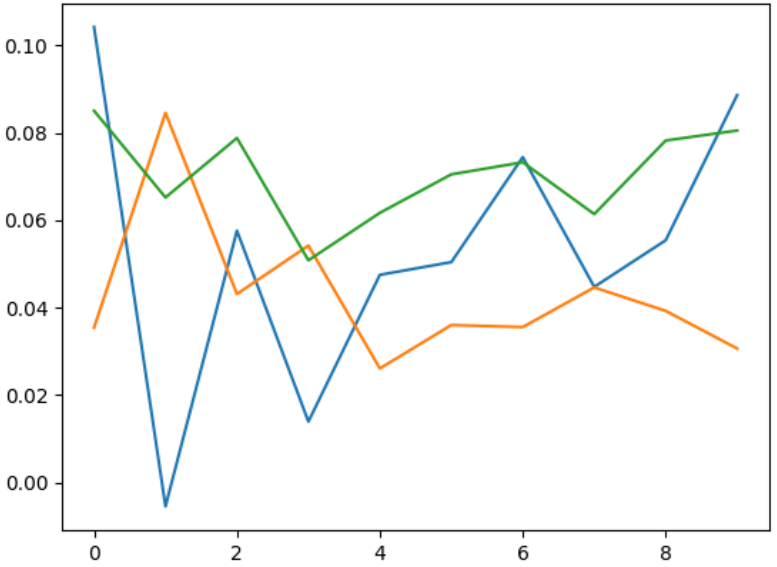
Szövegelemző- és annotálórendszer fejlesztése

Könyv műfaj meghatározás

Önálló Laboratórium

Eszközök - NLTK

- **Tokenizálás**
- **Stop szavak és speciális karakterek törlése**
- **Stemming**



- **Lemmatize**
- **N-grams**
- **Word cloud**
- **Vader**

Korpusz

Project Gutenberg

- **Több mint 60 ezer könyv**
- **Régi könyvek, régi nyelvezet**
- **Címkézett:**
  - Science-fiction: 2367
  - Detective: 990
  - Fantasy: 851
  - Western: 543
  - Romantic: 831

Bag of Words modell

- **Egyszerű**
  - Szótár formájú
  - Szövegkörnyezetet nem veszi figyelembe
  - Kb 50 ezer szó műfajonként

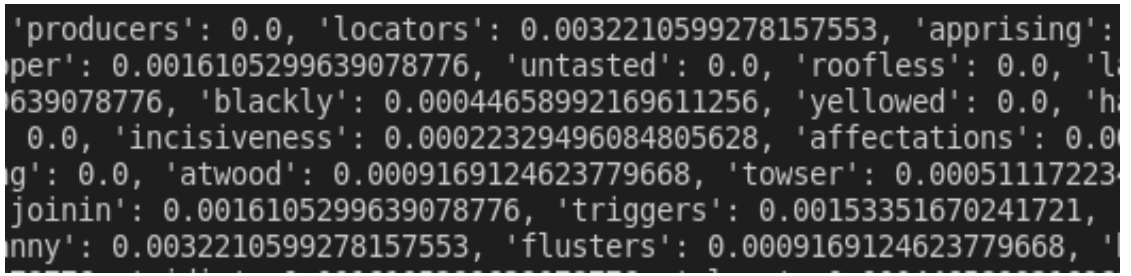

```
{"bow": 71, "ribbon": 23, "chin": 96, "bancroft": 1, "came": 418, "back": 429, "explaining": 16, "driven": 76, "since": 323, "base": 43, "mangan": 1, "peak": 43, "asking": 65, "conra
"new": 348, "team": 55, "horses": 213, "curtis": 7, "looked": 395, "critically": 23, "praising": 2, "points": 50, "approving": 8, "heartily": 45, "told": 347, "bought": 101, "riding"
"become": 180, "expert": 24, "horsewoman": 3, "big": 328, "cottonwood": 28, "tree": 116, "grew": 204, "beside": 232, "gate": 70, "little": 425, "plot": 18, "grass": 159, "enclosed":
"square": 137, "welcome": 84, "green": 167, "exclaimed": 176, "delight": 90, "entered": 193, "tiny": 90, "yard": 61, "stepping": 42, "mincingly": 1, "across": 334, "lifted": 160, "go
"played": 117, "cheeks": 108, "glad": 207, "laughed": 247, "signs": 85, "simply": 116, "must": 388, "walk": 150, "never": 413, "saw": 372, "anything": 305, "lovely": 47, "lawn": 14,
238, "bible": 22, "shadow": 134, "great": 374, "rock": 136, "weary": 52, "land": 213, "yes": 270, "replied": 231, "threw": 172, "open": 323, "door": 299, "knew": 373, "mexico": 84, "
389, "turf": 14, "feel": 223, "uncurled": 2, "whole": 289, "wrinkles": 15, "squints": 1, "around": 341, "eyes": 414, "socorro": 2, "springs": 36, "rambling": 15, "sequence": 7, "room
"part": 292, "northern": 54, "side": 352, "corral": 90, "low": 271, "struggling": 49, "tufts": 8, "weeds": 19, "along": 352, "top": 178, "trailed": 36, "edge": 187, "adding": 25, "ch
"determination": 50, "life": 373, "evade": 9, "overcome": 22, "death": 189, "opened": 220, "row": 58, "outside": 164, "doors": 77, "looking": 335, "toward": 298, "east": 165, "additi
"trickling": 9, "stream": 115, "water": 279, "willows": 21, "cottonwoods": 32, "beyond": 243, "spread": 119, "field": 76, "young": 355, "alfalfa": 20, "stretched": 130, "far": 375, "
58, "sheds": 10, "dazzling": 27, "white": 341, "vivid": 42, "sunshine": 62, "smote": 14, "sight": 261, "like": 431, "blow": 113, "eyeballs": 3, "large": 247, "room": 257, "gayly": 20
16, "hung": 155, "windows": 85, "paper": 104, "covered": 157, "ceiling": 29, "unpainted": 9, "shelves": 18, "pine": 69, "battered": 27, "desk": 48, "filled": 195, "books": 69, "round
"magazines": 11, "tobacco": 88, "pouches": 1, "pipes": 21, "housekeeper": 10, "peters": 7, "brought": 300, "pitcher": 10, "explained": 138, "ranch": 162, "took": 397, "appellation":
265, "years": 343, "party": 132, "spanish": 73, "explorers": 7, "unexpectedly": 27, "upon": 384, "pure": 64, "waters": 69, "almost": 351, "dead": 244, "thirst": 40, "suggestion": 85,
"golden": 66, "twenty": 175, "miles": 265, "farther": 121, "westward": 38, "accidentally": 8, "left": 395, "ajar": 4, "swung": 147, "way": 426, "could": 431, "hear": 270, "plainly":
"wandering": 54, "scarcely": 120, "heeding": 9, "said": 426, "absorbed": 40, "discussion": 29, "local": 38, "politics": 18, "dan": 24, "tillinghurst": 1, "right": 394, "sheriff": 56,
```


