

# **OTDK-dolgozat**

**2023**

Budapesti Corvinus Egyetem

**SZERZŐI ÉRTÉKELÉSEK A GOODREADS FELHASZNÁLÓK  
KÖRÉBEN**

**AUTHOR REVIEWS FROM GOODREADS USERS**

**JAKÓ FANNI**

**VIDA TAMÁS**

Konzulens: Drabancz Áron, El-Meouch Nedim Márton

Kézirat lezárásának dátuma: 2022. 04. 19.

## ABSZTRAKT

Dolgozatunkban a Goodreads könyvmegosztó- és értékelő oldalt vizsgáljuk, annak egy 6539 elemet tartalmazó mintáján. Kutatásunk középpontjában a szerzői értékelések és az irodalmi jelentőség kapcsolata áll. Arra a kérdésre keressük a választ, hogy a szerző értékelése csupán a könyveinek szórakoztató értéke alapján alakul ki vagy az irodalmi érték is szerepet játszik abban, valamint felhívjuk a figyelmet ezen hatások könyvpiacra gyakorolt hatásaira.

Véleményünk szerint már a könyvek 1-5 skálán történő értékelése is egy túlzott egyszerűsítése a kérdéskörnek, amely csak tovább erősödik azáltal, hogy a szerzőt külön nem lehet értékelni, az ő átlagos értékelése automatikusan számolódik a könyveire adott értékelésekből. Véleményünk szerint egy könyv esetében meg kell különböztetni a szórakoztató és az irodalmi faktort, melyek közül az első a könyv értékelésében kell megjelenjen, míg a második a szerzőnél. Hipotézisünk szerint a honlapot használó felhasználók számára a szórakoztatási faktor a jelentősebb, melynek egyik oka lehet a korosztályi eloszlás a felhasználók körében, valamint amelynek a következménye, hogy a magasan értékelt és sokszor ajánlott könyvek ezen tulajdonságot reprezentálják.

A kérdéseket a hangulatelemzés, a k-középpontú klaszterezés, főkomponenselemzés, valamint a hálózatelemzés módszertanával vizsgáljuk. Eredményeink alapján a jelentősebb díjak nem növelik szignifikánsan a művek értékelését, sem azok ajánlási gyakoriságait. A legmagasabbra értékelt könyvekben és a díjazott művekben nem tér el jelentősen a leggyakrabban használt szóállomány. A szerzők többsége tizedannyi könyvespolcra helyezett könyvvel rendelkezik, mint a kiemelkedően népszerű versenytársaik. Az ajánlási rendszer nem csupán az átlagos értékelési változók értékei alapján dolgozik, figyelembe veszi az értékelési számokat leíró változókat is. A szerzői értékelés ily módon történő kialakítása fals eredményeket ad, mivel az legnagyobb mértékben egy mű értékelése által adott. Az ajánlási hálózat kis sűrűségű, kisvilág-tulajdonságú, az ajánlási rendszer nem random, amögött valamilyen algoritmus húzódik, mely nem reprezentálja felül a díjazott műveket.

**Kulcsszavak:** Goodreads, főkomponenselemzés, klaszterezés, hangulatelemzés, hálózatelemzés

**JEL klasszifikáció:** C55, C38, D83

## TARTALOMJEGYZÉK

|        |                                                     |    |
|--------|-----------------------------------------------------|----|
| I.     | BEVEZETÉS (Jakó Fanni, Vida Tamás).....             | 1  |
| II.    | SZAKIRODALOM (Jakó Fanni, Vida Tamás) .....         | 3  |
| II.1.  | GOODREADS.....                                      | 3  |
| II.2.  | MÓDSZERTAN A SZAKIRODALOMBAN.....                   | 6  |
| III.   | MÓDSZERTAN (Jakó Fanni, Vida Tamás).....            | 8  |
| III.1. | HÁLÓZATI ALAPFOGALMAK .....                         | 8  |
| III.2. | HANGULATELEMZÉS .....                               | 9  |
| III.3. | K-KÖZÉPPONTÚ KLASZTEREZÉS .....                     | 9  |
| III.4. | FŐKOMPONENSELEMZÉS.....                             | 10 |
| IV.    | ADATBÁZIS KIALAKÍTÁSA (Jakó Fanni, Vida Tamás)..... | 12 |
| V.     | EREDMÉNYEK (Jakó Fanni, Vida Tamás).....            | 15 |
| V.1.   | VARIANCIAANALÍZIS .....                             | 15 |
| V.2.   | KLASZTEREZÉS .....                                  | 16 |
| V.3.   | FŐKOMPONENSELEMZÉS .....                            | 17 |
| V.4.   | AJÁNLÁSI RENDSZER, MINT HÁLÓZAT .....               | 19 |
| VI.    | ÖSSZEFOGLALÁS (Jakó Fanni, Vida Tamás) .....        | 23 |
|        | IRODALOMJEGYZÉK.....                                | 25 |

## ÁBRAJEGYZÉK

|                                                                                                                                         |    |
|-----------------------------------------------------------------------------------------------------------------------------------------|----|
| 1. ábra: Optimális klaszterszám meghatározása. Forrás: Saját szerkesztés .....                                                          | 10 |
| 2. ábra: Leggyakrabban előforduló szavak a díjazott művekben, valamint a legjobban értékelt könyvekben. Forrás: Saját szerkesztés ..... | 16 |
| 3. ábra: Kialakított klaszterek a főkomponensek terében. Forrás: Saját szerkesztés .....                                                | 17 |
| 4. ábra: Egyes változók a főkomponensek terében. Forrás: Saját szerkesztés .....                                                        | 18 |
| 5. ábra: Ajánlási gráf a vizsgált mintán. Forrás: Saját szerkesztés.....                                                                | 19 |
| 6. ábra: Az ajánlási gráf legnagyobb összefüggő komponense. Forrás: Saját szerkesztés .....                                             | 20 |

## I. BEVEZETÉS

*„The right book in the right hands at the right time can change the world.”*

Ezt az idézetet olvashatjuk a Goodreads felületén, amely a világ legnagyobb olyan weboldala, valamint applikációja, melyeken könyveket böngészhetünk, értékelhetjük, és ajánlhatjuk azokat (Goodreads, 2022). Dolgozatunkban ezen honlapot vizsgáljuk abból a szempontból, hogy a szerzői értékelések megfelelően kalibráltak-e, valamint hogy az ajánlási rendszer milyen hatásokat erősít.

A dolgozat célja, hogy megvizsgálja, a rendszer mennyire értékeli komplexen egy-egy mű és a szerző értékelését, valamint hogy ez összhangban áll-e az olvasói értékeléssel és gondolkozásmóddal. A két dolog bizonyos szempontból szorosan köthető egymáshoz, hiszen egy szerzőt a művei alapján ítélünk meg, azonban egy mű szórakoztató értéke és a szerző irodalmi minősítése között különbség van. Lehet egy könyv nagyon szórakoztató, de gyengén megírt, valamint nehezen befogadható, de irodalmilag jelentős.

A platform sajátosságaiból adódóan a szerzőt két számszerű értékkel jellemzi – a könyveinek értékeléseiből számított súlyozott számtani átlaggal, valamint azon felhasználók számával, akik követik az adott író. Véleményünk szerint már a könyvek 1-5 skálán történő értékelése is egy túlzott egyszerűsítése a kérdéskörnek, amely csak tovább erősödik azáltal, hogy a szerzőt külön nem lehet értékelni, az ő átlagos értékelése automatikusan számolódik a könyveire adott értékelésekből. Véleményünk szerint egy könyv esetében meg kell különböztetni a szórakoztató és az irodalmi faktort, melyek közül az első a könyv értékelésében kell megjelenjen, míg a második a szerzőnél.

Az alkalmazás a közösségi megosztás élménye köré épül: virtuális könyvespolcokat hozhatunk létre, barátok olvasási szokásait követhetjük, valamint ajánlásokat kaphatunk arra vonatkozóan, hogy mely könyvet olvassuk el az eddigi elolvasott könyveink, azok értékelése és a barátok tevékenysége alapján. Ezen javaslatok lehetőséget teremtenek arra, hogy az ajánlási rendszer alkotta hálózatot vizsgáljuk, mely a dolgozat egyik központi kérdésköre. Ennek segítségével keressük a választ arra a kérdésre, hogy ezen rendszer milyen szerepet játszik a szórakoztató faktor felülreprezentálásában.

Hipotézisünk szerint a honlapot használó felhasználók körében a szórakoztatási faktor a jelentősebb, melynek egyik oka lehet a felhasználók korosztályi eloszlása, valamint amelynek

a következménye, hogy a magasan értékelt és sokszor ajánlott könyvek ezen tulajdonságot reprezentálják.

Kutatási kérdéseink a következők: A jelentősebb díjak elnyerése hatással van-e a mű értékelésére, valamint annak ajánlási gyakoriságára? Mely mutatók különböztetik meg a legnépszerűbb szerzőket a többségtől? Mely szerzői teljesítményre vonatkozó mutatók írják le ugyanazon információt? A felület ajánlási rendszere milyen változók figyelembevételével dolgozik? A szerzői értékelés honlapon megvalósított módja megfelelően reprezentálja-e a szerzők közötti különbségeket?

A kérdések megválaszolásához egy 6539 elemű mintán vizsgáljuk meg a honlapon található könyveket és a hozzájuk kapcsolódó szerzőket. Ennek során az általánosan elérhető adatokon túl (pl.: átlagos értékelés, értékelésszám), valamint a szerzői adatokon túl (pl.: átlagos értékelés, értékelésszám, követőszám) figyelembe vesszünk olyan változókat is, melyek a könyv absztraktjának szövegezésére, az ajánlások számára, valamint a kapott díjak számára és minőségére vonatkoznak. Ezek segítségével vizsgáljuk, hogy a könyv – és ezáltal a szerző – magas értékelése a szórakoztató faktorból vagy az irodalmi értékből származik-e.

