دانشکده مهندسی کامپیوتر

استاد درس: دکتر اعتمادی

بهار ۱۴۰۰

گزارش پروژه فاز ۱
مبانی پردازش زبان و گفتار

پریسا یل‌سوار
شماره دانشجویی: ۹۶۵۲۲۰۸۷

۱ جمع آوری داده

به منظور جمع آوری خلاصه انیمه و ژانرهای مربوط از دو وبسایت anime-planet و AnimeSeries استفاده شده است. از این وبسایت‌ها به ترتیب عنوان انیمه، خلاصه انیمه و لیستی از ژانرهای آن گردآوری شده است.

کد جمع آوری داده در فایل NLP_Project_Data_Raw.ipynb به تفکیک هر وبسایت قرار دارد. در استخراج داده از همه وبسایت‌ها از ابزار BeautifulSoap برای خواندن و پارس کردن html دانلود شده استفاده شده است. اطلاعات عنوان انیمه، خلاصه انیمه و ژانرهای آن نیز با بررسی و استخراج html tag های مربوط به دست آمده است.

از وبسایت anime-planet داده انیمه و مانگا[1] استخراج شده که به ترتیب در فایل‌های anime-planet.csv و anime-planet-manga.csv ذخیره شده‌اند. از وبسایت AnimeSeries فقط داده انیمه جمع آوری شده که در فایل anime-series.csv ذخیره شده است. واحد برچسب گذاری ، خلاصه هر انیمه یا مانگا است. این برچسب‌ها به صورت آماده از صفحه هر انیمه یا مانگا استخراج می‌شوند و نیازی به برچسب گذاری مجدد نیست.

فایل‌های anime-planet.csv ، anime-planet-manga.csv و anime-series.csv به ترتیب دارای ۱۵۹۸۴ ، ۶۱۶۷۹ و ۸۸۰۰ سطر یا ردیف هستند.

۲ پیش پردازش‌های داده

برای تمیز کردن و آماده کردن داده برای فازهای بعدی ، پیش پردازش‌هاش زیر روی مجموعه داده انجام شده است.

۱.۲ تمیز کردن داده

کد موجود در فایل NLP_Project_Clean_Data.ipynb فایل‌های دخیر شده در پوشه raw را (anime-planet.csv ، anime-planet-manga.csv و anime-series.csv) تبدیل به یک فایل با ۸۶۴۶۳ سطر می‌کند.

برای تمیز کردن مجموعه دادگان ابتدا سطرهایی که مقادیر ژانر یا خلاصه آن‌ها None باشد، حذف می‌شوند. سپس سطرهایی که خلاصه آن‌ها تکراری و با دیگر خلاصه‌ها یکسان باشد نیز حذف می‌شوند. در آخر خلاصه‌هایی که تعداد جملاتشان از ۳ عدد کمتر باشند همراه با سطر مربوط حذف می‌شوند. (جملات با استفاده از nltk sent_tokenize تفکیک می‌شوند.) تعداد سطرهای دادگان باقی مانده ۳۷۲۸۳ است. حاصل این قسمت فایل anime_df.pkl است که در پوشه clean ذخیره شده است.

۲.۲ تفکیک اسناد

در فایل NLP_Project_Clean_Data.ipynb قسمت Visualize document count for each genre از خروجی قسمت قبل استفاده می‌کند تا داکیومت‌های مربوط به هر ژانر تفکیک شوند.

در مجموع ۶۱۲ ژانر وجود دارد که از این تعداد ۱۲ ژانر اول با بیشترین تعداد سند موجود برای ادامه پردازش‌ها جدا می‌شوند. خروجی این قسمت فایل document_level_df.pkl است که در پوشه clean ذخیره شده است. (برای دخیره دادگان از فرمت pickle استفاده می‌شود، چرا که فرمت csv قادر به نگهداری لیست‌های طولانی تو در تو نیست و موجب از دست رفتن اطلاعات می‌شد.)

[1] Manga

۳.۲ تفکیک جملات

کد موجود در فایل NLP_Project_Sentence_Broken.ipynb از مجموعه داده anime_df.pkl استفاده می‌کند و جملات هر داکیومنت را با استفاده از nltk sent_tokenize تفکیک می‌کند.

