Adatbányászat és gépi tanulás   
(beadandó)

# Cél meghatározása

A [Goodreads.com](https://www.goodreads.com/) könyvekról szóló oldalon text review-kat lehet írni a könyvekhez és értékelni lehet őket ott egy 1-től 5-ig terjedő skálán. Az innen nyert több mint 11 000 rekordot tartalmazó tábla fellelhető a [kaggle.com](https://www.kaggle.com/datasets/jealousleopard/goodreadsbooks/data) oldalon. Célom ezen adatbázis felhasználásával megvizsgálni, hogy az olvasmányok mennyi text review-t generáltak és mennyi az átlagos értékelésük. Egész pontosan szeretném megvizsgálni, hogy milyen kategóriák léteznek ezen a téren: Hogyan függ össze az, hogy egy könyv mennyi text review-t generál és az, hogy milyen értékelése. A kurzushoz megadott adatsorhoz képest a könyvek és értékeléseik közelebb állnak hozzám, ezért választottam más adatsort.

A feladathoz tartozó programkód és a megtisztított forrásfájlok megtalálhatók az alábbi github repozitóriumban: <https://github.com/ilyefalvi/adatbanyaszat_es_gepi_tanulas_elte.git>

# Adatok megismerése és bemutatása

Az adatbázis egy vesszővel tagolt .csv-fájl, azaz egyetlen táblából áll, ez 11127 db rekordot tartalmaz. A fájl első sorában fellelhetők a mezőnevek is, amelyek a mezőkről szólnak:

* bookID: ez egy egyedi azonosító szám (egész szám)
* title: a könyv címe (szöveg)
* authors: A könyv szerző /-jellel elválasztva (szöveg/tömb)
* **average\_rating**: A könyv értékeléseinek átlaga (0-től 5-ig terjed a skála) (tört szám)
* isbn: A könyv isbn száma (szöveg)
* isbn: A könyv 13 jegyű isbn-száma (szöveg)
* language\_code: A könyv milyen nyelven írodott (rövid szöveg)
* num\_pages: A könyv oldalszáma (egész szám)
* **ratings\_count**: A könyvre hány értékelés érkezett (egész szám)
* **text\_reviews\_count**: A könyvre hány **szöveges** értékelés érkezett (egész szám)
* publication\_date: A kiadás dátuma (dátum)
* publisher: a kiadás dátuma (szöveg)

# Adatok előkészítése, előfeldolgozása

A beolvasás során a következő hibákat észleltem, ezeket javítottam:

* A csv-fájl vesszőkkel van tagolva, mégis néhány rekordban idézőjel nélkül szerepelnek vesszők, amelyek elrontják így a formátumok. Ezek a hibák és az általam alkalmazott átírások:
  + 3350. sor: "Sam Bass Warner, Jr." -> "Jr. Sam Bass Warner"
  + 4704. sor: "David E. Smith (Turgon of TheOneRing.net, one of the founding members of this Tolkien website)/Verlyn Flieger/Turgon (=David E. Smith)" -> "David E. Smith/Verlyn Flieger/Turgon"
  + 5879. sor: "James Wesley, Rawles" -> "Rawles James Wesley"
  + 8981. sor: "James, Son & Ferguson" -> "Brown/Son/Ferguson"
* A rekordok közt néhány dátum nem létező dátumra utalt:
  + 8182.sor: "11/31/2000" -> "11/30/2000"
  + 11100.sor: "6/31/1982" -> "6/30/1982"

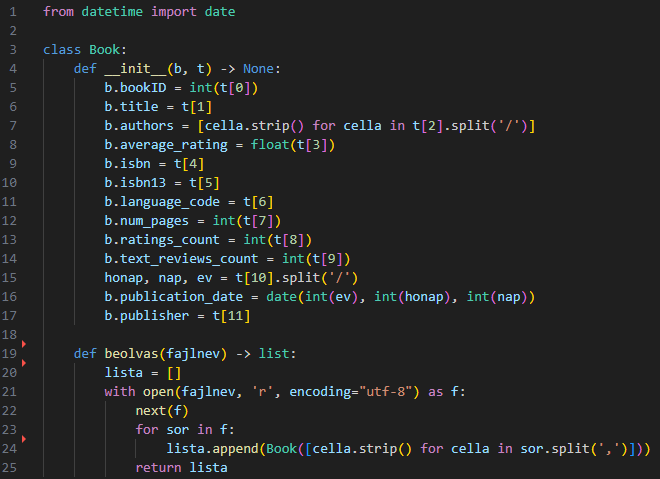
Az adatokkal ez leszámítva nem volt egyéb probléma.