A kérdéseket az ajánlási hálózat alapvető hálózati mutatóinak vizsgálatával, az absztraktok hangulatelemzésével, valamint a szerzők adatainak k-középpontú klaszterezésével és főkomponenselemzésével vizsgáljuk.

A dolgozat felépítése a következő: a II. fejezetben bemutatjuk a Goodreads felületét, valamint a szakirodalomban található kutatási eredményeket az alkalmazással és a módszertanokkal kapcsolatosan. Ezt követően a III. fejezetben bemutatjuk a felhasznált módszertanokat, majd a IV. fejezetben ismertetjük az adatbázis kialakítását. Az V. fejezetben bemutatjuk a kapott eredményeket, melyeket az VI. fejezetben összegzünk.

## II. SZAKIRODALOM

Az alábbi fejezetben bemutatjuk a Goodreads felületével kapcsolatos eddigi kutatási eredményeket, valamint a hangulatelemzés és a hálózati modellek alkalmazási területeit.

### II.1. GOODREADS

A Goodreads 2007 januárjában kezdte meg a működését, jelenleg az Amazon tulajdonában áll (Goodreads, 2022), 2019 júliusában pedig már 90 millió regisztrált felhasználóval rendelkezett (Statista, 2021). A felületet weboldalként és telefonos applikációként is használhatjuk. Lehetőségünk van rögzíteni a könyveket, melyeket már elolvastunk, és amik felkeltették az érdeklődésünket. Az egyes könyveket 1-5 skálán értékelhetjük, de szöveges értékelést is írhatunk róluk. Az oldal egy közösségi platformként is működik, barátnak jelölhetünk más felhasználókat, mellyel a fiatalabb generációkat igyekeznek megszólítani. Az olvasott könyvek értékelése, a barátnak jelölt felhasználók korábbi aktivitása, illetve az olvasásra jelölt könyvek alapján ajánl új könyveket a felhasználóknak a honlap. Emellett minden mű oldalán általános ajánlást is ad a rendszer a hasonló művekre.

A Goodreads weboldalon az egyes könyveket egy 1-től 5-ig terjedő skálán pontozhatjuk, a szerzők pontszámát pedig a könyveinek átlagos értékeléséből automatikusan, felhasználói beavatkozás nélkül alakítják ki. Problémaként merülhet fel, hogy egy-egy könyvet és szerzőt nem lehet csupán egyetlen mutatószám alapján értékelni. A könyvek pontszáma például nem veszi figyelembe az elnyert díjak számát és a felhasználók szöveges értékeléseinek sajátosságait, a szerző pontszámában sem tükröződik a követőinek száma és a műveinek irodalmi értéke.

A Goodreads felhasználók viselkedésével és a szöveges értékelések elemzésével kapcsolatban már több kutatás született. Az eddigi kutatások egyik fő kérdése az volt, hogy összefügg-e egy könyv idézettsége a Goodreads oldalon megfigyelhető népszerűségével. A weboldalon megjelenő mutatószámokat – például az értékelések számát és az átlagos értékelést – a Book Citation Index (BKCI) adatbázis alapján, a Google Book keresésekkel és a Wikipédián szereplő hivatkozások számával vetették össze. A korrelációs kapcsolat minden esetben közepesen gyenge volt, főleg 0,2 és 0,3 közötti korrelációs együtthatókat figyeltek meg. Az elemzést műfajonként külön-külön elvégezve arra az eredményre jutottak, hogy egyes kategóriákban (Arts, Humanities, Social Sciences) nagyobb eséllyel lehet meghatározni a könyv egyéb aspektusait: a kulturális és az oktatási értékét (Kousha, et al., 2017). Azt is megfigyelték, hogy általában nem a (BKCI szerint) sokat idézett könyvek a legnépszerűbbek a Goodreads-en, sőt jellemzően alacsonyabb az átlagos értékelésük ezeknek a könyveknek. A szöveges értékelések

vizsgálatával arra a következtetésre jutottak, hogy ennek az lehet az oka, hogy a sokat hivatkozott könyvek tudományosabbak, és nehezebben befogadhatóak egy átlagos olvasó számára (Wang, et al., 2019). A weboldalon megtalálható történelmi témájú könyvek idézettségét nemcsak a BKCI adatbázis alapján, hanem informálisabb forrásokat is hozzáadva, altmetrikák segítségével határozták meg. A művek átlagos értékelésével azonban itt is csak gyenge pozitív irányú korrelációt állapítottak meg, a korrelációs együttható értéke 0,212 (Zuccala, 2015). A kutatásokat figyelembe véve tehát az idézettség nem befolyásolja azt, hogy mennyire népszerű egy könyv a Goodreads applikáción, az olvasók nem értékelik magasabbra a tudományos szempontból jelentős műveket. Ezek alapján a tudományos műveket író szerzők értékelése alacsonyabb lesz, mint amennyi abban az esetben lenne, ha a tudományos ismeretterjesztés területén nyújtott munkásságukat is értékelni lehetne közvetlenül a szerző értékelése által.

A felhasználók által írt szöveges értékeléseket nyelvi szempontból is többen vizsgálták. A kutatások egy részében manuálisan, egyesével olvasták el a kiválasztott mintában szereplő szövegeket. A vizsgált értékelések 86,1%-a az olvasási élményre reflektált, 68%-uk pedig kimondottan egy érzelmi reakciót említett meg. A felhasználók az ajánlások során gyakran említettek meg más szerzőket és olvasókat, de karaktereket is (Driscoll & Rehberg Sedo, 2019). Ez is mutatja, hogy a felhasználók egy része nem érzi elegendő véleménykinyilvánításnak az 5-s skálán történő értékelést, az olvasás során kapott élményt és az érzelmi reakciókat nem tudják egy számértékbe sűríteni.

A kutatások egy másik részében digitálisan, egy program segítségével elemezték az értékeléseket. Az LIWC program azt mutatta ki, hogy magasabb a vizsgált szövegekre jellemző a funkciósavak (kötőszavak, névutók stb.) és a pozitív töltetű szavak aránya (Hajibayova, 2019). Más programok használata során arra az eredményre jutottak, hogy a szövegek hangvétele (a negatív vagy pozitív töltetű szavak aránya, -1 és 1 közötti mutatószámként) és az értékelések átlaga között pozitív korrelációs kapcsolat van. A legtöbb töltéssel rendelkező szó pedig már a szöveg első mondatában megjelenik (Wang, et al., 2019). Az szöveges értékelések egy másfajta összetételét is vizsgálták, amely során az ajánlások jellegét állapították meg: tudományos ajánlás, személyes vélemény, rövid vagy hosszabb összefoglaló a könyv tartalmáról (Kousha, et al., 2017). A szöveges értékelések vizsgálatára azért van szükség, hogy tisztább képet kapjunk arról, hogy mit gondolnak a felhasználók, ne egy mutatószámba sűrítsék gondolataikat.

Az eredmények alapján elmondható, hogy a könyv értékelésének 1-5 skálán történő értékelése sok esetben túlzóan egyszerűsítő, mely nem tudja mind a szórakoztató, mind az érzelmi, mind

az irodalmi faktort megragadni. A felhasználók ezért a szöveges értékelés részben tudják kifejeteni részletesebben is véleményüket, amely azonban nem kerül bele sem a könyv értékelésének kalkulációjába, és ezáltal a szerző értékelésében sem jelenik meg. Kutatásunkban mi ehhez hasonlóan a szerzők értékelésénél fogjuk figyelembe venni a mű irodalmi értékét, amelyhez az absztrakt szövegét használjuk forrásul.

A Goodreads oldalon a kifejezetten sokat olvasó felhasználók úgynevezett könyvtárosok lehetnek. Több kérdés irányult a könyvtáros – átlagos felhasználó, illetve a férfi – női viselkedés közötti különbségek feltárására is. A könyvtáros csoport tagjai aktívabbak, több könyvet olvasnak és értékelnek, viszont az átlagos értékelésük alacsonyabb, mint a többi felhasználóé (Wang, et al., 2019). A férfiak és nők értékeléseinek számában és átlagában nincs jelentős eltérés, viszont érdekesség, hogy a férfiak és nők is jellemzően az azonos nemű szerzők könyveit értékelik magasabbra (Thelwall, 2019). A nemek közti viselkedés különbségeinek meghatározásának korlátja, hogy a privát felhasználókról nem derülnek ki személyes információk, a nyilvános felhasználók többsége, körülbelül háromnegyede pedig nő (Thelwall & Kousha, 2017). A kutatásokból kiderül, hogy a több könyvet olvasó könyvtárosok általában kritikusabbak a művek pontozásánál, azaz az átlagos felhasználók hajlamosabbak túlértékelni azokat. Ennek is az lehet az oka, hogy egy átlagos olvasó a könyvek szórakoztató szerepét helyezi előtérbe az irodalmi értéke helyett, ami a szerzők értékelésénél is megjelenik, és torzíthatja a szerző pontszámát. A többet olvasó és ezáltal könyvtárossá váló felhasználók viszont jobban össze tudják hasonlítani a valós értéket a művekben, így referenciaként szolgálhatnak. Mivel azonban a könyvtárosok értékelése nincs súlyozva az értékelések számítása során, ezért ezek a hatások eltörpülnek a mű és a szerző értékelésében is.

2012-ben a Goodreads applikációt összekapcsolták a Facebook felületével, amely következtében megnőtt a „barátok” funkciót használók száma. Thelwall és Kousha (2017) azt vizsgálták, hogy az olvasott könyvek száma összefügg-e a felhasználó barátainak számával. Eredményeik alapján a felhasználók egy része gyakran csak az egyik funkciót használja. Arra a megállapításra jutottak, hogy a Goodreads elsődleges céljaként se a könyvek rögzítése, se a szociális kapcsolatok kiépítése nem azonosítható, a weboldal teljesen hibrid.

