

# cb-02-szymanski-tag\_based\_features\_research

January 4, 2019

This notebook is devoted to the task of adding tag based features to the feature vectors of content based recommendation models.

## 1 Important

make\_features has to be run before running any notebook cell

## 2 Imports

```
In [1]: import numpy as np
import pandas as pd

In [2]: book_tags = pd.read_csv('../data/raw/book_tags.csv')
tags_data = pd.read_csv('../data/raw/tags.csv')
with open("../data/external/genres.txt") as file:
    goodreads_genres = [line.rstrip('\n') for line in file]
```

### 2.1 Data description

```
In [3]: book_tags.head()
```

```
Out[3]:
```

	goodreads_book_id	tag_id	count
0	1	30574	167697
1	1	11305	37174
2	1	11557	34173
3	1	8717	12986
4	1	33114	12716

The data contains information about what tags were assigned to a specific book and how many times was it assigned - the count column in the above presented data frame.

```
In [4]: tags_data.tag_name
```

```
Out[4]:
```

0	-
1	--1-
2	--10-
3	--12-
4	--122-
5	--166-

6		--17-
7		--19-
8		--2-
9		--258-
10		--3-
11		--33-
12		--4-
13		--5-
14		--51-
15		--6-
16		--62-
17		--8-
18		--99-
19	--available-at-raspberrys--	
20		-2001--
21		-calif--
22		-d-c--
23		-dean
24		-england-
25		-fiction
26		-fictional
27		-fictitious
28		-football-
29		-george
	...	
34222		
34223		
34224		
34225		--
34226		-
34227	---	
34228		
34229		--
34230		
34231	--	
34232		
34233		
34234		-
34235		
34236		
34237		
34238		-
34239		
34240		
34241		
34242		
34243		
34244	moonplus-reader	

```

34245          -
34246          -
34247          hildrens
34248
34249
34250
34251
Name: tag_name, Length: 34252, dtype: object

```

Unfortunately, some tags are defined in other languages than english and some tags contain no specific information as for example --5-. That is why only tags representing genres will be kept as book features. The considered set of features is presented in the cell below

```
In [5]: goodreads_genres
```

```

Out[5]: ['10th-century',
        '11th-century',
        '12th-century',
        '13th-century',
        '14th-century',
        '15th-century',
        '16th-century',
        '17th-century',
        '1864-shenandoah-campaign',
        '18th-century',
        '1917',
        '19th-century',
        '1st-grade',
        '20th-century',
        '21st-century',
        '2nd-grade',
        '40k',
        'abandoned',
        'abuse',
        'academia',
        'academic',
        'academics',
        'accounting',
        'accra',
        'action',
        'activism',
        'adaptations',
        'addis-ababa',
        'addition',
        'adolescence',
        'adoption',
        'adult',
        'adult-colouring-books',
        'adult-fiction',

```

'adventure',  
'adventurers',  
'aeroplanes',  
'africa',  
'african-american',  
'african-american-literature',  
'african-american-romance',  
'african-literature',  
'agender',  
'agriculture',  
'aircraft',  
'airliners',  
'airships',  
'albanian-literature',  
'alchemy',  
'alcohol',  
'alexandria',  
'algeria',  
'algiers',  
'algorithms',  
'aliens',  
'alternate-history',  
'alternate-universe',  
'alternative-medicine',  
'amateur-sleuth',  
'amazon',  
'ambulance-service',  
'ambulances',  
'american',  
'american-civil-war',  
'american-classics',  
'american-fiction',  
'american-history',  
'american-novels',  
'american-revolution',  
'american-revolutionary-war',  
'americana',  
'amish',  
'amish-fiction',  
'amish-historical-romance-fiction',  
'ancient',  
'ancient-history',  
'androgyny',  
'angels',  
'anglo-saxon',  
'angola',  
'animal-fiction',  
'animals',

