára. A 12. és 13. ábra ezt a funkciót mutatja meg részleteibe menően. A jövőt tekintve egy hétnapos időintervallumot kérdezek le, amit a visszamérések alapján viszonylag pontosan prediktál a beépített funkció. A megtippelt értékekhez alsó és felső határértékek is tartoznak, egyfajta intervallumot megadva a becsült értékhez. Látható, hogy nem csupán a megtanult trend alapján ábrázol egy egyenes vonalat, hanem más sajátosságokat is figyelembe vesz.



12. ábra - A Power BI forecast funkciója 7 napra vetítve

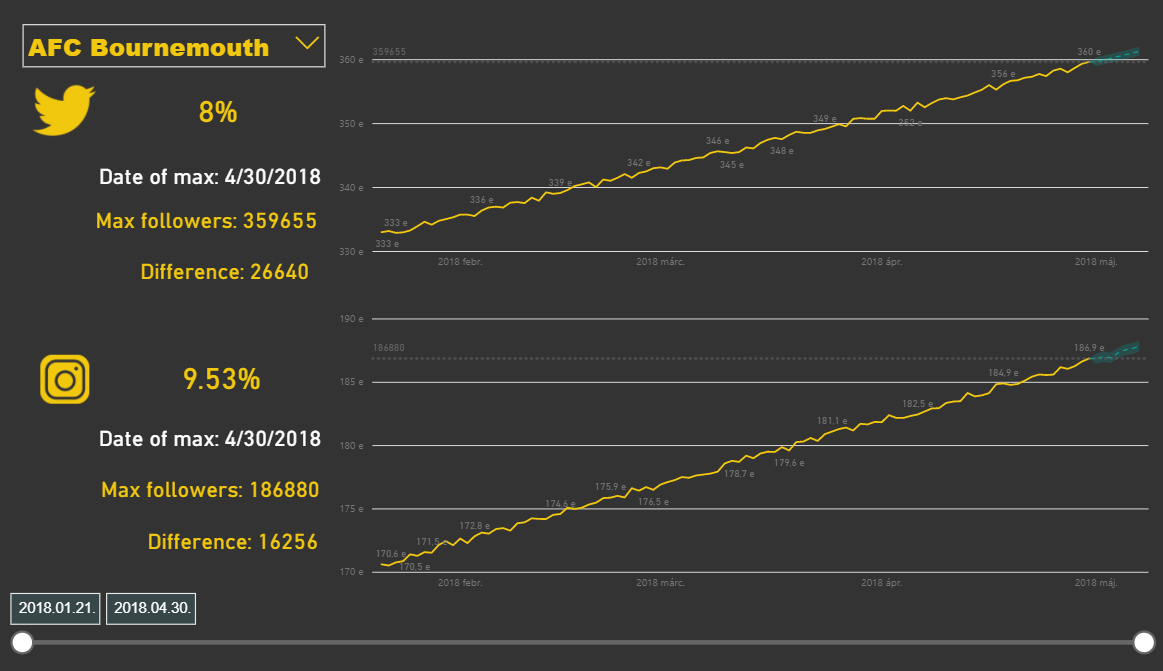
A 13. ábrán látható, hogy az egyes adatpontok fölé húzva az egeret, látható az előrejelzett érték egy-egy dátumhoz rendelve.



13. ábra - A forecast részletező nézete

* Csúszka szeletelő  
  Ugyanaz igaz rá, mint az első dashboard-nál, tehát segíti az időben történő szűrést.
* Kártya  
  Kártyákra helyeztem el a vizsgált időtartományban lévő főbb adatokat, például a maximális követői számot, annak dátumát, valamint a kezdő-és végdátum közti változást különbségként és százalékosan.
* Logók  
  A két közösségi oldal logója statikusan, képként került fel a riportra.

Az időtartományban kereső felületet a 14. ábra jeleníti meg.



14. ábra - Az időtartományban kereső felület dashboard-ja

## A megjelenítés értékelése

Az adatgyűjtő modulban ismertetett mechanizmussal gyűjtött adatok megjelenítésére terveket dolgoztam ki, melyekből három dashboard-ot implementáltam is. A megjelenítéshez megismertem a Power BI-t és annak asztali változatát, valamint az abban rejlő lehetőségeket. Új adatforrásokat vontam be a megjelenítés támogatására. Az elkészült felületeket több iterációban javítottam, módosítottam.