A Goodreads értékelések vizsgálatát arra is használták, hogy megpróbálják előre jelezni a sikeres könyveket. Az egyik modell 88,72%-os pontossággal tudta megállapítani, hogy egy hónap múlva mely könyvek fognak az Amazon bestsellerek közé tartozni (Maity, et al., 2017). Egy másik kutatás keretrendszer biztosít arra, hogy a Goodreads könyvek adatai segítségével prediktálják, hogy a weboldal „Readers Choice Awards” eseményén hány szavazatot fognak kapni az egyes könyvek. A modell jól magyaráz, a korrelációs együttható 0,61, és a szórás

viszonylag kicsi. A modellben fontos szerepet játszanak a felhasználók viselkedései (pl. kapott értékelések átlaga, az elolvasott könyvek sokszínűsége) és a szerzők presztízse is (Maity, et al., 2018). A kutatások tehát nagy pontossággal jelezték előre, hogy mely könyvek lesznek népszerűek, ami utal a Goodreads felhasználók viselkedésére: a populáris, szórakoztató könyveket helyezik előtérbe.

A kutatások során nem vizsgálták azt, hogy a felhasználók hogyan értékelik a művek irodalmi értékét és szórakoztató szerepét, ezért esett választásunk ezen kérdéskör vizsgálatára.

## **II.2. MÓDSZERTAN A SZAKIRODALOMBAN**

Egyre több tudományág próbálkozik hálózatok segítségével megérteni a komplex rendszereket. A hálózatok csúcsokból és őket összekötő élekből álló rendszerek. Attól függően, hogy számít-e a csúcsok közötti kapcsolat iránya, az élek lehetnek irányítottak vagy irányítatlanok. A gráf lehet összefüggő, de előfordulhat, hogy több komponensre bomlik szét, ahol két komponens között nem fut él. A hálózatok komponenseinek jellemzőiből és a gráfok különböző csúcsokra vonatkozó (lokális) és globális mutatószámai alapján következtethetünk a tulajdonságaira. Ilyenek a csúcsok központiságára (például csúcsok fokszáma) és az élek sűrűségére (például klaszterezési együttható) utaló mutatószámok.

A hálózatok alkalmazásának egyik jellemző tudományága a biológia. Több kutatás született arról, hogy különböző hálózatelméleti módszerek segítségével próbálják megérteni az emberi agy működését. Ennek egyik kiindulópontja a kisvilág-tulajdonságú hálózat, amelynek jellemzője, hogy annak ellenére, hogy az egyes csúcsoknak nem feltétlenül van közvetlen szomszédja, két-két csúcs közötti átlagos távolság mégis kicsi (Watts & Strogatz, 1998). A kisvilág-tulajdonság és a többi hálózatokhoz kapcsolódó mutatószám elemzésével lehetőség nyílt az agyi működések vizsgálatára (Telesford, et al., 2011). Az agy vizsgálatát továbbfejlesztve hálózatok segítségével a különböző agyi rendellenességeket is vizsgálni lehet, mint például az Alzheimer-kórt, a traumás agysérülést és az epilepsziát (Stam, 2014).

A hálózatelemzés nemcsak a biológiában, hanem más területeken is megjelent. Készült olyan kutatás, amely a városi közlekedési hálózatok optimális kialakításával foglalkozik (Ding, et al., 2019), egy Elba szigetével foglalkozó esettanulmány keretein belül pedig a turizmust is hálózatok segítségével vizsgálták (Baggio, 2017). Hálózatok elemzésével közgazdaságtudományi és pénzügyi témák vonatkozásában is foglalkoztak. A módszer segítségével lehetőség van nemzetközi pénzügyi hálózatok elemzésére, amely a globális pénzügyi stabilitás megteremtésében is fontos szerepet játszhat (Oatley, et al., 2013).

A Goodreads applikáció az egyes könyvekhez a többi felhasználó viselkedése és pontozása alapján ajánl új, hasonló könyveket. Ezek a könyvek egy hálózatot alkotnak. Kutatásunkban a gráfokkal kapcsolatos mutatószámok segítségével vizsgáljuk meg az adatbázisunkban szereplő könyvek hálózatának jellemzőit.

A hangulatelemzés (sentiment analysis) szövegek elemzésével foglalkozó tudományterület, amely segítségével nagy mennyiségű szöveg hangulatát, érzelmi töltetét és jellemzőit lehet megbecsülni számítógépes programok segítségével. Több kutatás foglalkozik a hangulatelemzés eszközeinek vizsgálatával és összevetésével. Léteznek lexikonalapú elemzések, amelyek emberi beavatkozást igényelnek egy kezdeti szótár létrehozása által. A gépi tanulási módszerek megjelenésével és térnyerésével a szövegek hangulatának vizsgálata gyakran teljesen automatikusan történik manapság, például a naiv Bayes-modell által (Medhat, et al., 2014).

A hangulatelemzés tudományterületén belül többféle szöveget elemeztek a különböző módszerek segítségével. Egyik kutatási irány a közösségi médiában megjelenő posztok és kommentek vizsgálata. A különböző módszerek felhasználásával Twitter posztok hangulatát is vizsgálták. Az elemzésben a hangulatelemzés további lehetőségeire is felhívták a figyelmet: például ennek segítségével ki lehet szűrni azokat a termékeket és szolgáltatásokat, amelyekkel a vásárlók nem elégedettek, és meg lehet figyelni különböző trendeket az egyes piacokon (Kharde & Sonawane, 2016).

Kutatásunkban a könyvek absztraktjainak hangulatát vizsgáljuk meg pozitív és negatív jelentésű szavak elemzésével, ez alapján becsüljük meg a könyv hangvételt és irodalmi értékét. Arra szeretnénk választ kapni, hogy hogyan befolyásolja egy könyv és a szerzőjének értékelését a mű hangulata, van-e különbség a pozitív és negatív témájú könyvek megítélésében.

### III. MÓDSZERTAN

Az alábbi fejezetben bemutatjuk az általunk használt módszertanokat: a hálózatelemzési alapfogalmakat, a hangulatelemzést, a k-középpontú klaszterezést, valamint a főkomponenselemzést. A hálózatelemzést a könyvek ajánlási hálózatára, a hangulatelemzést az absztraktok vizsgálatára fogjuk alkalmazni. A k-középpontú klaszterezést, valamint a főkomponenselemzést a szerzői változók közötti kapcsolat vizsgálatokhoz fogjuk felhasználni.

#### III.1. HÁLÓZATI ALAPFOGALMAK

A hálózatok csúcsokból, és csúcsokat összekötő irányított vagy irányítatlan élekből álló rendszerek. Egy komponensbe tartoznak azok a csúcsok, amelyek között létezik legalább egy út. Két csúcs távolsága az a legrövidebb út, amely során az egyik csúcsból a másikba juthatunk.

A gráfok jellemzőinek bemutatásához az alábbi jelöléseket és mutatószámokat használjuk:

$N$  a hálózat csúcsainak száma,  $L$  pedig a hálózatban szereplő élek száma

$k_i$  az  $i$ -edik csúcshoz tartozó fokszám, mely megadja, az adott csúcsból hány él indul

Az irányítatlan gráfok átlagos fokszáma:  $\overline{k_u} = \frac{2L}{N}$ , az irányítottaké pedig  $\overline{k_d} = \frac{L}{N}$ .

A hálózat egy komponensében két csúcs átlagos távolsága:  $\bar{d} = \frac{2}{N(N-1)} \sum_{i \neq j} d_{i,j}$

A gráf átmérője  $\delta = \max(d_{i,j})$ , azaz a két legtávolabbi csúcs távolsága.

A hálózat sűrűsége az élek és az összes lehetséges él hányadosa:  $P = \frac{2L}{N(N-1)}$ .

Azt, hogy az egyes csúcsok mennyire töltnek be központi szerepet a hálózatban, a különböző centralitás mutatókkal jellemezhetjük. Ezek a csúcsok fokszáma, a közelség centralitás (az adott csúcsból induló legrövidebb utak hosszának reciprokösszege), a közöttség centralitás (az adott csúcs a két-két csúcsot összekötő legrövidebb útvonalak hányadrésében szerepel) és a sajátvektor központiság (a közvetlen szomszédoktól függő közelségi mérték). Ezen mutatók csúcsokra vonatkozó mutatók, így a teljes rendszer vizsgálatokhoz ezekre jelen vizsgálat során nem térünk ki.

A csúcsokra jellemző lokális klaszterezési együttható azt vizsgálja meg, hogy a csúcs közvetlen szomszédjai mekkora arányban vannak összekötve egymással. A hálózatot az átlagos lokális klaszterezési együttható mellett a globális klaszterezési együtthatóval is jellemezhetjük, amely a gráfban szereplő háromszögek és az összes kiválasztható összefüggő csúcshármas arányát mutatja meg (Katy, et al., 2007).

## **III.2. HANGULATELEMZÉS**

A hangulatelemzés célja, hogy a nemstrukturált formában lévő adatokat, mint például híreket vagy esetünkben absztraktokat strukturált formában leírja. A szövegek hangulatelemzése által számszerűsített értéket kaphatunk arról, hogy az adott szövegnek milyen a tartalma.

A hangulatelemzés egyik legkorábbi formája a szavakhoz való számérték rendelése, mely egyszerűsége miatt a mai napig széles körben alkalmazott. A módszer használatához a szövegeket szavakra bontjuk, majd minden szó esetében eltávolítjuk a toldalékokat. A szótővezett szóhoz ezt követően hozzárendeljük valamely hangulati szótár (pl.: AFINN, Bing, Loughran, NRC) által adott hangulati értékét. Az adott szöveget leíró számot ezen szavakhoz rendelt hangulati értékek átlagaként határozzuk meg.