'anime',  
'anthologies',  
'anthropology',  
'anthropomorphic',  
'anti-racist',  
'antietam-campaign',  
'antiquities',  
'antisemitism',  
'apocalyptic',  
'apple',  
'applied-mathematics',  
'appomattox-campaign',  
'archaeology',  
'architecture',  
'arithmetic',  
'army-of-northern-virginia',  
'army-of-tennessee',  
'army-of-the-potomac',  
'art',  
'art-and-photography',  
'art-books-monographs',  
'art-design',  
'art-history',  
'arthurian',  
'artificial-intelligence',  
'asexual',  
'asia',  
'asian-literature',  
'asmara',  
'aspergers',  
'astrology',  
'astronomy',  
'atheism',  
'atlanta-campaign',  
'atlases',  
'australia',  
'autobiography',  
'aviation',  
'aviation-history',  
'azeroth',  
'babylon-5',  
'back-to-school',  
'baha-i',  
'bande-dessinée',  
'bangladesh',  
'banking',  
'banks',  
'banned-books',

'barter',  
'baseball',  
'basketball',  
'batman',  
'battle-of-britain',  
'battle-of-gettysburg',  
'battles-of-the-american-civil-war',  
'bdsm',  
'beading',  
'beauty-and-the-beast',  
'beer',  
'belgian',  
'belgium',  
'belief',  
'benghazi',  
'benin',  
'beverages',  
'biblical',  
'biblical-fiction',  
'bicycles',  
'biography',  
'biography-memoir',  
'biology',  
'bird-watching',  
'birds',  
'bisexual',  
'bisexual-romance',  
'bizarro-fiction',  
'black-literature',  
'boarding-school',  
'bolivia',  
'bolsheviks',  
'bombers',  
'bookkeeping',  
'books-about-books',  
'booze',  
'botswana',  
'boys-love',  
'brain',  
'brazil',  
'brazzaville',  
'brewing',  
'british-literature',  
'buddhism',  
'buffy-the-vampire-slayer',  
'bulgaria',  
'bulgarian-literature',  
'burkina-faso',

'burundi',  
'buses',  
'business',  
'butch-femme',  
'cabo-verde',  
'cairo',  
'calculation',  
'calculus',  
'cameroon',  
'campus',  
'canada',  
'canadian-literature',  
'canon',  
'canterbury-theses',  
'cape-town',  
'carolinas-campaign',  
'carolingian',  
'cars',  
'cartography',  
'cartoon',  
'castile',  
'category-romance',  
'catholic',  
'cats',  
'central-africa',  
'central-african-republic',  
'chad',  
'challenged-books',  
'chancellorsville-campaign',  
'chapter-books',  
'chemistry',  
'chess',  
'chick-lit',  
'childrens',  
'childrens-classics',  
'china',  
'chinese-literature',  
'chivalry',  
'choose-your-own-adventure',  
'christian',  
'christian-contemporary-fiction',  
'christian-fantasy',  
'christian-fantasy-dystopian',  
'christian-fiction',  
'christian-fiction-amish',  
'christian-historical-fiction',  
'christian-living',  
'christian-non-fiction',

'christian-romance',  
'christian-romance-historical',  
'christianity',  
'christmas',  
'church',  
'church-history',  
'church-ministry',  
'cinderella',  
'cisgender',  
'cities',  
'civil-war',  
'civil-war-confederacy',  
'civil-war-eastern-theater',  
'civil-war-history',  
'civil-war-navy',  
'civil-war-western-theater',  
'class',  
'class-issues',  
'classic-literature',  
'classical-music',  
'classical-studies',  
'classics',  
'clean-romance',  
'climate-change',  
'climate-change-fiction',  
'climbing',  
'cocktails',  
'coding',  
'coin-collecting',  
'collections',  
'college',  
'colouring-books',  
'comedy',  
'comic-book',  
'comic-strips',  
'comics',  
'comics-manga',  
'coming-of-age',  
'comix',  
'communication',  
'comoros',  
'complementary-medicine',  
'computation',  
'computer-reference',  
'computer-science',  
'computers',  
'conservation',  
'consumer-economics',



'contemporary',  
'contemporary-romance',  
'contemporary-romance-clean',  
'cookbooks',  
'cooking',  
'coptic-language',  
'coptology',  
'copts',  
'counselling',  
'counter-culture',  
'counting',  
'couture',  
'cozy-mystery',  
'crafts',  
'crime',  
'criticism',  
'crochet',  
'cross-dressing',  
'cthulhu-mythos',  
'cuisine',  
'culinary',  
'cult-classics',  
'cults',  
'cultural',  
'cultural-heritage',  
'cultural-studies',  
'curation',  
'currency',  
'cw-bentonville',  
'cw-iuka-corinth',  
'cw-stones-river',  
'cw-west-va',  
'cyberpunk',  
'cycling',  
'czech-literature',  
'dakar',  
'danish',  
'dark',  
'dark-fantasy',  
'dc-comics',  
'death',  
'democratic-republic-of-the-congo',  
'demons',  
'denmark',  
'design',  
'detective',  
'diary',  
'dictionaries',