این جملات نیز مجدد پیش پردازش می‌شوند و تمام علائم نگارشی ، کلمات پرتکرار زبان انگلیسی (stopwords) ، html tag هایی که ممکن است باقی مانده باشد و نیم فاصله (half space) و فاصله (space) های متوالی حذف می‌شوند. تمام اعداد نیز با توکن <NUM> جایگزین می‌شوند. (تمام این موارد با کمک regex تشخیص و جایگزین می‌شوند.) کلمات یا حروفی که با الفبای انگلیسی نوشته نشده باشند، مانند کلمات ژاپنی نیز از جملات حذف می‌شوند. در آخر برای حذف کردن تک حروف و خلاصه‌های کوتاه ، جملات با طول کوچکتر ۲ و مجموعه جملات با تعداد کوچکتر از ۳ حذف می‌شوند.

مجموعه داده sentence_df.pkl پس از این پردازش‌ها ۳۶۸۲۹ سطر دارد و در پوشه sentence_broken ذخیره شده است.

تفکیک کردن ژانرهای موجود در قسمت Visualize sentence count for each genre انجام می‌شود.

از بین همه ژانرهای موجود فقط ۱۲ عدد با بیشترین تعداد جملات در نظر گرفته می‌شوند. مجموعه داده به تفکیک برچسب با نام sentence_level_df.pkl در پوشه sentence_broken ذخیره شده است.

۴.۲ تفکیک کلمات

کد موجود در فایل NLP_Project_Word_Broken.ipynb از فایل قسمت قبل sentence_df.pkl استفاده می‌کند و کلمات را با استفاده از nltk word_tokenize تفکیک می‌کند. در این مرحله به دلیل اینکه مجموعه داده عاری از کلمات پرتکرار و غیر انگلیسی ست؛ فقط lemmatization روی کلمات انجام می‌شود (با استفاده از nltk WordNetLemmatizer) و کلمات با طول ۱ و کمتر نیز حذف می‌شوند. (کلمات با طول ۱ همان حروف الفبا هستند که ارزش چندانی برای هدف این پروژه نخواهند داشت.)

حاصل این پردازش‌ها با نام word_df.pkl در پوشه word_broken ذخیره شده است.

در آخر فرآیند تفکیک برچسب‌ها در قسمت Visualize word count for each tag انجام می‌شوند و فقط ۱۲ برچسب با بیشترین تعداد کلمات انتخاب می‌شوند. مجموعه داده به تفکیک برچسب با نام word_level_df.pkl در پوشه word_broken ذخیره شده است.

۳ استخراج آمار

در این قسمت آمار و اطلاعات خواسته شده از مجموعه داده استخراج شده و تنایج آن آورده شده است.