Az egyes szótárak különböző típusú szövegek vizsgálatára készültek. Az AFINN szótár -5 és 5 közötti értékekkel jellemzi az adott szó pozitív vagy negatív voltát (Nielsen, 2011), míg a legegyszerűbb felépítésű Bing szótár pozitív, negatív jelzőket ad az egyes szavaknak (Hu & Liu, 2004). A Loughron szótár elsődleges célja a pénzügyi szavak kategorizálása (Loughran & McDonald, 2011), míg az NRC szótár a félelem – megnyugvás érzelmi tengelyen kategorizálja a szavakat (harag, várakozás, undor, félelem, öröm, negatív, pozitív, szomorúság, meglepetés, bizalom) (Mohammad & Tuney, 2013). Az egyes szótárak a felismert szavak számában is eltérnek, 2 500 – 13 900 szóig terjedve, viszont többségük felülreprezentálja a negatív kifejezéseket. Ezek mellett több külön célra létrehozott szótár is létezik, mint például az IMDb leírások szövegeinek kategorizálására (Stine, 2018).

Céljainkhoz az AFINN szótár által megvalósított számszerűsítés áll a legközelebb, így az absztraktok elemzésére ezen szótárt alkalmazzuk. Ennek fő előnye, hogy egy 11 fokú skálán értékeli a szavakat a kategorizálás helyett, így lehetővé téve a szöveg átlagos értékének számítását. Kutatások alapján a pozitív és a semleges szavak detektálásában ezen hangulati szótár a legpontosabb (Islam & Zibran, 2017). A szótár készítésének célja eredetileg a mikroblogok szövegeinek elemzése volt (Nielsen, 2011), mely nagyban hasonlít a könyvek bemutatóinak irodalmi stílusához: röviden és tömören írva kell az olvasó érdeklődését felkelteni.

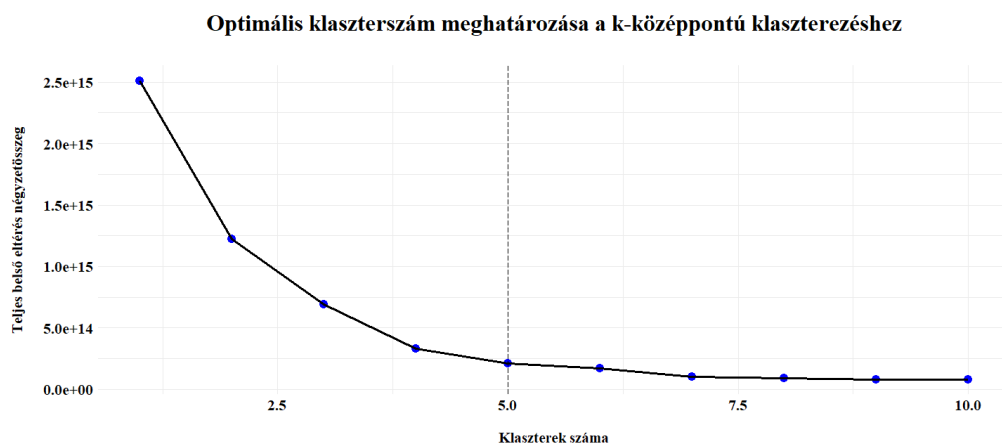
## **III.3. K-KÖZÉPPONTÚ KLASZTEREZÉS**

A K-középpontú klaszterezés célja, hogy az adatokat K darab csoportba sorolja úgy, hogy a csoporton belül az adatok minél homogénebbek legyenek. A módszer a tanító nélküli tanulás csoportjába sorolható, melynek jelentése, hogy nem használ fel külső változót, melynek értékét

prediktálni kívánja. A módszer csupán az adatokat rendezi csoportokba úgy, hogy a csoportokon belül az adatok eltérése minél kisebb legyen.

Az algoritmus random kiválasztott klaszterközéppontokkal indul. Minden adatpont hozzárendelésre kerül a hozzá euklideszi távolságban legközelebb lévő középponthoz. Ezt követően az egyes klaszterekhez hozzárendelt adatok mértani középpontja felülírja a klaszterközéppontot. Az algoritmus újra lefut mindaddig, míg a klaszterközéppontok változásának mértéke az egyes lépések során nem változik.

Az így kialakított K darab klaszter jellemezhető a klaszterközépponttól mért távolság négyzetösszegével. K növelésével ez az érték folyamatosan csökken, viszont egy bizonyos érték felett a klaszterszám növelése nem ad információ többletet. Az optimális klaszterszám meghatározása az úgynevezett „könyök módszerrel” történik, mely a K klaszterszám függvényében ábrázolja a belső eltérési négyzetösszeget és itt keresi az a pontot, ahol a hozzáadott 1 klaszter már nem csökkenti jelentősen ezt az értéket (Thorndike, 1953).



1. ábra: Optimális klaszterszám meghatározása. Forrás: Saját szerkesztés

Az 1. ábrán látható a „könyök módszer” alkalmazása a Goodreads adatbázisra, mely alapján a szerzők 5 klaszterbe sorolhatóak.

Az egyes klaszterek ezt követően jellemezhetőek lesznek a középpontjuk értékével, a klaszterbe tartozó elemek számával. A módszer előnye, hogy nem él előzetes hipotézisekkel az adatok tulajdonságairól, nem elvárás az adatok lineáris függetlensége annak alkalmazásához.

### III.4. FŐKOMPONENSELEMZÉS

A főkomponenselemzésnek több célja lehet, az adatok kevesebb dimenzióba való sűrítése, az adatok egyszerűsítése, valamint az egyes változók közötti kapcsolat kimutatása (Abdi & Williams, 2010).

A módszertan alkalmazása előtt a Kaiser-Meyer-Olkin és a Bartlett tesztet kell elvégezni. A Kaiser-Meyer-Olkin index a korrelációs mátrix felhasználásával számítható és megmutatja, hogy az egyes változók mekkora mértékben magyarázhatóak a változók kimaradó részével. Az index 0 és 1 közötti értéket vehet fel. A nagyobb értékek az adatok nagyobb tömörítési lehetőségét mutatják, ezáltal indukálva a dimenziócsökkentésre való nagyobb alkalmazhatóságot (Kaiser, 1970). A szakirodalom a 0,5-nél magasabb értékek esetén ajánlja a dimenziócsökkentési eljárások alkalmazását a vizsgált adattáblára (Kaiser & Rice, 1974).

A Bartlett teszt nullhipotézise, hogy a korrelációs mátrix megegyezik az identitás mátrixszal, az adatok „gömbölyűek”, azaz az adathalmaz nem használható dimenziócsökkentésre (Bartlett, 1951).

A főkomponenselemzés egymással korreláló numerikus változók felhasználásával végezhető el. A módszer célja, hogy az eredeti változószámnál kevesebb egymásra merőleges változóba, úgynevezett komponensbe sűrítse az adatokat úgy, hogy az adathalmaz varianciájának (azaz a benne tárolt információmennyiségnek) minél nagyobb hányadát megőrizze. Az információsűrítésnek köszönhetően a változók és a megfigyelések közötti hasonlóságok könnyebben felismerhetőek (Wold, et al., 1987).

A komponensek meghatározását a módszer a korrelációs mátrix sajátértékeinek és sajátvektorainak segítségével határozza meg. A készített új, egymásra ortogonális komponensek több változó által leírt közös adatot reprezentálják, ezért látens változóknak nevezzük őket. A látens változók egymással már nem korrelálnak, az új dimenziók mentén a variancia monoton csökken. Az információtömörítés fokát, azaz a készítendő látens változók számát a magyarázott variancia arányában szokás meghatározni, melyet az adott dimenzióhoz tartozó sajátérték reprezentál. Abban az esetben szokás a komponenszt a modellbe vonni, ha annak sajátértéke 1-nél nagyobb, azaz a magyarázott variancia a 10%-t meghaladja. Ezen hozzájárulások kumulált értékei adják a főkomponenselemzés általános magyarázóerejét a variancia terén, mely 50% felett már elfogadható (Wold, et al., 1987).

#### IV. ADATBÁZIS KIALAKÍTÁSA

Az alábbi fejezetben bemutatjuk az általunk használt adatbázisokat, azok kialakítását, tisztítását, valamint általános jellemzőit.

A Goodreads weblapok letöltését a *BeautifulSoup* nevű Python package segítségével végeztük el, mely a honlap linkje alapján tölti le annak tartalmát, melyből aztán kinyerhetőek a szükséges információk. A letöltést 2022. 03. 17-én végeztük el. Az adatok elemzését ezt követően az R programcsomag használatával készítettük el.

A letöltésre használt könyvek listáját a *Best Books of the 21st Century* 9405 elemű lista alapján alakítottuk ki. Ennek a listának vettük egy 6543 elemű mintáját (~67%), amelyeknél a letöltés nem ütközött hálózati akadályokba. A kiválasztott elemek a lista rangsorában egyenletes módon oszlanak el, így mintavételi hibákból adódó torzításokra különösebb módon nem kell ügyelni.

Minden kiválasztott könyv esetében letöltöttük a címét, a szerzőjét, az átlagos értékelését, az értékeléseinek, valamint a szöveges értékeléseinek számát, a kiadási éveit, az eredeti nyelvét, a kapott díjainak listáját, az absztraktját, a könyvsorozathoz való tartozását, hogy mely könyveket ajánlja, hogy hány ember jelölte, hogy jelenleg olvassa, illetve, hogy szeretné a későbbiekben elolvasni. Emellett a szerzők honlapjáról letöltöttük a követők számát, az értékelését, az értékelések és a szöveges értékelések számát, valamint a szerző által írt művek számát.