'diets',  
'dinosaurs',  
'disability',  
'disability-studies',  
'disabled-communities',  
'discipleship',  
'disease',  
'divination',  
'division',  
'divorce',  
'djibouti',  
'doctor-who',  
'dogs',  
'domestic-science',  
'drag',  
'dragonlance',  
'dragons',  
'drama',  
'drawing',  
'dressmaking',  
'drinking',  
'dungeons-and-dragons',  
'dungeons-and-dragons-manuals',  
'dutch-literature',  
'dying-earth',  
'dyscalculia',  
'dystopia',  
'earth',  
'earth-sciences',  
'eastern-africa',  
'eastern-philosophy',  
'ecclesiology',  
'ecology',  
'economic-development',  
'economics',  
'education',  
'edwardian',  
'egypt',  
'egyptian-literature',  
'egyptology',  
'electrical-engineering',  
'elizabethan-period',  
'emergency-services',  
'engineering',  
'english-civil-war',  
'english-literature',  
'entrepreneurship',  
'environment',

'epic',  
'epic-fantasy',  
'epic-poetry',  
'equatorial-guinea',  
'eritrea',  
'erotic-historical-romance',  
'erotic-horror',  
'erotic-paranormal-romance',  
'erotic-romance',  
'erotica',  
'esoterica',  
'esp',  
'espionage',  
'essays',  
'ethiopia',  
'ethnic',  
'ethnic-studies',  
'ethnicity',  
'ethnography',  
'eunuch',  
'european-history',  
'european-literature',  
'evangelism',  
'evolution',  
'f-f-f',  
'f-m-f',  
'fables',  
'fae',  
'fairies',  
'fairy-tale-retellings',  
'fairy-tales',  
'faith',  
'family',  
'family-law',  
'fandom',  
'fantasy',  
'fantasy-of-manners',  
'fantasy-romance',  
'far-right',  
'fashion',  
'fat',  
'fat-acceptance',  
'fat-studies',  
'feminism',  
'feminist-studies',  
'feminist-theory',  
'femslash',  
'ferdinand-and-isabella',

'ferries',  
'fiction',  
'field-guides',  
'fighters',  
'figure-skating',  
'film',  
'finance',  
'financial-management',  
'finnish-literature',  
'fire-engines',  
'fire-services',  
'firefighters',  
'fitness',  
'flash-fiction',  
'folk-tales',  
'folklore',  
'food',  
'food-and-drink',  
'food-and-wine',  
'food-history',  
'food-preservation',  
'food-writing',  
'foodie',  
'football',  
'forgotten-realms',  
'foster-children',  
'foster-parents',  
'fostering',  
'foursome',  
'fractured-fairy-tales',  
'france',  
'freetown',  
'freight',  
'french-literature',  
'french-revolution',  
'frugal',  
'functional-analysis',  
'funnies',  
'funny',  
'futurism',  
'futuristic',  
'futuristic-romance',  
'gabon',  
'gaborone',  
'game-design',  
'gamebooks',  
'games',  
'gaming',

'gaming-fiction',  
'gardening',  
'gastronomy',  
'gay',  
'gay-erotica',  
'gay-fiction',  
'gay-for-you',  
'geek',  
'gemstones',  
'gender',  
'gender-and-sexuality',  
'gender-identity',  
'gender-roles',  
'gender-studies',  
'genderfluid',  
'genderfuck',  
'genderqueer',  
'genetics',  
'geoffrey-chaucer',  
'geography',  
'geology',  
'geometry',  
'georgian',  
'georgian-romance',  
'german-literature',  
'germany',  
'ghana',  
'ghost-stories',  
'ghosts',  
'gliders',  
'global-warming',  
'gnosticism',  
'go',  
'god',  
'goddess',  
'gods',  
'golden-age-mystery',  
'google',  
'goth',  
'gothic',  
'gothic-horror',  
'gothic-revival',  
'gothic-romance',  
'government',  
'grad-school',  
'graffiti',  
'graphic-literature',  
'graphic-non-fiction',