- **Korpusz 4:1 arányban tanító- és teszthalmaz**

TF-IDF

- **Műfaj specifikus szavak súlyozása**
- **TF: term frequency**
- **IDF: inverse document frequency**

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$



TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

Tulajdonság vektor

- **Műfajoknál rendezés és összefésülés**
- **Összehasonlítás a könyvek szavaival**

```
western:  {'cowman': 0.027379009386433915, 'greasewood': 0.022547419494710286,
105299639078776, 'cowboys': 0.015407352298515883, 'cañon': 0.0153351670241721,
4494769675170898, 'punchers': 0.013801650321754889, 'chaparral': 0.013753686935
s': 0.012884239711263021, 'mustangs': 0.012884239711263021, 'cayuses': 0.012884
'sonora': 0.011919862010913568, 'shorely': 0.011273709747355143, 'corrals': 0.0
469, 'puncher': 0.010734616916920469, 'policer': 0, 'marple': 0, 'houseboat': 0
le': 0, 'airport': 0, 'mademoiselle': 0, 'boathouse': 0, 'plutocrat': 0, 'noddy
 0, 'sleuths': 0, 'blackmailers': 0, 'mantelpiece': 0, 'jabez': 0, 'clues': 0,
yer': 0, 'ozma': 0, 'munchkin': 0, 'distaff': 0, 'trolls': 0, 'wroth': 0, 'prin
erambulators': 0, 'weisinger': 0, 'fatima': 0, 'genie': 0, 'midmost': 0, 'pumpk
': 0, 'dulcinea': 0, 'sancho': 0.00051117223413907, 'hamete': 0, 'fanaticals':
```