# Alkalmazandó modell meghatározása

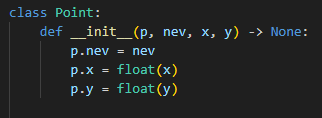
Az elemzés során unsupervised learninget, egész pontosan **K-means clustering**-et alkalmaztam a text\_review és az avereage\_rating mezők közötti összefüggés vizsgálatára. Ehhez külön függvényeket és programot írtam pythonban, az eredményt a LaTeX TikZ könyvtárának segítségével ábrázoltam. Utólag már persze látom, hogy scikit-tel gyorsabb lett volna, de így legalább gyakoroltam a programozást és jobben megértettem a modellt.

A továbbiakban a program kódja következik:

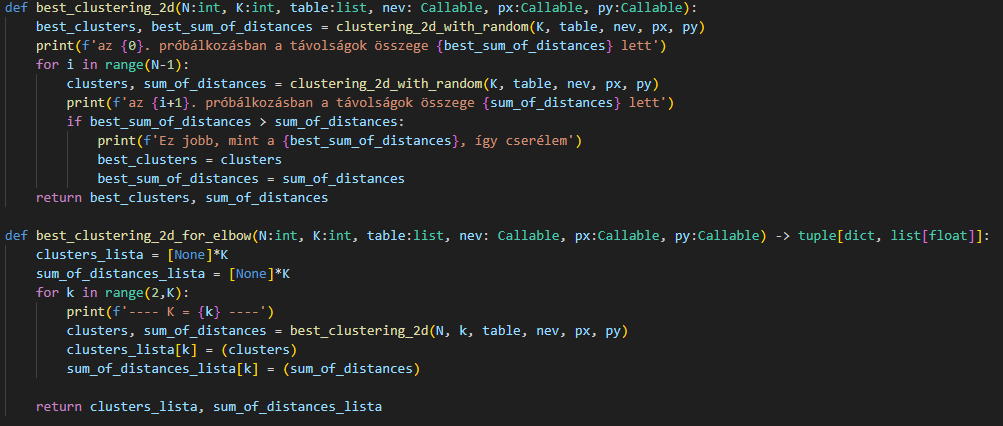
## Book class

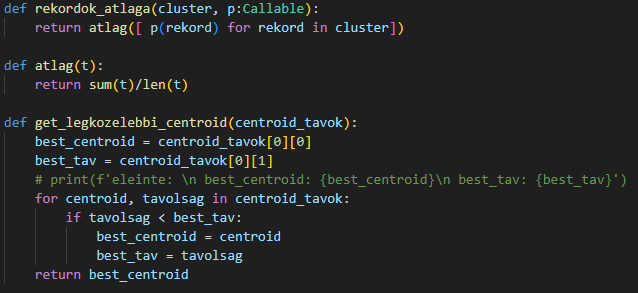
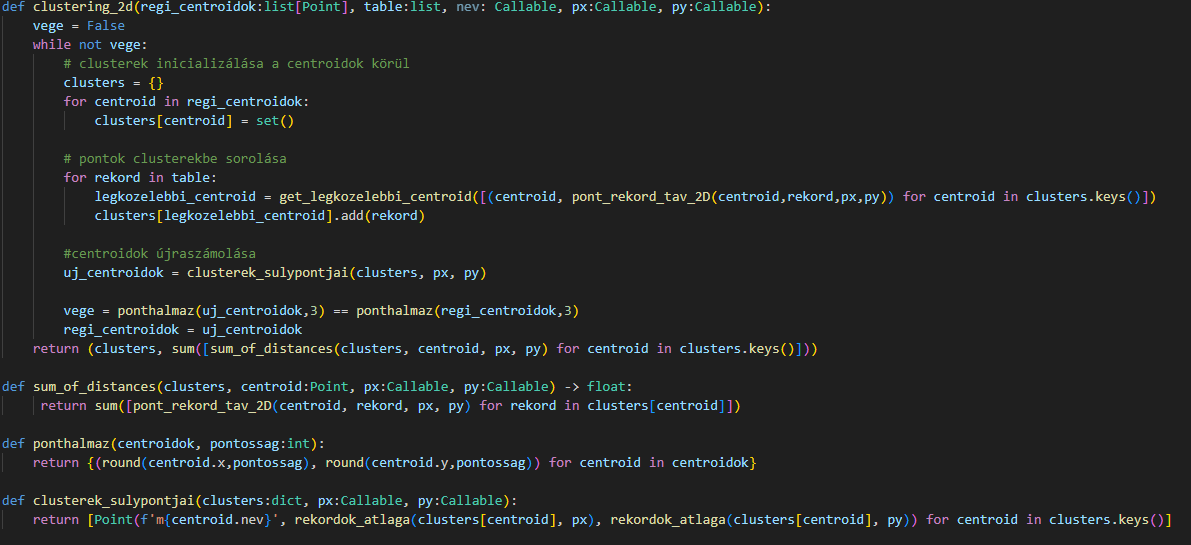