Ezt követően az absztrakt és a díjak változókból képeztünk új változókat is. Az absztrakt esetében a hangulatelemzéssel kapott (AFINN szótár) átlagos hangulati értéket alakítottuk ki. A díjak változóból képeztünk egy összes díj száma változót, valamint dummy változókat arra vonatkozóan, hogy az egyes kiemelt irodalmi díjakat elnyerte-e az adott mű.

Számos irodalmi díjat osztanak ki minden évben, amellyel a szerzők és könyveik elismertté válhatnak. Nem volt elérhető olyan szakirodalom, amely kifejezetten azzal foglalkozna, hogy mely irodalmi díjak a legmeghatározóbbak, így könyvekkel foglalkozó internetes források segítségével az alábbi irodalmi díjakat választottuk ki: *Pulitzer Prize*, *Booker Prize*, *Edgar Awards*, *National Book Critics Circle Award*, *National Book Awards*, *Costa Book Award*, *Women's Prize for Fiction*, *PEN America Literary Awards*, *British Book Awards*.

A Goodreads 2009-től minden évben lehetőséget ad a felhasználóknak, hogy minden kategóriában megszavazzák a kedvenc könyveiket, a legjobbak pedig a *Goodreads Choice Awards* díjban részesülnek.

Ahogy a legtöbb vizsgált változó, úgy a szerzők követőinek száma is balra ferde eloszlást követ. A mintában a legtöbb, 488 908 követővel Colleen Hoover rendelkezik, ezzel szemben a szerzők felének 466-nál kevesebb, 75%-uknak pedig 3464-nál kevesebb követőjük van.

A listán szerepel 6 olyan könyv is, amit senki nem pontozott eddig, pontszámuk 0, mégis a 21. század legjobb könyvei közé kerültek, ráadásul nem is a lista végén szerepeltek, mely utal a felületen található ajánlási listák pontatlan kivitelezésére. Ezeket a könyveket eltávolítottuk a mintából, és a maradék 6539 adattal dolgoztunk tovább.

A legrosszabbra értékelt könyv 1,89-es, a legjobb pedig maximális, 5-ös értékeléssel rendelkezett. A könyvek az átlagos értékelés alapján azonban egy szűk skálán mozognak: a könyvek felének az átlagos pontszáma 3,76 és 4,18 közé esik, de az interdecilis terjedelem értéke is csak 0,83. Maximális pontszámmal 117 darab könyv rendelkezik, de ezeket maximum 13 ember értékelte. Az értékelők alacsony száma miatt ezen pontszámok fals képet adnak, azonban ezen értékeket nem szűrtük ki az adatbázisból, hiszen pont a pontozási rendszer gyengeségeire kívánjuk felhívni a figyelmet.

A két értékelés közötti különbség esetében, tehát a könyv értékelése – a szerző értékelése változó esetében az adatok 50%-a esik a  $[-0,09; 0,09]$  intervallumba, míg a 90%-a  $[-0,35; 0,35]$  intervallum eleme. Abban az esetben, ha minden szerző esetében csak a legtöbbet értékelt könyv segítségével vizsgáljuk ezt a különbséget, az adatok 90%-a  $[-0,33; 0,27]$  intervallumba esik, míg az adatok fele a  $[-0,08; 0,07]$  eleme. Ezen értékek is mutatják, hogy bár a szerzők értékelése az összes könyvük értékelésének súlyozott átlaga alapján alakul, az eltérő értékelésszám miatt ezt 0,08-os pontossággal a legtöbbször értékelt, azaz a legolvasottabb könyv értékelése adja.

Ezen eredmény tovább erősíti a kritikánkat a szerzői átlagos értékelés információtartalma ellen, hiszen a szerző ilyen típusú megítélését a legtöbb ember által olvasott, azaz megítélésünk szerint a legnagyobb szórakoztató faktorról rendelkező könyv adja.

A könyvek értékeléseinek száma is jobbra elnyúló eloszlást követ. A medián értéke 3 140, csupán a könyvek fele kapott ennél több pontozást. A harmadik kvartilis értéke már 18 420, de van olyan könyv is, amit több mint 5 millióan értékelték (Twilight). Egy könyv átlagosan 2 073 szöveges értékelést kapott, mely változó mediánja 301.

Az adatbázisunk 1990 és 2021 között kiadott könyveket tartalmaz, a könyvek felét 2007 és 2013 között adták ki. A könyvek közül 5 915 angol nyelvű, 252 pedig más nyelven íródott. A vizsgált mintában nagy arányú azoknak a könyveknek a száma, amelyek egy sorozatban szerepelnek, az összes könyv 32,8%-a ilyen.

Az adatbázisban szereplő könyveknek a 65%-a egyetlen díjban sem részesült, de van 47 könyv, amely legalább 10 díjat nyert. A legtöbb, 29 díjat John Green *The Fault in Our Stars* című műve kapta. A nevesebb díjak közül a „*National Book Critics Circle Award*” címben részesült a legtöbb könyv, 124 db, ezzel szemben a „*Goodreads Choice Awards*” díjat 631 könyv is megkapta.

A legfelső 10%-ra értékelt könyvek között 35 *Goodreads Choice Awards* díjas könyv szerepel, míg az összes többi díj közül csak a *Booker Prize* reprezentált a mintában. A díjazott művek alacsony száma miatt 1-1 díj esetében ez a véletlen műve is lehetne, azonban 8 díj esetében erre már kis esélyt látunk.

A kiemelt díjak közül összesen egyet 316, összesen kettőt 70, míg összesen hármat 11 mű kapott, mely összességében is jelentősen elmarad a Goodreads által adott díjak számától. Ez is mutatja a vizsgált díjak minőségét.

A vizsgált könyvek 4 280 különböző szerzőhöz köthetőek, így kialakítottunk egy szerzői adatbázist is, mely az eddig már bemutatott szerzői adatok mellett tartalmazza a könyvek átlagos érzelmi értékét, a könyvekre mutató ajánlások, valamint a díjak számát, a kiemelt díjak közül minimum egy elnyerését, a könyvek átlagos, valamint maximális értékelését, illetve hogy a szerző írt-e könyvsorozatot, valamint hogy hány ember tette a könyvespolcra (épp olvasott vagy olvasni kívánt könyvek közé) a szerző műveit. A klaszterezést, a főkomponenselemzést ezen adatbázison végeztük el.

A könyvek esetében letöltött ajánlások segítségével elkészítettünk egy hálózati adatbázist is, melyben a csúcsokat azon vizsgált könyvek adták, melyek ajánlottak legalább egy olyan könyvet, melyet vizsgáltunk, vagy őket ajánlotta legalább egy ilyen mű oldala. Így alakult ki egy 3 530 csúcsú gráfunk, melyben az élek az ajánlásokat mutatják. A hálózati mutatókat ezen adatbázison vizsgáltuk.

## V. EREDMÉNYEK

Az alábbi fejezetben bemutatjuk a varianciaanalízis, a hangulatelemzés, a klaszterezés, a főkomponenselemzés, valamint a hálózati elemzés során kapott eredményeinket.

### V.1. VARIANCIAANALÍZIS

Varianciaanalízis segítségével azt vizsgáltuk, hogy az egyes kategóriaváltozók által kialakított csoportokon belül a könyvek átlagos értékelése 5%-os szignifikanciaszinten eltér-e egymástól. A tesztet csak abban az esetben végeztük el, ha a Levene-teszt (Levene, 1960) alapján a csoportokon belül a variancia egyezőségének hipotézisét nem lehet elvetni.

Ennek során arra az eredményre jutottunk, hogy a *Pulitzer Award*, a *British Book Awards* és a *National Book Awards* díj elnyerése sem növeli szignifikánsan a könyv értékelését, míg a *Women's Prize for Fiction* díj elnyerése szignifikánsan csökkent, a *Goodreads Choice Awards* elnyerése szignifikánsan növeli azt. Ezen eredmények tovább erősítik azt a megfigyelést, hogy a könyvek értékelései nem az irodalmi érték alapján alakulnak ki. A Goodreads díj eltérő eredménye következhet abból is, hogy a felhasználók jelölése és szavazatai alapján ítélik oda a díjat, mely egybevág azzal, hogy az adott könyvet mennyire értékelték. Ezáltal az okság megfordul, a magasabb értékelés miatt kerül odaítélésre a díj, nem pedig fordítva.

Szintén megvizsgáltuk az ajánlási gyakoriság esetében is az ezen változók szerinti csoportokat. Ebben az esetben sem találtunk szignifikáns pozitív hatásokat, melyeket a jelentősebb díjak indukálnának.

Összegezve sem az adott mű ajánlásainak számában, sem az átlagos értékelésében nem tükröződik az, hogy az adott mű kapott-e díjat a kiemelt jelentősebb irodalmi elismerések közül. Ez arra enged következtetni, hogy nem csupán az értékelések, valamint az ebből kalkulált szerzői értékelések, de a rendszer ajánlási rendszere sem tudja kiemelni az irodalmi értéket, a honlap csupán a szórakoztatási faktor tekintetében tud reprezentatív ajánlásokat tenni.

Az AFINN szótár segítségével készített hangulati értékek alapján végzett varianciaanalízis az *Edgar Award* esetében szignifikánsan alacsonyabb hangulati értéket mutatott ki a díjazott művek esetében, mint a nem díjazottak csoportjában. Ez a hangulati megbontás a *National Book Critics Circle Award* esetében fordított, ezen díj a pozitívabb hangvétellű műveket díjazza.

A könyv értékelése alapján készített kategóriák alapján a 4,5 feletti értékelésű könyvek átlagos érzelmi világa 0,32, míg az ez alatti művek hangulati értéke 0,18, mely eltérés 5%-on szignifikáns.