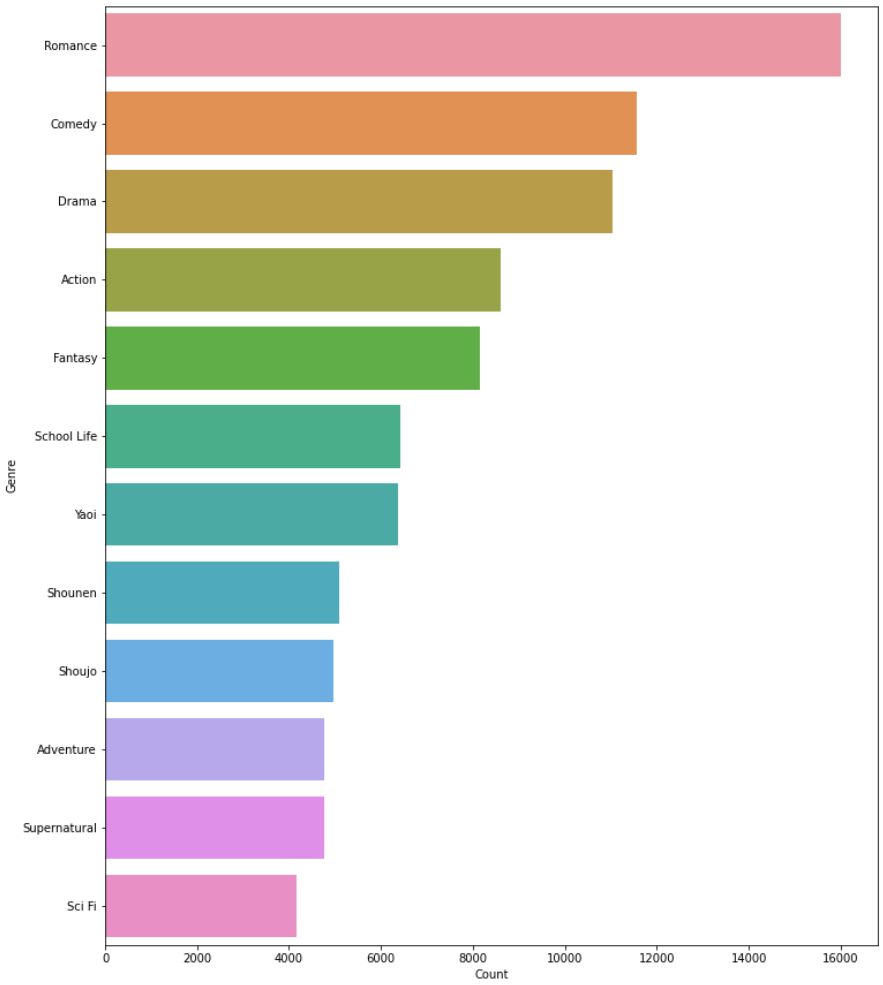
۱.۳ تعداد واحد داده

کد به دست آوردن تعداد داکیومنت برای هر برچسب در قسمت

Visualize document count for each genre فایل NLP_Project_Clean_Data.ipynb قرار دارد.

در کل ۶۱۲ برچسب وجود دارد که از این بین ۱۲ برچسب با بیشترین تعداد داکیومنت انتخاب شده‌اند.

خروجی آن در قالب جدول و نمودار در ادامه آمده است.