# Továbbfejlesztési lehetőségek

Dolgozatom utolsó részfeladataként számításba vettem az esetleges továbbfejlesztési irányokat. Az alábbi lehetőségeket érzem reális fejlesztési célnak a jövőre nézve.

## Több adatforrás bevonása

Az elemzések változatosságát és a különböző trendek azonosítását nagyban segítené, ha további plusz közösségi oldalakat lehetne bevonni a gyűjtésbe. Ilyen extra adatforrás lehetne például YouTube, ahol jelentős számban képviseltetik magukat futballklubok is. Ez az eddig használt források által gyűjtött webes linkek alapján is megállapítható. Az ott elért nézettségi és követési adatok, kommentek hasznosak volnának elemzési szempontból.

Nem jelent teljesen új forrást, de mindenképpen fontos lépés lenne a Facebook-os adatgyűjtés eredményes újraindítása. A dolgozat készítésekor ez átmenetileg szünetel, de elképzelhető, hogy a Cambridge Analytica-botrány elülése után valamilyen formában újból fel lehetne használni a legnépszerűbb közösségi oldalt is a feladathoz.

## Több attribútumtípus gyűjtése, más elemzési szempontok

Fokozná a vizsgált elemek változatosságát, ha a meglévő forrásokból több attribútumtípus is gyűjtésre kerülne, esetleg számított értékként hozzáadódna a meglévő halmazhoz.

Érdekes lenne például az egyes csapatok követőinek nemek vagy életkor szerinti megoszlását vizsgálni, bár jelenleg egyik API sem tesz lehetővé ilyen irányú elemzést. A legnagyobb esély az országot szerinti változatosságot lenne megfigyelni, de jelenleg ez sem megoldott.

Szintén hasznos funkció lehetne a szentiment analízis bevonása a feladatba, amellyel meg lehetne határozni egy-egy klubra, hogy éppen milyen a megítélése a felhasználók visszajelzései alapján, vagy az általa közétett tartalmak szövege szerint. Erre jelenleg a gyűjtött hashtag-ek halmaza lenne a legalkalmasabb, ám számos nyelvi és formai nehézség miatt végül nem készítettem el ezt a modult. A nyelvi nehézségeket a csapatok sokfélesége jelenti, míg a formait az, hogy a hashtag-ek kizárólag egybeírva szerepelnek. Ezt automatizáltan tokenizálni nem lehet. Szentiment analízisre alkalmas lehetne viszont a közzétett tartalmakra adott kommentek halmaza, ám ezek gyűjtése nem képezte részét a dolgozatomnak.

## Más megjelenítési formák és eszközök bevonása

Látványos opció volna, ha aze gyes dashboard-ok színei tükröznék a megjelenített csapat színeit (paraméteresen megadva). Erre jelenleg nem találtam megoldást a Power BI kínálatában.

Szintén javítana az ábrázolás minőségén, ha az időben változó értékeknél a változás irányát külön szín demonstrálná. Tehát a növekedést zöld, a csökkenést piros színnel jelenítené meg a riport. Erre akadnak alternatívák egy továbbfejlesztett kártya típus formájában letölthető plugin-ként, de használatát túl körülményesnek ítéltem meg.

# Összegzés

A közösségi oldalakon, szabadon hozzáférhető publikus tartalom kimeríthetetlen forrást jelent különböző adatbányászati feladatok megoldásához. Gyűjtésükre nehéz egységes és általános megoldást találni, ugyanakkor forrás-specifikus megvalósítások elkészítése különösebb nehézség nélkül megvalósítható.

A közösségi oldalak sosem látott népszerűségnek örvendenek. Minden valamire való cég igyekszik megjelenni a lehető legtöbb platformon. Nincs ez másképp sportcsapatok esetében sem. Az oldalaikról kinyerhető népszerűségi információk hasznosak lehetnek például potenciális befektetők számára.

Ahhoz, hogy felhasználható adataim legyenek, adatgyűjtést végeztem. Ennek elvégzésére adatgyűjtő modult készítettem, amely három különböző közösségi oldalról képes adott input névre illeszkedő profiladatokat gyűjteni. Megismertem a Facebook, a Twitter és az Instagram publikus API-ját. Elkészítettem egy-egy gyűjtőmodult mind a három közösségi oldalhoz. Elkészítettem egy fő modult, ami összefésüli az egyes gyűjtőmodulok eredményeit.