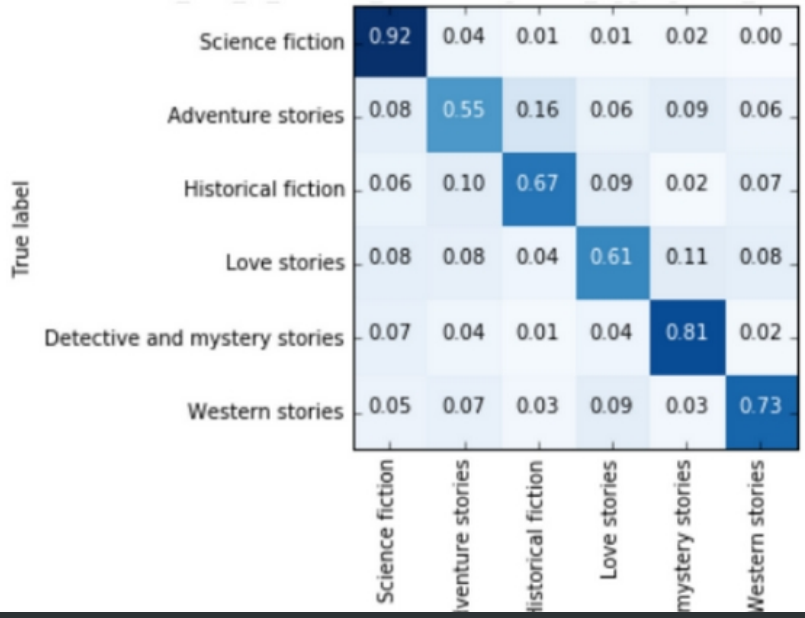
```
fantasy:  {'courtiers': 0.021370751109832494, 'glinda': 0.01816894506683841, 'flotila
893899168758784, 'ozma': 0.013386361147989032, 'munchkin': 0.012825137694238879, 'dis
00973553538035566, 'munchkins': 0.00961885327067916, 'mayst': 0.009127064419083432, '
zier': 0.007802020246446784, 'kline': 0.007481330321639345, 'perambulators': 0.007481
6745, 'midmost': 0.006412568847119439, 'pumpkinhead': 0.006412568847119439, 'numa': 0
68847119439, 'jinns': 0.006412568847119439, 'policer': 0, 'marple': 0, 'houseboat': 0
rt': 0, 'mademoiselle': 0.0016960913579232137, 'boathouse': 0, 'plutocrat': 0, 'noddy
wood': 0, 'sleuths': 0, 'blackmailers': 0, 'mantelpiece': 0, 'jabez': 0, 'clues': 0,
a': 0, 'sancho': 0.00033921827158464276, 'hamete': 0, 'fanaticals': 0, 'cide': 0, 'ri
```

```
detective    0.010226753190491378
fantasy    0.00133362403222096
romance    0.007078779007931591
scifi    0.00900460617347924
western    0.09530348374405267


detective    0.008932362310086008
fantasy    0.006091147538180619
romance    0.004962259702219757
scifi    0.014180704491934298
western    0.017911062105815983


detective    0.009978844211219121
fantasy    0.0009319400636828472
romance    0.014018996135961147
scifi    0.004956365371600169
western    0.0788920333308398
```

Eredmények

- **140 tulajdonság szó / műfaj**

```
szórás
detective : 0.040675295962263 73
romance : 0.04131302589379723
fantasy : 0.03631950044936429
scifi : 0.008917506320195416
western : 0.01657029624971752

átlag
detective : 0.6762962962962963
romance : 0.3754491017964072
fantasy : 0.5751831998931427
scifi : 0.9514767932489451
western : 0.9639143730886851
```




```
teszt:  2

{'detective': {'western': 0.09090909090909091, 'scifi': 0.1422222222222222, 'detective': 0.7163636363636363, 'romance': 0.04040404040404041,
'fantasy': 0.010101010101010102}, 'fantasy': {'western': 0.04878048780487805, 'scifi': 0.14634146341463414, 'detective': 0.13414634146341464,
'romance': 0.054878048780487805, 'fantasy': 0.6158536585365854}, 'romance': {'western': 0.18562874251497005, 'scifi': 0.15568862275449102,
'detective': 0.10359281437125747, 'romance': 0.4053892215568863, 'fantasy': 0.1497005988023952}, 'scifi': {'western': 0.02109704641350211,
'scifi': 0.9514767932489452, 'detective': 0.016877637130801686, 'romance': 0.004219409282700422, 'fantasy': 0.006329113924050633}, 'western':
{'western': 0.9541284403669725, 'scifi': 0.03669724770642202, 'detective': 0.009174311926605505, 'romance': 0.0, 'fantasy': 0.0}}
teszt:  3
```

Eredmények 2

- **60 szó / műfaj**

```
detective : 0.6665319865319865
romance : 0.3199600798403193
fantasy : 0.5351674224682774
scifi : 0.9282700421940927
western : 0.9425076452599388


detective: 0.0215751102270717 83
romance : 0.04408241361589028 4
fantasy : 0.048717181174963 36
scifi : 0.014015384718259482
western : 0.024723404197440 65
```

- **30 szó / műfaj**

```
detective : 0.7244444444444444
romance : 0.226546906187624 77
fantasy : 0.475028963133222 85
scifi : 0.8985935302390998
western : 0.8746177370030581

detective : 0.034244261643343085
romance : 0.033437751044106866
fantasy : 0.052015511646999 33
scifi : 0.020286345690129715
western : 0.032073665081656016
```

Fejlesztési lehetőségek, hibalehetőségek

- Rövid korpusz
- Bag of Words hibái → N-grams
- Szófaj szerinti keresés
- Tulajdonságok normalizálása
- Könyvsorozatok nyelvezete
- Tulajdonnevek szűrése

Köszönöm szépen a figyelmet!