'graphic-novels',  
'graphic-novels-comics',  
'graphic-novels-comics-manga',  
'graphic-novels-manga',  
'graphica',  
'greece',  
'greek-mythology',  
'green',  
'grimm',  
'growth-mindset',  
'guidebook',  
'guides',  
'guinea',  
'guinea-bissau',  
'habsburg-empire',  
'habsburg-spain',  
'hackers',  
'halloween',  
'harare',  
'hard-boiled',  
'hard-science-fiction',  
'harem',  
'harlequin',  
'harlequin-blaze',  
'harlequin-desire',  
'harlequin-heartwarming',  
'harlequin-historical',  
'harlequin-kimani-romance',  
'harlequin-medical-romance',  
'harlequin-nocturne',  
'harlequin-presents',  
'harlequin-romance',  
'harlequin-romantic-suspense',  
'harlequin-teen',  
'health',  
'health-care',  
'helicopters',  
'herbs',  
'heritage-preservation',  
'heroic-fantasy',  
'hierarchy',  
'high-fantasy',  
'high-school',  
'hinduism',  
'hip-hop',  
'historical',  
'historical-fantasy',  
'historical-fiction',

'historical-mystery',  
'historical-romance',  
'historical-romance-clean',  
'history',  
'history-and-politics',  
'history-civil-war-eastern-theater',  
'history-of-science',  
'hockey',  
'holiday',  
'holland',  
'holocaust',  
'home-economics',  
'horror',  
'horse-racing',  
'horses',  
'horticulture',  
'hot-air-balloons',  
'how-to',  
'hqn',  
'hugo-awards',  
'huguenots',  
'human-capital',  
'human-development',  
'human-ecology',  
'human-resources',  
'humanities',  
'humor',  
'hungarian-literature',  
'hungary',  
'hydrogeology',  
'illness',  
'income-tax',  
'india',  
'indian-literature',  
'indigenous-history',  
'indonesian-literature',  
'informatics',  
'information-science',  
'innnumeracy',  
'inspirational',  
'intensive-care',  
'international',  
'international-development',  
'international-literature',  
'international-relations',  
'internet',  
'interracial-romance',  
'intersex',

'iran',  
'ireland',  
'irish-civil-war',  
'irish-literature',  
'islam',  
'islamic-terrorism',  
'islamism',  
'israel',  
'italian-literature',  
'italy',  
'ivory-coast',  
'japan',  
'japanese-history',  
'japanese-literature',  
'jazz',  
'jewellery',  
'jewellery-making',  
'jewish',  
'johannesburg',  
'josei',  
'journal',  
'journaling',  
'journalism',  
'judaica',  
'judaism',  
'juvenile',  
'kazakhstan',  
'kenya',  
'khartoum',  
'kigali',  
'kinshasa',  
'knitting',  
'komik',  
'ku-klux-klan',  
'labor',  
'lais',  
'landscaping',  
'language',  
'lapidary',  
'latin-american',  
'latin-american-history',  
'latin-american-literature',  
'law',  
'lds',  
'lds-fiction',  
'lds-non-fiction',  
'leadership',  
'lebanon',



'led-zeppelin',  
'legal-thriller',  
'lenin',  
'leningrad',  
'lesbian',  
'lesbian-fiction',  
'lesbian-romance',  
'lesbotronic',  
'lesotho',  
'lgbt',  
'lgbt-memoir',  
'liberia',  
'librarianship',  
'library-science',  
'libya',  
'light-novel',  
'linguistics',  
'literary-criticism',  
'literary-fiction',  
'literature',  
'local-history',  
'logic',  
'london-underground',  
'love',  
'love-inspired',  
'love-inspired-historical',  
'love-inspired-suspense',  
'love-story',  
'lovecraftian',  
'loveswept',  
'low-fantasy',  
'lusaka',  
'luxemburg',  
'm-f-f',  
'm-f-m',  
'm-m-aliens',  
'm-m-contemporary',  
'm-m-f',  
'm-m-fantasy',  
'm-m-historical-romance',  
'm-m-horror',  
'm-m-m',  
'm-m-m-f',  
'm-m-m-m',  
'm-m-mystery',  
'm-m-new-adult',  
'm-m-office-romance',  
'm-m-paranormal',