A releváns találatok biztosítása érdekében végiggondoltam a profilok validálás lehetőségeit. Több megoldást is számításba véve kialakítottam a legpontosabb eredményt biztosító megoldást. Az így összegyűjtött adatok tárolására olyan megoldást javasoltam, amely elősegíti a későbbi feldolgozást. A modulok segítségével lehetővé tettem statisztikák generálását és meg is vizsgáltam azokat.

A megjelenítő modulhoz látványterveket készítettem, amelyek alapján implementálhatók az egyes felületek. Áttekintettem a megjelenítésben rejlő lehetőségeket és igényeket.

Megismertem a Power Bi vizualizációs eszközt és annak képességeit. Implementáltam három dashboard-ot a megjelenítési tervek alapján, amelyek segítségével a gyűjtött adatok értelmezhető formába hozhatók.

Végül megoldásokat javasoltam a szoftver továbbfejlesztésére. Átnéztem a felmerülő lehetőségeket és azok megvalósíthatóságát.

# Függelék

## A cache-ben tárolt csapatok és főbb mutatóik

(Gyűjtés időpontja: 2018. 04.05.)

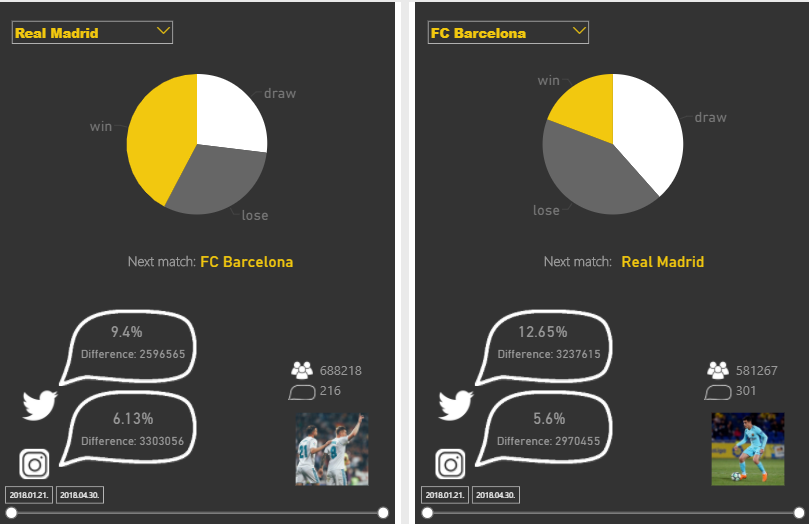
|  |  |  |  |
| --- | --- | --- | --- |
| team | facebook\_likes | twitter\_followers | instagram\_followers |
| 1. FC Köln | 754034 | 590424 | 152611 |
| 1. FSV Mainz 05 | 441916 | 171919 | 44867 |
| 1899 Hoffenheim | 274598 | 176032 | 68820 |
| AC Milan | 24827964 | 6160093 | 3854248 |
| ACF Fiorentina | 2045951 | 587158 | 227757 |
| AFC Bournemouth | 354723 | 332210 | 170530 |
| AS Monaco | 5004969 | 1928366 | 760077 |
| AS Nancy | 143797 | 80138 | 20667 |
| AS Roma | 8821951 | 1506869 | 1223651 |
| AS Saint-Étienne | 818632 | 651933 | 101138 |
| Angers SCO | 171420 | 85065 | 28166 |
| Arsenal FC | 37947647 | 12279656 | 226296 |
| Atalanta | 200054 | 250742 | 75315 |
| Athletic Bilbao | 53273 | 13100 | 166179 |
| Atlético Madrid | 13746278 | 45547 | 3865185 |
| Bayer Leverkusen | 2700359 | 398553 | 309958 |
| Bayern München | 43597794 | 4146627 | 11259013 |
| Bologna FC | 546748 | 139608 | 55488 |
| Bor. Mönchengladbach | 980938 | 451140 | 216438 |
| Borussia Dortmund | 15409612 | 3069987 | 4723298 |
| Burnley FC | 11602 | 298029 | 122449 |
| CA Osasuna | 249374 | 207467 | 31445 |
| CD Alavés | 110939 | 167100 | 44765 |
| CD Leganés | 52375 | 204076 | 30268 |
| Cagliari Calcio | 337003 | 312354 | 117049 |
| Celta Vigo | 474104 | 363307 | 104299 |
| Chelsea FC | 47864196 | 10841090 | 10474711 |
| Chievo Verona | 121099 | 239155 | 32695 |
| Crystal Palace | 1060249 | 640507 | 221745 |
| Deportivo La Coruña | 399589 | 309797 | 73641 |
| Dijon FCO | 128352 | 74328 | 18772 |
| EA Guingamp | 226021 | 236980 | 22206 |
| Eintracht Frankfurt | 619026 | 412687 | 122597 |
| Empoli FC | 104395 | 127623 | 34187 |
| Espanyol Barcelona | 399678 | 334849 | 78459 |
| Everton FC | 3042846 | 1381271 | 510329 |
| FC Augsburg | 247432 | 184989 | 60247 |
| FC Barcelona | 103521788 | 25527780 | 53079727 |
| FC Crotone | 121867 | 18396 | 40691 |
| FC Ingolstadt 04 | 86026 | 72160 | 27278 |
| FC Lorient | 241772 | 246674 | 38502 |