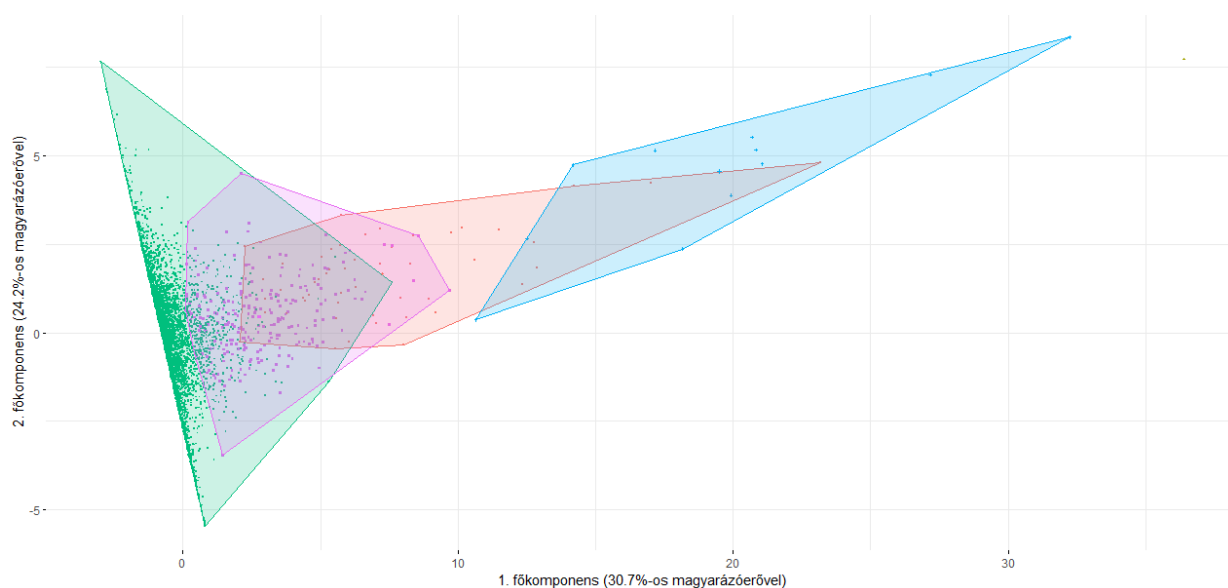
## V.2. KLASZTEREZÉS

Az 5 klaszterbe 236, 50, 3969, 11, valamint 1 elem került. Ezek alapján van 1 jelentősen eltérő szerző, aki nem meglepő módon *J. K. Rowling*. Az ő ajánlásainak, szerzői értékeléseinek, valamint szöveges értékeléseinek száma oly módon meghaladja az összes többi szerzőnél tapasztalt értéket, mely semelyik másik szerzőhöz nem hasonlítható.

16

A legnagyobb elemszámú csoport semleges érzelmi világgal, 0,8 ajánlással, 1,19 díjjal, 3,95-ös szerzői és átlagos könyvekre vonatkozó értékeléssel, 50 000-nél kevesebb értékeléssel, 5 000-nél kevesebb szöveges értékeléssel, 1 300 követővel és tizedannyi könyvespolcra helyezéssel jellemezhető, mint a többi klaszter szerzői.

Ezek alapján a szerzők a felületen a könyvespolcra helyezett művek számával, a követőszámmal és az ajánlások számával tudnak kiemelkedni a tömegből. Emiatt fontos kérdés, hogy az ajánlások algoritmusai alapján határozza meg az ajánlott műveket, valamint emiatt kritika, hogy a szerzői értékelés miért nem korrigálja a könyvek átlagos értékét a követőszámmal és a díjak számával.



3. ábra: Kialakított klaszterek a főkomponensek terében. Forrás: Saját szerkesztés

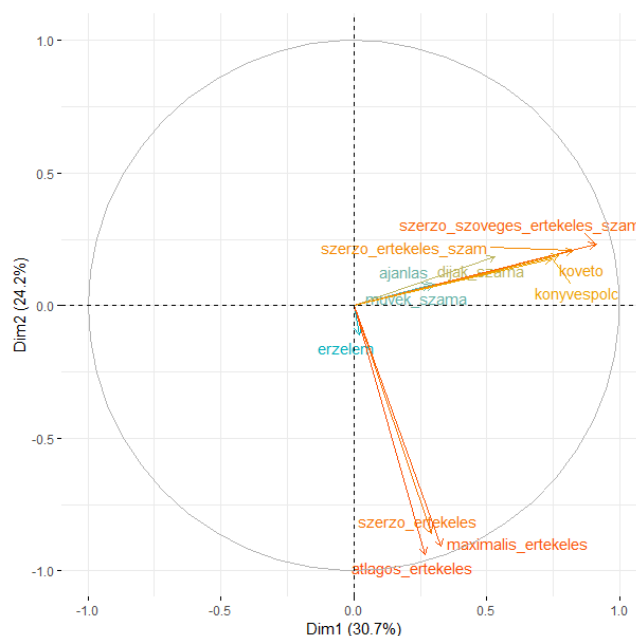
### V.3. FŐKOMPONENSELEMZÉS

A Kaiser-Meyer-Olkin érték 0,71, mely meghaladja a 0,5-es kritikus értéket, így az adatbázis használható főkomponenselemzésre, a vizsgált változók (melyek megegyeznek az előző részben vizsgáltakkal) nem függetlenek egymástól. A Bartlett teszt alapján az adatok nem gömbölyűek, van értelme a főkomponenselemzésnek, lehet az adatokat sűríteni. A sajátértékek közül 2 volt 1 felett, így 2 főkomponenst állítottunk elő, mellyel közel 51%-os varianciamegőrzést értünk el. A komponens mátrix alapján a két komponens az alábbiak szerint alakult ki, mely a 4. ábrán is látható:

1. a szerző díjainak száma, a felhasználók által későbbi olvasásra jelölések száma, a szerző követőinek száma, a szerző értékeléseinek és szöveges értékeléseinek száma
2. a szerző értékelése, a szerző művének legmagasabb értékelése, a szerző műveinek átlagos értékelése

Ezek mellett az alábbi változók játszottak kis szerepet a főkomponensek kialakításában, azaz ezen változók térnek el a legnagyobb mértékben a többiektől, ők tartalmaznak plusz információt, melyet az egyes főkomponensekbe nem lehetett sűríteni: az átlagos érzelmvilága a szerzői absztraktoknak, szerző műveinek száma, a szerző művei ajánlásainak száma.

Ez alapján az ajánlási rendszer nem csupán ezen változók értékei alapján dolgozik, azonban nagyobb szerepet játszik a kialakításában az első főkomponens, mint a második.



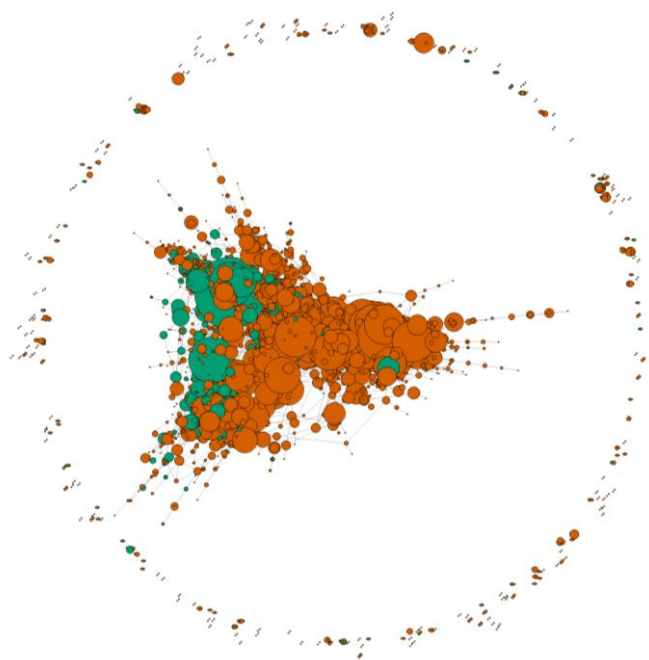
4. ábra: Egyes változók a főkomponensek terében. Forrás: Saját szerkesztés

Az első látens változó, mely a teljes információmennyiség 30,7%-át magyarázza, azon változók alapján került kialakításra, melyet véleményünk szerint reprezentálni kellene a szerzők értékelésében: a díjak, a könyvespolcra helyezések, a követők és az értékelések száma. Ezen változók a főkomponenselemzés alapján egymáshoz hasonlóak, emiatt elegendő lenne egy ilyen értéket az értékelés korrigálására használni. A korrelációs mátrix értékei 0,6 körül csoportosulnak, az értékek mindegyike közepes vagy erős lineáris együttmozgást mutat ezen változók között.

A második látens változó megerősíti azt a korábban már többször leírt eredményt, hogy a szerzői értékelés ily módon történő kialakítása fals eredményeket ad, hiszen a legjobban értékelt könyv és a szerző értékelése között 0,78-as korrelációs együtthatót találunk, mely alapján a többi mű hatása a szerzői értékelésre csekély.

#### V.4. AJÁNLÁSI RENDSZER, MINT HÁLÓZAT

A könyvek ajánlási hálózata egy irányított gráfot alkot, ahol a csúcsok a könyveket, az élek pedig az ajánlásokat jelentik. Az ajánlások gyakran az adatbázisunkon kívül eső művekre mutattak, ezeket a könyveket nem ábrázoltuk ebben a gráfban. A hálózat csúcsai közé 3 530 könyv került, melyek között 6 743 él fut összesen, mely az 5. ábrán is látható. Ez azt mutatja, hogy a 6 539 könyv esetében csupán 3 530 (~54%) mutatott a mintán belüli elemre vagy mutatott rá mintabeli elem.



5. ábra: Ajánlási gráf a vizsgált mintán. Forrás: Saját szerkesztés

A gráf sűrűsége ezek alapján kicsi, az összes lehetséges élnek csak 0,05%-a szerepel a hálózatban. Ez olyan szempontból nem meglepő, hogy az ajánlási rendszer célja, hogy a milliós nagyságrendű könyvet tartalmazó honlapon segítse a felhasználót abban, hogy olvasnivalót válasszon. A felhasználási okok miatt a kimenő éleknek egy természetes felső korlátja van.