'm-m-romance',  
'm-m-romantic-suspense',  
'm-m-science-fiction',  
'm-m-shapeshifters',  
'm-m-sports-romance',  
'm-m-supernatural',  
'm-m-urban-fantasy',  
'm-m-young-adult',  
'madagascar',  
'magic',  
'magical-realism',  
'magick',  
'mail-order-brides',  
'malabo',  
'malawi',  
'mali',  
'management',  
'managers',  
'manga',  
'manga-romance',  
'manhwa',  
'mannerpunk',  
'maps',  
'marathi',  
'maritime',  
'marriage',  
'martial-artist',  
'martial-arts',  
'martyr',  
'martyrdom',  
'marvel',  
'mary-i',  
'material-culture',  
'mathematics',  
'mauritania',  
'mauritius',  
'media-tie-in',  
'medical',  
'medicine',  
'medieval',  
'medieval-history',  
'medieval-romance',  
'medievaesque',  
'medievalism',  
'memoir',  
'menage',  
'mental-health',  
'mental-illness',

'mermaids',  
'metallurgy',  
'metaphysics',  
'microhistory',  
'middle-english-literature',  
'middle-grade',  
'military-fiction',  
'military-history',  
'military-romance',  
'military-science-fiction',  
'mills-and-boon',  
'mineralogy',  
'mira',  
'mixed-martial-arts',  
'mmorpg',  
'modern',  
'modern-classics',  
'mogadishu',  
'mombasa',  
'money',  
'money-management',  
'monrovia',  
'mormonism',  
'moroccan',  
'morocco',  
'moscow',  
'motorcycle',  
'motorcycling',  
'mountaineering',  
'movies',  
'mozambique',  
'multicultural-literature',  
'multiple-partners',  
'multiplication',  
'murder-mystery',  
'muscovy',  
'museology',  
'museums',  
'music',  
'music-biography',  
'musicals',  
'musician-erotica',  
'musicians',  
'muslimah',  
'muslims',  
'mystery',  
'mystery-thriller',  
'mysticism',

'mythology',  
'nairobi',  
'namibia',  
'native-american-history',  
'native-americans',  
'natural-history',  
'nature',  
'nazarene',  
'nazi-party',  
'near-future',  
'neo-medieval',  
'nerd',  
'neuroscience',  
'new-adult',  
'new-adult-romance',  
'new-age',  
'new-testament',  
'new-weird',  
'new-york',  
'niger',  
'nigeria',  
'noir',  
'non-fiction',  
'nordic-noir',  
'norman',  
'north-american-history',  
'northern-africa',  
'novella',  
'novels',  
'nsfw',  
'number',  
'number-theory',  
'numeracy',  
'numismatics',  
'nursery-rhymes',  
'nursing',  
'nutrition',  
'occult',  
'occult-detective',  
'old-english-literature',  
'old-testament',  
'omegaverse',  
'oral-history',  
'organizational-culture',  
'origami',  
'ornithology',  
'outdoors',  
'overland-campaign',

'paganism',  
'pakistan',  
'palaeontology',  
'palaeozoology',  
'paranormal',  
'paranormal-mystery',  
'paranormal-romance',  
'paranormal-urban-fantasy',  
'parenting',  
'patternmaking',  
'peak-oil',  
'pediatricians',  
'pediatrics',  
'peninsula-campaign',  
'personal-development',  
'personal-finance',  
'petrograd',  
'philip-ii-of-spain',  
'philosophy',  
'photography',  
'physics',  
'picture-books',  
'picu',  
'pirates',  
'planetary-romance',  
'planetary-science',  
'planets',  
'plantagenet',  
'plants',  
'plays',  
'plus-size',  
'poetry',  
'poland',  
'police',  
'polish-literature',  
'political-development',  
'political-science',  
'politics',  
'polyamorous',  
'polyamory',  
'polyandry',  
'polygamy',  
'polygyny',  
'pop-culture',  
'popular-science',  
'pornography',  
'portugal',  
'portuguese-literature',