## Points class

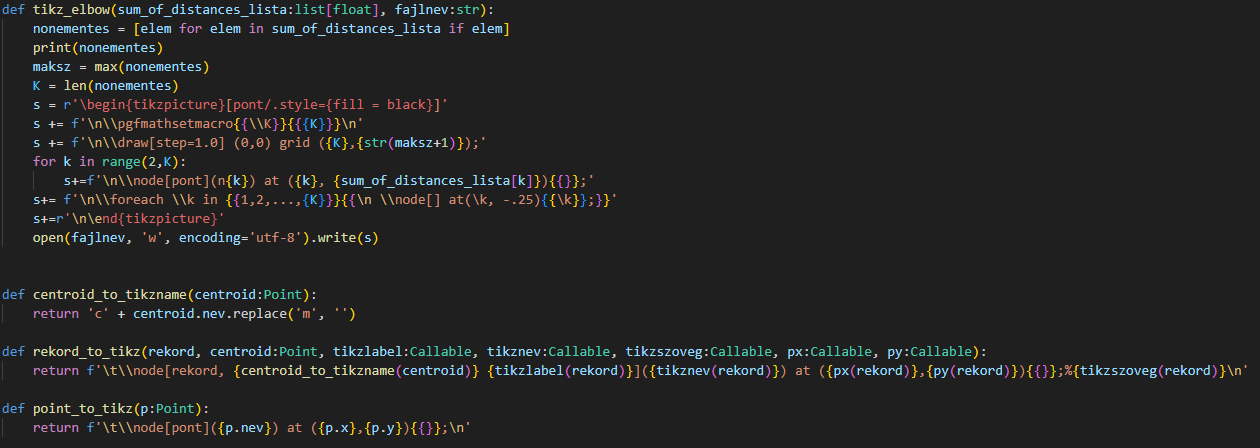


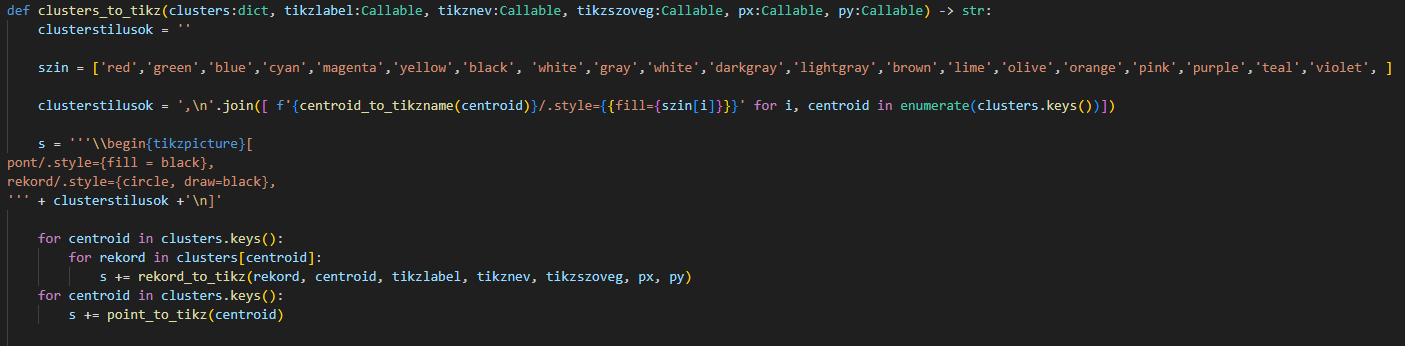
## k\_means\_clustering.py függvényei





### A TikZ-szel kapcsolatos kiíró parancsok

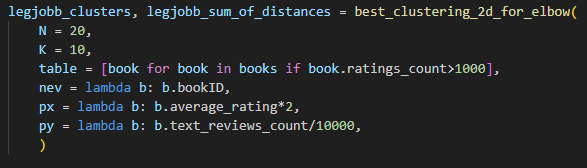


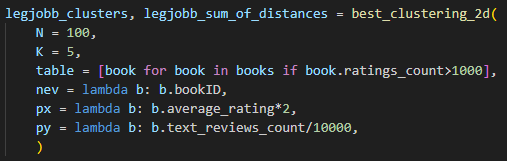


# Vezérlő paraméterek meghatározása modelben

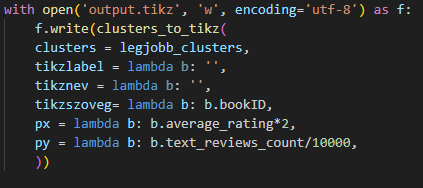
Először is leszűrtem az adatokat aszerint, hogy csak az 1000-nél több ratings\_count-tal rendelkező adatokkal dolgozzak, így végül az adatok felével dolgoztam csak, de ez is elég számottevő az adatbázis mérete miatt.

Mivel a text\_reviews\_count 0 és 100000 közé esett, az average rating pedig 0 és 5 közé, ezért az elsőt 10000-zel osztottam, az utóbbit pedig 2-vel szoroztam, hogy egy 10x10-es négyzetben ábrázolhatóak legyenek az adatok.

Lefuttattam 20-szori samplingezéssel K=2,3,4,5,6,7,8,9 alkalommal a klaszterezést:  
  
Aztán az „elbow method” során arra jutottam, hogy 5 klaszterre lenne érdemes bontani az adatokat. Itt tűnt úgy, hogy 5-nél több klaszter már nem csökkenti tovább olyan nagyon szignifikánsan a pontok centroidokhoz való közelségösszegeit. Itt aztán N=100-szor teszteltem és elmentettem a legkisebb távolságösszegű klaszterezést:

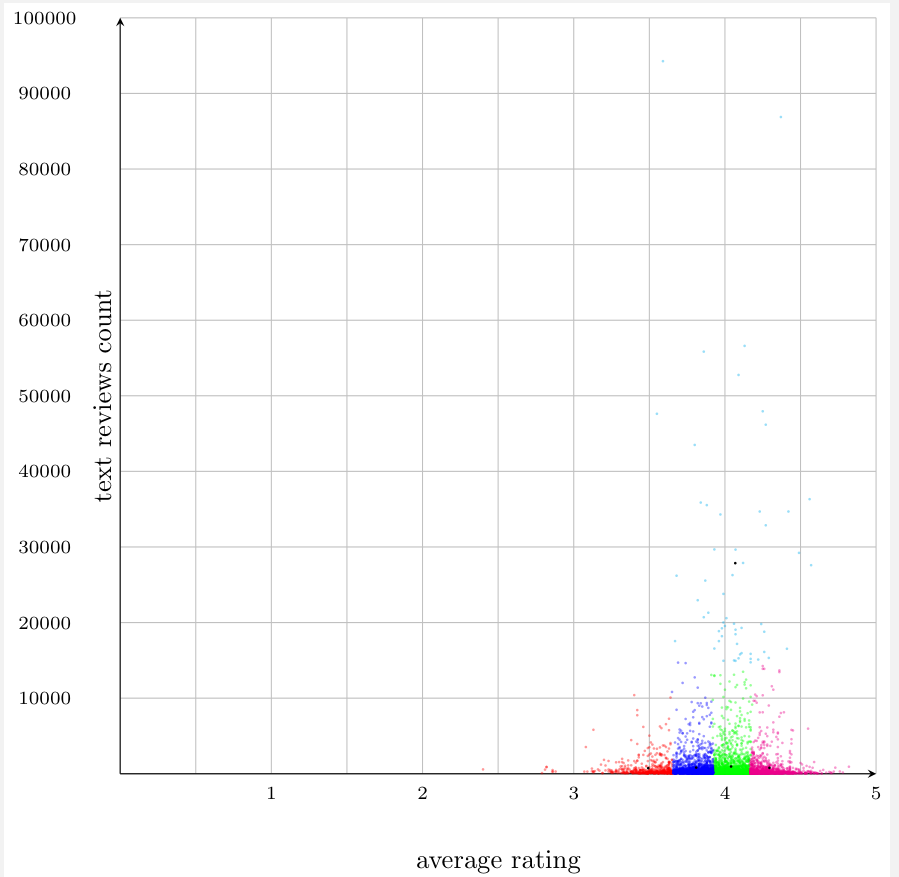


Kiírtam az adatokat



## Eredmények bemutatása és részletes összehasonlítása

Az eredmény a következő grafikon lett:



A fekete pontok jelölik a centroidokat, az öt szín pedig az öt különböző clustert. Az ábráról az olvasható le, hogy azok a könyvek, amelyeket legalább ezren értékeltek, a 3-5 közötti tartományba esnek az átlagos értékelés terén. Mégis a 3,5-4,5 közötti könyvek azok, amelyek nagyobb számú text review-val bírnak (kék és zöld színnel jelölt klaszterek), illetve létezik egy ötödik cián színű klaszter is, amely a kifejezetten sok text review-t generáló könyveket tartalmazza. További érdekesség tehát, hogy a magas text-reviewt generáló könyvek általában nem tartoznak a legmagasabbra értékelt könyvek közé, noha nem is kapnak rossz értékeléseket.

# Saját tapasztalatok levonása, saját gondolatok az elemzés kapcsán

Mindig is érdekelt, hogy milyen kapcsolat van a számszerű értékelés és a szöveges értékelés mennyisége között. Ez a felhőszerű cián klaszter most visszaigazolja azt, hogy érdemes volt ezt a k means clustering algoritmusával megvizsgálni, mert a nagy text review-t (tehát gondolatokat) generáló könyvek nem pusztán legjobb könyvek, hanem jól elkülönülő kategóriát alkotnak a könyvek között. Tehát azt lehetne mondani, hogy az oldalnak leginkább érdekes nagy text-reviewt generáló könyvek nem feltétlenül a „legjobb” könyvek egyben, hanem kell oda valami más is. A csúcstartó könyv egyébként a Twilight című könyv.

Személyes fejlődésemben nagyon sokat számított ennek a beadandónak az elkészítése. Egy barátom javasolta a tikz használatát a grafikonok készítésében, és most azt találtam, hogy nagyon szép képeket lehet vele készíteni – noha elég nehéz volt néha megtalálni benne a hibákat. Kicsit hamar választottam ugyanakkor irányt, mert utólag látom, hogy a scikit nevezetű libraryvel sokkal könnyebben meg lehetett volna oldani mindent. De így úgy érzem, hogy a saját függvények fejlesztésével jobban megértettem a modell matematikai hátterét is.