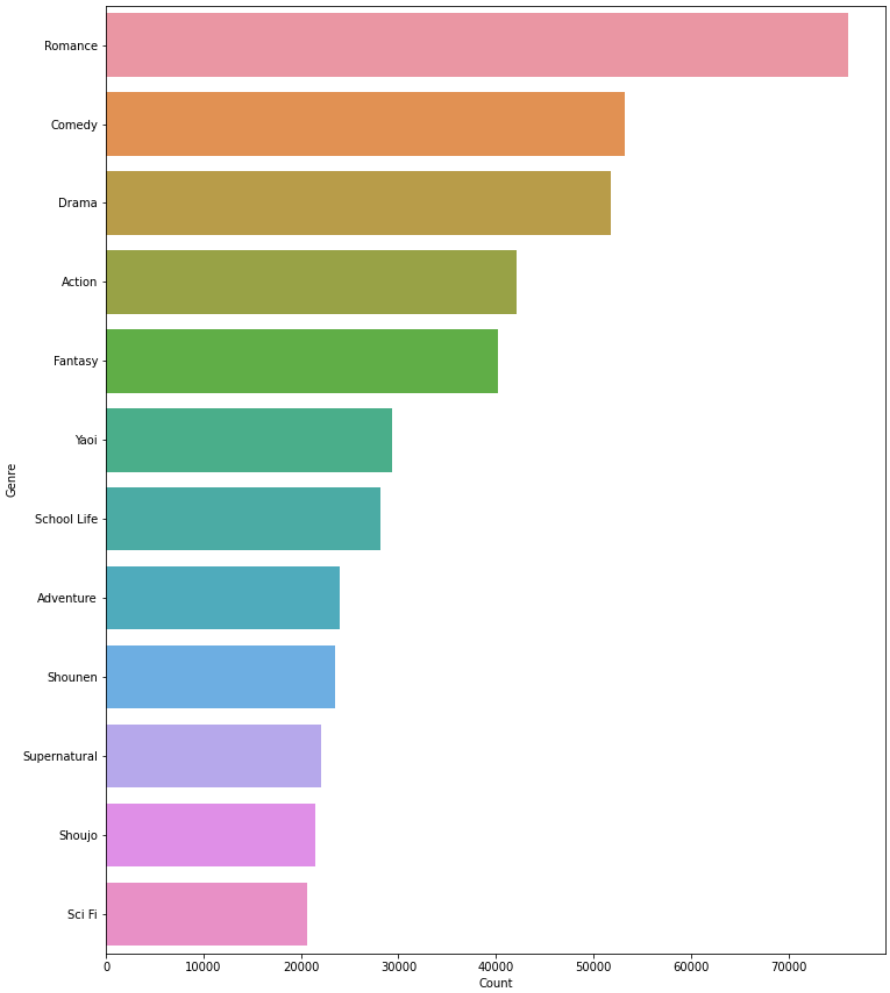
شکل ۱: نمودار تعداد اسناد

	Genre	Count
0	Action	8613
1	Adventure	4773
2	Comedy	11567
3	Drama	11041
4	Fantasy	8144
5	Romance	15996
6	School Life	6428
7	Sci Fi	4173
8	Shoujo	4978
9	Shounen	5094
10	Supernatural	4763
11	Yaoi	6380

شکل ۲: جدول تعداد اسناد

۲.۳ تعداد جملات

کد به دست آوردن تعداد جملات برای هر برچسب در قسمت `Visualize sentence count for each genre` فایل `NLP_Project_Sentence_Broken.ipynb` قرار دارد. در کل ۶۱۲ برچسب وجود دارد که از این بین ۱۲ برچسب با بیشترین تعداد داکیومنت انتخاب شده‌اند. خروجی آن در قالب جدول و نمودار در ادامه آمده است.



شکل ۳: نمودار تعداد جملات

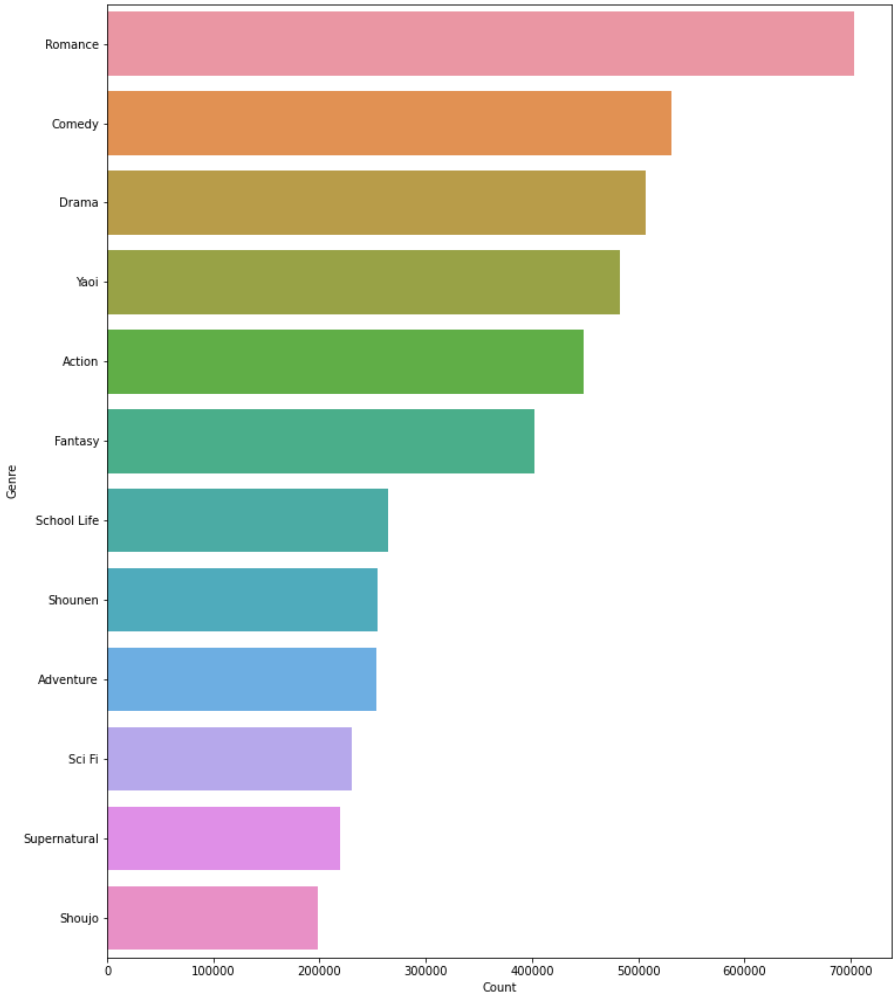
	Genre	Count
0	Action	42067
1	Adventure	23959
2	Comedy	53225
3	Drama	51836
4	Fantasy	40236
5	Romance	76083
6	School Life	28163
7	Sci Fi	20595
8	Shoujo	21475
9	Shounen	23499
10	Supernatural	22080
11	Yaoi	29322