'post-apocalyptic',  
'poverty',  
'prayer',  
'pre-k',  
'pre-raphaelite',  
'prehistoric',  
'prehistory',  
'preservation',  
'presidents',  
'pretoria',  
'princesses',  
'productivity',  
'professors',  
'programming',  
'programming-languages',  
'prostitution',  
'psychiatry',  
'psychoanalysis',  
'psychological-thriller',  
'psychology',  
'public-transport',  
'pulp',  
'pulp-adventure',  
'pulp-noir',  
'punk',  
'punx',  
'puzzles',  
'quantum-mechanics',  
'queer',  
'queer-lit',  
'queer-studies',  
'quilting',  
'rabat',  
'rabbits',  
'race',  
'racing',  
'railway-history',  
'railways',  
'read-for-college',  
'read-for-school',  
'real-person-fiction',  
'realistic-fiction',  
'realistic-young-adult',  
'recreation',  
'recruitment',  
'reference',  
'regency',  
'regency-romance',

'regency-romance-clean',  
'relationships',  
'religion',  
'republic-of-the-congo',  
'research',  
'retellings',  
'reverse-harem',  
'road-trip',  
'robots',  
'rock-n-roll',  
'role-playing-games',  
'roman',  
'roman-britain',  
'romance',  
'romania',  
'romanian-literature',  
'romanovs',  
'romantic',  
'romantic-suspense',  
'romanticism',  
'royal-air-force',  
'rus',  
'russia',  
'russian-empire',  
'russian-federation',  
'russian-history',  
'russian-literature',  
'russian-revolution',  
'rwanda',  
'saint-helena',  
'sao-tome',  
'sao-tome-and-principe',  
'satanism',  
'scandinavian-literature',  
'school',  
'school-stories',  
'science',  
'science-fiction',  
'science-fiction-fantasy',  
'science-fiction-romance',  
'science-nature',  
'scooters',  
'scores',  
'scotland',  
'scripture',  
'seinen',  
'self-help',  
'semiotics',

```

'senegal',
'sequential-art',
'serbian-literature',
'sewing',
'sex-and-erotica',
'sex-work',
...]

```

## 2.2 How to represent tags as features?

The question is how those tags should be converted to features. The following ideas are considered:

- append tags counts to existing feature vectors
- normalize the tags count in order to measure ‘how much fictional’ is the considered book

The problem of the first approach is that one book might have been assigned a 100 times and another one a 1000 times. For example the first one got the comic-book tag assigned a 100 times and the second one got tagged as comic-book 300 times. Now the first book seems like a pure comic-book but in terms of quantities the second book is ‘more’ comic-book than the first even though it is just partly a comic book.

The first step is to check the average amount of unique tags assigned to a single book.

```

In [6]: book_tags_names = book_tags.merge(tags_data)
        book_tags_names = book_tags_names[book_tags_names.tag_name.isin(goodreads_genres)]
        tags_assigned_count = book_tags_names.groupby(
            'goodreads_book_id')['tag_id'].apply(np.unique).apply(len).reset_index()['tag_id']

```

```

In [7]: tags_assigned_count.describe()

```

```

Out [7]: count      10000.000000
        mean         20.712700
        std           5.206311
        min           3.000000
        25%          17.000000
        50%          20.000000
        75%          24.000000
        max          43.000000
        Name: tag_id, dtype: float64

```

On average a single book has 20 different tags assigned. This makes it an relevant feature as having 20 tags overall is not overspecific, but provides useful insights at the same time. Additionally, the small dimensionality allows omitting heavy computations.

## 2.3 Feature extraction result analysis

```

In [8]: tag_features = pd.read_csv('../features/tag_based_features.csv', index_col='book_id')

```

```

In [9]: tag_features.apply(sum, axis=1).head()

```



```
Out[9]: book_id
      1      1.0
      2      1.0
      3      1.0
      4      1.0
      5      1.0
      dtype: float64
```

```
In [10]: all(tag_features.apply(sum, axis=1).apply(round, 1) == 1)
```

```
Out[10]: True
```

All values sum up to 1 in each row which means that the tags count were normalized correctly. The reason why the sum was rounded up is because while extracting features computations were made on floating numbers which do not provide perfect accuracy.

### 3 Bibliography