| FC Metz | 201665 | 133142 | 28968 |
| FC Nantes | 348541 | 442596 | 75153 |
| FC Schalke 04 | 2897519 | 659242 | 406423 |
| Genoa CFC | 283516 | 287724 | 79835 |
| Girondins Bordeaux | 791754 | 316017 | 53075 |
| Granada CF | 356038 | 322071 | 42576 |
| Hamburger SV | 53701 | 658923 | 168524 |
| Hertha BSC | 359482 | 314663 | 91540 |
| Hull City | 1014946 | 392137 | 105652 |
| Inter | 16879743 | 1481759 | 1669263 |
| Juventus | 30333370 | 810341 | 8368040 |
| Lazio Roma | 787633 | 509451 | 217296 |
| Leicester City | 6618060 | 1079759 | 1865348 |
| Lille OSC | 748551 | 613193 | 84041 |
| Liverpool FC | 5058874 | 9066224 | 5038804 |
| Manchester City | 27856572 | 5279780 | 5674329 |
| Manchester United | 73779232 | 15852334 | 19729490 |
| Middlesbrough FC | 443069 | 261664 | 65881 |
| Montpellier HSC | 295575 | 304104 | 43711 |
| Málaga CF | 1057873 | 617882 | 73224 |
| OGC Nice | 441887 | 359447 | 122225 |
| Olympique Lyon | 2723691 | 1401101 | 490632 |
| Olympique Marseille | 4816164 | 2662901 | 62104 |
| Paris Saint-Germain | 32207564 | 5772182 | 10488852 |
| Pescara Calcio | 130641 | 49975 | 30252 |
| RB Leipzig | 344037 | 139145 | 112064 |
| Real Betis | 833292 | 518127 | 223907 |
| Real Madrid | 105975730 | 27568120 | 53804831 |
| Real Sociedad | 1197989 | 16891 | 82258 |
| SC Bastia | 289236 | 295024 | 28242 |
| SC Freiburg | 238481 | 240198 | 32915 |
| SD Eibar | 70531 | 169743 | 33298 |
| SM Caen | 162562 | 176034 | 28292 |
| SSC Napoli | 4004165 | 1261523 | 722957 |
| SV Darmstadt 98 | 158874 | 90215 | 42030 |
| Sampdoria | 296429 | 310218 | 85848 |
| Sassuolo Calcio | 255992 | 239937 | 78193 |
| Sevilla FC | 1652075 | 894158 | 336239 |
| Southampton FC | 1728262 | 888212 | 313374 |
| Sporting Gijón | 286333 | 43536 | 38722 |
| Stade Rennais | 434213 | 301415 | 56994 |
| Stoke City | 1304226 | 873809 | 284775 |
| Sunderland AFC | 1698777 | 800637 | 176227 |
| Swansea City | 1348861 | 850533 | 204508 |
| Torino FC | 71798 | 21804 | 50186 |
| Tottenham Hotspur | 8618244 | 2504752 | 1525344 |
| Toulouse FC | 356198 | 443214 | 45385 |
| UD Las Palmas | 229500 | 240227 | 93150 |
| US Palermo | 324196 | 181747 | 60742 |
| Udinese Calcio | 459724 | 264258 | 63625 |
| Valencia CF | 3334130 | 1053559 | 366087 |
| VfL Wolfsburg | 1118994 | 266195 | 251145 |
| Villarreal CF | 880413 | 377320 | 105750 |
| Watford FC | 392107 | 394577 | 173197 |
| Werder Bremen | 1002100 | 20864 | 143871 |
| West Bromwich Albion | 796628 | 764095 | 165393 |
| West Ham United | 2297916 | 1330110 | 474809 |
| Santos | 3556643 | 2598717 | 522496 |
| Fenerbahce | 9717237 | 6301733 | 2760795 |
| Galatasaray | 12978703 | 273306 | 3936375 |
| AL Ahly SC | 11434090 | 2432527 | 2884457 |
| Besiktas | 6053268 | 3302602 | 1710888 |
| New York Red Bulls | 1016787 | 193904 | 111516 |
| Los Angeles Galaxy | 1848454 | 423337 | 453795 |
| Boca Juniors | 8300064 | 3097118 | 1971370 |
| River Plate | 8554728 | 2589838 | 1542978 |