Az 5. és 6. ábrán az egyes csúcsok mérete a csúcs fokszámával arányos, míg a zöld szín jelképezi, hogy az adott mű kapott-e elismerést a kiemelt minőségű díjak közül (zöld, ha igen és sárga, ha nem). Az 5. ábrán látható, hogy a külső köríven is helyezkednek el zöld csúcsok, mely azt mutatja, hogy a minákon belül több olyan díjnyertes alkotás is volt, mely az ajánlási rendszer „szélén” helyezkedik el. Emellett néhány kiemelt díjnyertes művet is találunk a rendszerben.

A csúcsok átlagos fokszáma 1,91, azaz átlagosan ennyi él indul ki egy csúcsból, azonban a fokszámeloszlás kétmódusú, így az átlag csak tájékoztató jellegű. A gráf nem összefüggő, 250 komponensre oszlik. A legnagyobb komponens 2 814 csúcsú, a második legnagyobb pedig

csupán 18 csúcsból áll. Ezen értékek még kisebbek lennének abban az esetben, ha a kimaradt 3 009 nulla fokszámú csúcsot hozzáadnánk a gráfhoz. Az adatok mintavétele okozhatja azt, hogy a gráf ilyen kis részekre bomlik szét, emiatt a következőkben a hálózat legnagyobb komponensének mutatószámait mutatjuk be.



6. ábra: Az ajánlási gráf legnagyobb összefüggő komponense. Forrás: Saját szerkesztés

A 6. ábrán már jobban kivehető, hogy bár több olyan díjnyertes mű van, melyet a rendszer többszörösen ajánl, az ilyen mód legnépszerűbb könyvek mégsem ezen kategóriából kerülnek ki. Az ajánlási rendszer ezek alapján nem reprezentálja felül a díjazott műveket. Az ajánlási rendszer célja minden közösségi felületen, hogy megkönnyítse a felhasználók dolgát a keresésben. Legyen szó akár könyvekről, akár zenéről, akár ismerősökről vagy posztokról, a mai világban nincs arra lehetőség, hogy a napi szinten ezres, tízezres, milliós nagyságrendekben keletkező újdonságok közül saját magunk válasszunk, így az ilyen típusú ajánlási rendszerek szerepe és ezáltal a felelőssége is megnő. A Goodreads felületénél maradván, amennyiben a háttérben meghúzódó algoritmus a magasabb értékelést veszi figyelembe és az irodalmi értéket nem, a rendszer hozzájárul ahhoz a folyamathoz, hogy a szerzőknek minél inkább a szórakoztatásra helyezték a nagyobb hangsúlyt, ha egy ilyen piacon sikeresek szeretnének lenni.

A hálózat legnagyobb komponense 2 814 csúcsból és 6 173 élből áll, azaz a gráf sűrűsége 0,078%, a csúcsok átlagos fokszáma pedig 2,19, ennyi él indul ki átlagosan egy csúcsból. Ez azt jelenti, hogy egy olyan képzeletbeli felhasználó, aki ezen könyvek bármelyikéből elindulva az ajánlások mentén halad, el fog jutni a másik 2 813 könyv mindegyikéhez. Ezek alapján ezen 2 814 mű egy összetett világot alkot a Goodreads felületén.

Az átlagos lokális klaszterezési együtttható értéke 0,137, azaz egy átlagos csúcsnak két szomszédja 13,7% eséllyel van összekötve egymással is. Ez a gyakorlatban azt jelenti, hogy ha az első könyv ajánlja a másodikat és a harmadikat is, az 13,7%-ban indukálja azt, hogy a második is ajánlani fogja a harmadikat vagy fordítva. Az ajánlások számának korábban már említett volta miatt ez az érték befolyásolja azt, hogy az ajánlások alkotta hálózat több kisebb csoportra esik szét, melyek csoporton belül mutatnak egymásra vagy a rendszer egy nagy hálózatot alkot. A 13,7%-s értékből arra következtetünk, hogy a Goodreads teljes felületét megvizsgálva a második eset áll fent. A globális klaszterezési együtttható értéke 0,179, azaz az összes összefüggő csúcshármassok 17,9%-ánál mind a három él reprezentálva van.

A hálózat átmérője 34, azaz a két legtávolabbi csúcs közötti legrövidebb út 33 másik csúcsot tartalmaz. Abban az esetben, ha a hálózat egy ajánlási láncot alkotna, ezen érték 2 813 lenne, melyhez viszonyítva a 34 alacsony, azaz a hálózat kisvilág-tulajdonságú. Összehasonlításképpen az emberiség ismeretségi hálózatának átmérőjét a kutatások 4-7 közé teszik (Watts, 1999).

A legnagyobb fokszerű csúcsban 45 él található, a fokszerűségi mediánja 3, a csúcsok felében tehát legfeljebb három él található. Ezen értékek eltérése is mutatja, hogy a fokszerűeloszlás nem egyenletes, azaz az ajánlási rendszer nem random, amögött valamilyen algoritmus húzódik. A Goodreads minden könyvhöz 15-20 másik könyvet ajánl. Van olyan könyv, amely ajánlásai közül 13 az adatbázisunkon belülre mutat, illetve olyan könyv is, amelyre 44 másik könyv is hivatkozik.

Létrehoztunk egy olyan hálózatot is, amelyben minden olyan könyv szerepel, amelyre az adatbázisunkban szereplő könyvekről ajánlások mutatnak. A gráf csúcsainak száma 65 526, éleinek száma pedig 102 425. A gráf egy csúcsából átlagosan 1,56 él indul ki. Ennek oka, hogy a vizsgált művek esetében mind a 15-20 ajánlott könyv szerepel a hálózatban, viszont az általunk vizsgált művek által ajánlott könyvekből már nem mutat ki él, mely a csúcsoknak több mint 90%-ra igaz. Az átlagos 1,56 él tehát ebben az esetben sem reprezentatív.

A hálózat ebben az esetben 61 komponensre bomlik. A legnagyobb komponensben 64 638 könyv szerepel, mely az összes csúcs több mint 98%-a. Ebből arra következtetünk, hogy a leszakadó komponensek oka csupán a mintavétel, a teljes ajánlási hálózat összefüggő. A második legnagyobb komponens csúcsainak száma mindössze 37. A következőkben a hálózat legnagyobb komponensének jellemzőit mutatjuk be.

A hálózat átmérője 35, melyet összehasonlítva az előző gráffal láthatóvá válik, hogy bár a csúcsok száma jelentősen nőtt (~2300%), a hálózat átmérője csak eggyel lett nagyobb, mely egy újabb bizonyíték a hálózat kisvilág-tulajdonságára.

Az átlagos lokális klaszterezési együtttható értéke 0,064, azaz egy átlagos csúcsnak két szomszédja 6,4% eséllyel van összekötve egymással is, a globálisan vizsgálva a csúcshármasok 3,1%-ánál mind a három él reprezentálva, azaz ekkora arányban ajánlják kölcsönösen egymást. Ezen értékek tovább erősítik azt a feltételezést, hogy a teljes ajánlási rendszer összefüggő hálózatot alkot.

Az egy csúcsból kiinduló élek számának maximuma 18, ez mutatja, hogy ennyi könyvet rendel a Goodreads az egyes művekhez ajánlásként. Ezzel szemben az egy csúcsba befutó élek számának maximuma 108, van olyan könyv, amit 108 másik könyv adatlapján ajánl a weboldal, ugyanakkor a csúcsok 74%-ba legfeljebb egy él fut be.

Összegezve tehát a vizsgált mintánk alapján a Goodreads ajánlási rendszere nem random, a teljes ajánlási rendszer feltehetően összefüggő és a rendszer nem kisebb csoportok kialakítására törekszik, hanem a teljes könyvtár bejárására.

Az ajánlási rendszer ezek alapján nem reprezentálja felül a díjazott műveket, mely esetében kiemelendő a rendszer írókra gyakorolt hatása a szórakoztatás és a gondolati, valamint irodalmi tartalmak átadása közötti választásban.

## VI. ÖSSZEFOGLALÁS

Dolgozatunkban a Goodreads honlapot vizsgáljuk abból a szempontból, hogy a szerzői értékelések megfelelően kalibráltak-e, valamint hogy az ajánlási rendszer milyen hatásokat erősít. A dolgozat célja, hogy megvizsgálja, a rendszer mennyire értékeli komplexen egy-egy mű és a szerző értékelését, valamint hogy ez összhangban áll-e az olvasói értékeléssel és gondolkozásmóddal.

A könyv értékelésének vizsgálata során több anomáliát találtunk. 3 díj esetében találtuk azt, hogy a díj elnyerése nem növeli szignifikánsan a könyv értékelését, míg a *Women's Prize for Fiction* díj elnyerése szignifikánsan csökkenti, a *Goodreads Choice Awards* elnyerése szignifikánsan növeli azt. Ezen eredmények alapján a könyvek értékelései nem az irodalmi érték alapján alakulnak ki. A Goodreads díj eltérő eredménye következhet abból is, hogy a felhasználók jelölése és szavazatai alapján ítélik oda a díjat, mely egybevágh azzal, hogy az adott könyvet mennyire értékelték. Ezáltal az okság megfordul, a magasabb értékelés miatt kerül odaítélésre a díj, nem pedig fordítva. Az ajánlási gyakoriság esetében sincs kimutatható pozitív hatása a jelentős díjnak az ajánlási rendszerre. Ez arra enged következtetni, hogy nem csupán az értékelések, valamint az ebből kalkulált szerzői értékelések, de a rendszer ajánlási rendszere sem tudja kiemelni az irodalmi értéket, a honlap csupán a szórakoztatási faktor tekintetében tud reprezentatív ajánlásokat tenni.