شکل ۴: جدول تعداد جملات

۳.۳ تعداد کلمات

کد به دست آوردن تعداد کلمات برای هر برچسب در قسمت `Visualize word count for each genre` فایل `NLP_Project_Word_Broken.ipynb` قرار دارد. در کل ۶۱۲ برچسب وجود دارد که از این بین ۱۲ برچسب با بیشترین تعداد داکیومنت انتخاب شده‌اند.

خروجی آن در قالب جدول و نمودار در ادامه آمده است.



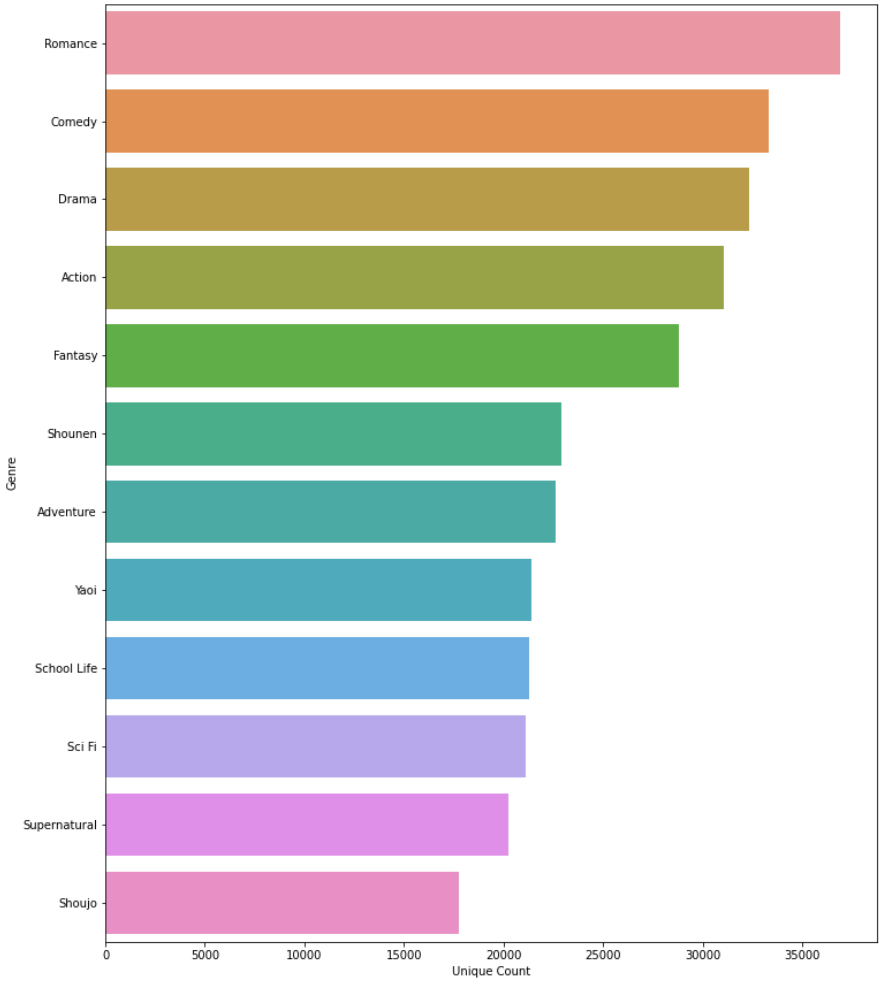
شکل ۵: نمودار تعداد کلمات

	Genre	Count
0	Action	448231
1	Adventure	253509
2	Comedy	531391
3	Drama	506772
4	Fantasy	402722
5	Romance	702923
6	School Life	264173
7	Sci Fi	230653
8	Shoujo	198394
9	Shounen	254330
10	Supernatural	218998
11	Yaoi	482913