## Az elkészült dashboard-ok

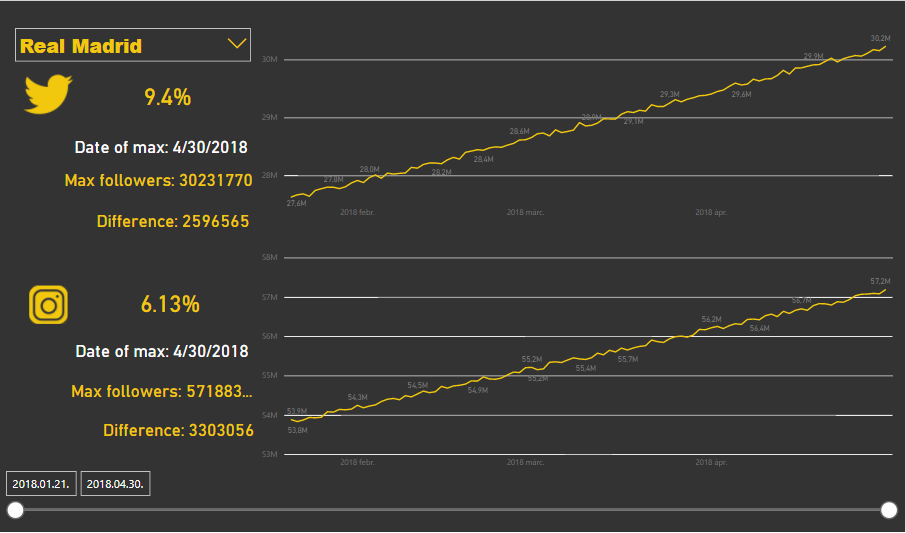
### Általános áttekintő felület



### Összehasonlító felület



### Időtartományban kereső felület



# Irodalomjegyzék

1. <http://www.digitalhungary.hu/kozossegi-media/Mennyi-idot-toltunk-el-a-kozossegi-mediaval/1722/>, [Hozzáférés: 2018. március]

1. <http://pandas.pydata.org/pandas-docs/version/0.16.2/>, [Hozzáférés: 2018. március]

1. <https://docs.djangoproject.com/en/1.11/ref/utils/> [Hozzáférés: 2018. március]

1. <https://developers.facebook.com/docs/graph-api> [Hozzáférés: 2018. március]

1. <https://developers.facebook.com/docs/graph-api/using-graph-api/> [Hozzáférés: 2018. április]
2. <https://www.theguardian.com/technology/2017/nov/08/twitter-to-roll-out-280-character-tweets-to-everyone> [Hozzáférés: 2018. április]

1. <https://dev.twitter.com/oauth> [Hozzáférés: 2018. április]

1. <https://github.com/bear/python-twitter> [Hozzáférés: 2018. április]

1. <https://dev.twitter.com/overview/api/users> [Hozzáférés: 2018. április]

1. <https://help.twitter.com/en/using-twitter/url-shortener> [Hozzáférés: 2018. április]

1. <http://mashable.com/2017/09/25/instagram-800-million> [Hozzáférés: 2018. április]

1. <https://github.com/facebookarchive/python-instagram> [Hozzáférés: 2018. április]

1. <https://docs.microsoft.com/hu-hu/power-bi/desktop-quickstart-learn-dax-basics> [Hozzáférés: 2018. április]

1. <https://appsource.microsoft.com/en-us/marketplace/apps?product=power-bi-visuals> [Hozzáférés: 2018. április]

1. <https://github.com/laszlogabor/Football-results/tree/master/data> [Hozzáférés: 2018. április]

1. <https://www.facebook.com/> [↑](#footnote-ref-1)
2. <https://twitter.com/> [↑](#footnote-ref-2)
3. <https://www.instagram.com/> [↑](#footnote-ref-3)
4. <https://app.moqups.com/> [↑](#footnote-ref-4)