A művek absztraktjainak vizsgálata során arra az eredményre jutottunk, hogy az *Edgar Award* esetében szignifikánsan alacsonyabb hangulati érték figyelhető meg a díjazott művek esetében, mint a nem díjazottak csoportjában. Ez a hangulati megbontás a *National Book Critics Circle Award* esetében fordított, ezen díj a pozitívabb hangvételi műveket díjazza, míg a legjobbra értékelt könyvek csoportjában szignifikánsan magasabb a könyvek hangulati értéke. Ennek ellenére a legmagasabbra értékelt könyvekben legtöbbet előforduló szavak és a díjazott művek leggyakoribb szavai között nem található jelentősebb eltérés.

A szerzői adatok 5 klaszterba sorolódnak, mely egy jelentősen eltérő szerzőt, *J. K. Rowling*-ot emeli ki a többi szerző közül az ajánlásainak, szerzői értékeléseinek, valamint szöveges értékeléseinek számával. Kiemelkedik még pozitív irányban 11 szerző a díjak össz-számában, a művek számában, a szerzői értékelések számában, de legfőképp a követőszámában és a könyvespolcra helyezett könyvek számában.

A legtöbb mű semleges érzelmi világgal, 0,8 ajánlással, 1,19 díjjal, 3,95-ös szerzői és átlagos könyvekre vonatkozó értékeléssel, 50 000-nél kevesebb értékeléssel, 5 000-nél kevesebb szöveges értékeléssel, 1 300 követővel és tizedannyi könyvespolcra helyezéssel jellemezhető,

mint a kiemelkedő szerzők. A szerzők a felületen a könyvespolcra helyezett művek számával, a követőszámmal és az ajánlások számával tudnak kiemelkedni a tömegből. Emiatt fontos kérdés, hogy az ajánlások algoritmusa mi alapján határozza meg az ajánlott műveket, valamint emiatt kritika, hogy a szerzői értékelés miért nem korrigálja a könyvek átlagos értékét a követőszámmal és a díjak számával.

A főkomponenselemzés alapján a szerzői absztraktok átlagos érzelmvilága, a szerző műveinek száma, valamint az ajánlásainak száma nem írható le az átlagos értékelések, valamint az értékelésszámok által kialakított tengelyeken, ezen változók térnek el a legnagyobb mértékben a többiektől, ők tartalmaznak plusz információt, melyet az egyes főkomponensekbe nem lehetett sűríteni. Az ajánlási rendszer nem csupán az átlagos értékelési változók értékei alapján dolgozik, figyelembe veszi az értékelési számokat leíró változókat is, azonban ezen korrekciók a honlapon a szerzőknél megjelenített értékelésnél nem jelennek meg. A főkomponenselemzés alapján a szerzői értékelés így módon történő kialakítása fals eredményeket ad, hiszen a legjobban értékelt könyv és a szerző értékelése között 0,78-as korrelációs együtthatót találunk, mely alapján a többi mű hatása a szerzői értékelésre csekély.

Az ajánlási hálózat kis sűrűségű, ezáltal segíti a felhasználók választását a milliós nagyságú könyvtárból. A hálózat kisvilág-tulajdonságú, bár az éleknek csak kis százaléka reprezentált a hálózatban a csúcsok közötti átlagos távolság alacsony. A hálózat fokszámoszlása nem egyenletes, azaz az ajánlási rendszer nem random, amögött valamilyen algoritmus húzódik. A leszakadó komponensek oka csupán a mintavétel, a teljes ajánlási hálózat összefüggő, a rendszer nem kisebb csoportok kialakítására törekszik, hanem a teljes könyvtár bejárására. Az ajánlási rendszer nem reprezentálja felül a díjazott műveket, mely esetében kiemelendő a rendszer írókra gyakorolt hatása a szórakoztatás és a gondolati, valamint irodalmi tartalmak átadása közötti választásban.

Az ajánlási rendszer célja minden közösségi felületen, hogy megkönnyítse a felhasználók dolgát a keresésben. Legyen szó akár könyvekről, akár zenéről, akár ismerősökről vagy posztokról, a mai világban nincs arra lehetőség, hogy a napi szinten ezres, tízezres, milliós nagyságrendekben keletkező újdonságok közül saját magunk válasszunk, így az ilyen típusú ajánlási rendszerek szerepe és ezáltal a felelőssége is megnő. A Goodreads felületénél maradva, amennyiben a háttérben meghúzódó algoritmus a magasabb értékelést veszi figyelembe és az irodalmi értéket nem, a rendszer hozzájárul ahhoz a folyamathoz, hogy a szerzők minél inkább a szórakoztatásra helyezték a nagyobb hangsúlyt, ha egy ilyen piacon sikeresek szeretnének lenni.

## IRODALOMJEGYZÉK

- Abdi, H. & Williams, L. J., 2010. Principal Component Analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), pp. 433-459.
- Baggio, R., 2017. Network science and tourism—the state of the art. *Tourism Review*.
- Bartlett, M. S., 1951. The effect of standardization on a Chi-square approximation in factor analysis. *Biometrika*, 38. kötet, pp. 337-344.
- Ding, R. és mtsai., 2019. Application of complex networks theory in urban traffic network researches. *Networks and Spatial Economics*, 19(4), pp. 1281-1317.
- Driscoll, B. & Rehberg Sedo, D., 2019. Faraway, so Close: Seeing the intimacy in Goodreads reviews.. *Qualitative Inquiry*, 25(3), pp. 248-259.
- Goodreads, 2022. *Goodreads*. [Online] Available at: <https://www.goodreads.com/about/us> [Hozzáférés dátuma: 2022. 04. 11].
- Hajibayova, L., 2019. Investigation of Goodreads' reviews: Kakutanied, deceived or simply honest?. *Journal of Documentation*.
- Hu, M. & Liu, B., 2004. Mining and summarizing customer reviews. In: *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining*. ACM.
- Islam, M. R. & Zibran, M. F., 2017. *A Comparison of Dictionary Building Methods for Sentiment Analysis in Software Engineering Text* ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM), IEEE.
- Kaiser, H. F., 1970. A second generation little jiffy. *Psychometrika*, 35. kötet, pp. 401-415.
- Kaiser, H. F. & Rice, J., 1974. Little jiffy, mark IV. *Educational and Psychological Measurement*, 34. kötet, pp. 111-117.
- Katy, B., Sanyal, S. & Vespignani, A., 2007. Network science. *Annual Review of Information Science & Technology*, 41(12), pp. 537-607.
- Kharde, V. A. & Sonawane, S., 2016. Sentiment analysis of twitter data: a survey of techniques. *International Journal of Computer Applications*, 139(11), pp. 5-15.
- Kousha, K., Thelwall, M. & Abdoli, M., 2017. Goodreads reviews to assess the wider impacts of books. *Journal of the Association for Information Science and Technology*, 68(8), pp. 2004-2016.

- Levene, H., 1960. Robust Tests for Equality of Variances. In: Olkin, szerk. *Contributions to Probability and Statistics*. Stanford: Stanford University Press, 25. fejezet, pp. 278-292.
- Loughran, T. & McDonald, B., 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66. kötet, pp. 35-65.
- Maity, S. K., Kumar, A. & Mullick, A., 2018. *Understanding Book Popularity on Goodreads*. Sanibel Island, FL, USA, Association for Computing Machinery.
- Maity, S. K., Panigrahi, A. & Mukherjee, A., 2017. *Book reading behavior on goodreads can predict the amazon best sellers*. Sydney, Advances in Social Networks Analysis and Mining.
- Medhat, W., Hassan, A. & Korashy, H., 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4), pp. 1093-1113.
- Mohammad, S. M. & Tuney, P. D., 2013. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3), pp. 436-465.
- Nielsen, F. A., 2011. A new ANEW: Evaluation of a world list for sentiment analysis in microblogs. In: M. Rowe, M. Stankovic, A. Dadzie & M. Hardey, szerk. *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*. Herakleion, pp. 93-98.
- Oatley, T., Winecoff, W. K., Pennock, A. & Danzman, S. B., 2013. The political economy of global finance: A network model. *Perspectives on Politics*, 11(1), pp. 133-153.
- Stam, C. J., 2014. Modern network science of neurological disorders. *Nature Reviews Neuroscience*, 15(10), pp. 683-695.
- Statista, 2021. *Statista*. [Online] Available at: <https://www.statista.com/statistics/252986/number-of-registered-members-on-goodreadscom/> [Hozzáférés dátuma: 2022. 04. 11.].
- Stine, R. A., 2018. Sentiment Analysis. *Annual Review of Statistics and its Applications*, 6(1), pp. 17.1-17.22.
- Telesford, Q. K. és mtsai., 2011. The brain as a complex system: using network science as a tool for understanding the brain. *Brain connectivity*, 1(4), pp. 295-308.
- Thelwall, M., 2019. Reader and author gender and genre in Goodreads *Journal of Librarianship and Information Science*, 51(2), pp. 403-430.
- Thelwall, M. & Kousha, K., 2017. Goodreads: A social network site for book readers.. *Journal of the Association for Information Science and Technology*, 68(4), pp. 972-983.

- Thorndike, R. L., 1953. Who Belong in the Family?. *Psychometrika*, 18(4), pp. 267-276.
- Wang, K., Liu, X. & Han, Y., 2019. Exploring Goodreads reviews for book impact assessment.. *Journal of Informetrics*, 13(3), pp. 874-886.
- Watts, D. J. & Strogatz, S. H., 1998. Collective dynamics of 'small-world' networks. *nature*, 393(6684), pp. 440-442.
- Wold, S., Esbensen, K. & Geladi, P., 1987. Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3), pp. 37-52.
- Zuccala, A. A., 2015. Altmetrics for the humanities: Comparing Goodreads reader ratings with citations to history books. *Aslib Journal of Information Management*.