شکل ۶: جدول تعداد کلمات

۴.۳ تعداد کلمات منحصر به فرد

کد به دست آوردن تعداد کلمات منحصر به فرد برای هر برچسب در قسمت
`Visualize word count for each genre` فایل `NLP_Project_Word_Broken.ipynb` قرار دارد.
در کل ۶۱۲ برچسب وجود دارد که از این بین ۱۲ برچسب با بیشترین تعداد داکیومنت انتخاب شده‌اند.
خروجی آن در قالب جدول و نمودار در ادامه آمده است.



شکل ۷: نمودار تعداد کلمات منحصر به فرد

	Genre	UCount
0	Action	31032
1	Adventure	22598
2	Comedy	33330
3	Drama	32305
4	Fantasy	28824
5	Romance	36874
6	School Life	21285
7	Sci Fi	21088
8	Shoujo	17764
9	Shounen	22923
10	Supernatural	20244
11	Yaoi	21398

شکل ۸: جدول تعداد کلمات منحصر به فرد

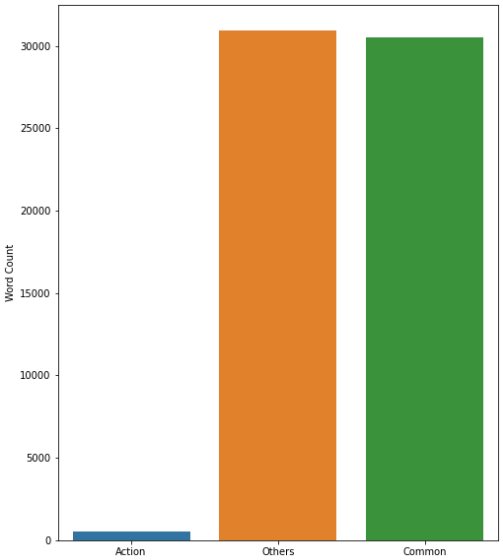
۵.۳ تعداد کلمات منحصر به فرد مشترک و غیر مشترک بین برچسب‌ها

برای به دست آوردن کلمات مشترک و غیرمشترک همان ۱۲ برچسب برتر که از مراحل قبل انتخاب شده‌اند ، استفاده می‌شوند. به منظور به دست آوردن کلمات ، یک برچسب به ترتیب از بین ۱۲ برچسب موجود انتخاب می‌شود و مابقی برچسب‌ها با یکدیگر جمع می‌شوند و تشکیل یک گروه جدید می‌دهند. سپس کلمات مشترک بین این دو گروه و کلمات غیرمشترک به صورت مجزا برای هر دو گروه استخراج می‌شوند.

کد این بخش در فایل `NLP_Project_Analysis` قسمت `Common and non common words between tags` قرار دارد.

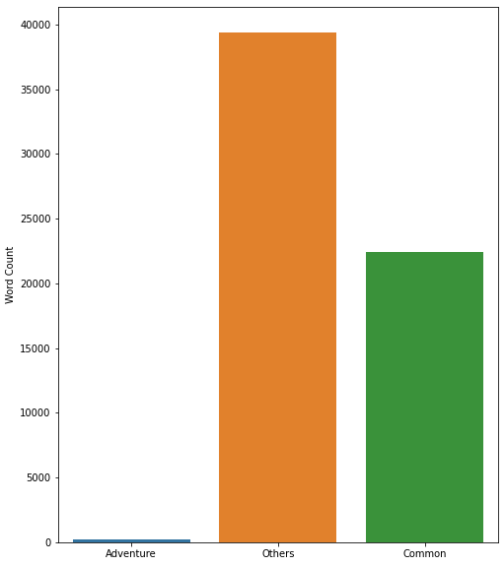
نتایج به دست آمده در ادامه به تفکیک ژانر آورده شده‌اند:

Action ۱.۵.۳



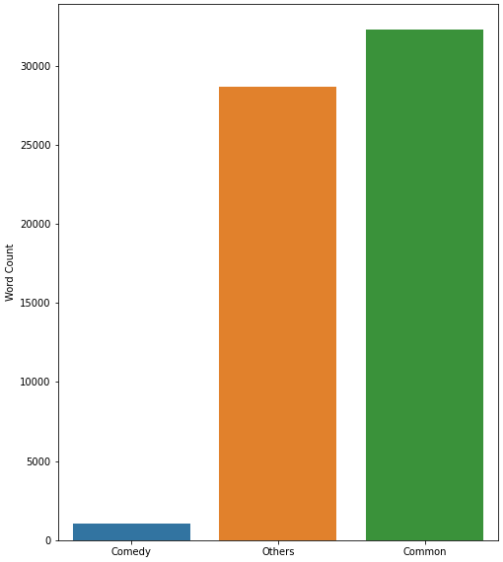
شکل ۹: نمودار تعداد کلمات مشترک و غیرمشترک

Adventure ۲.۵.۳



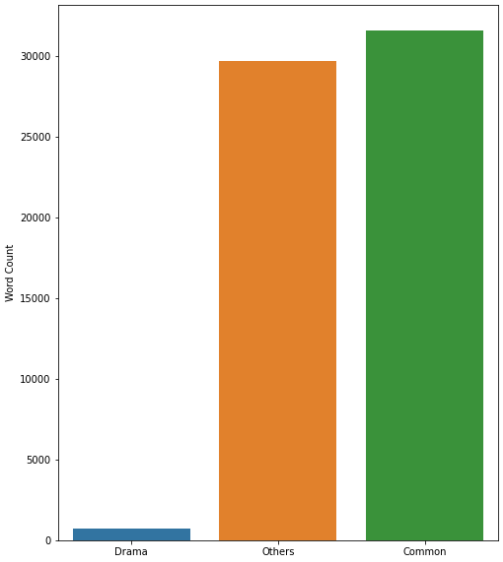
شکل ۱۰: نمودار تعداد کلمات مشترک و غیرمشترک

Comedy ۳.۵.۳



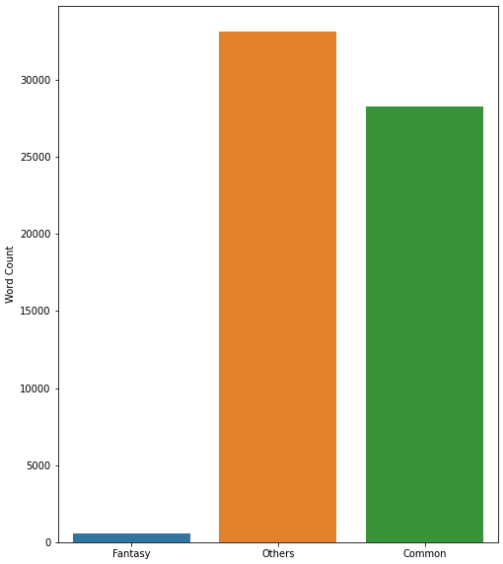
شکل ۱۱: نمودار تعداد کلمات مشترک و غیرمشترک

Drama ۴.۵.۳



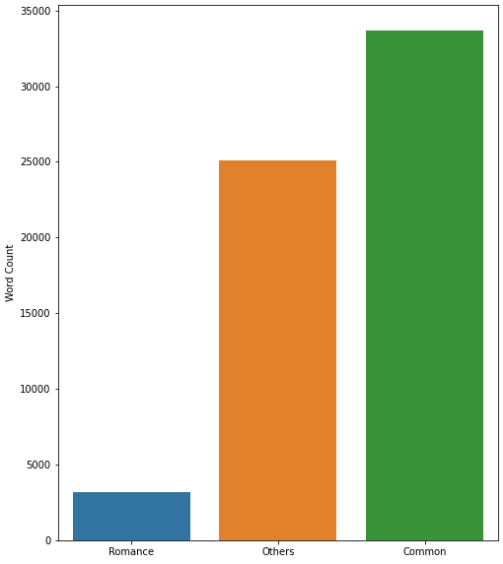
شکل ۱۲: نمودار تعداد کلمات مشترک و غیرمشترک

Fantasy ۵.۵.۳



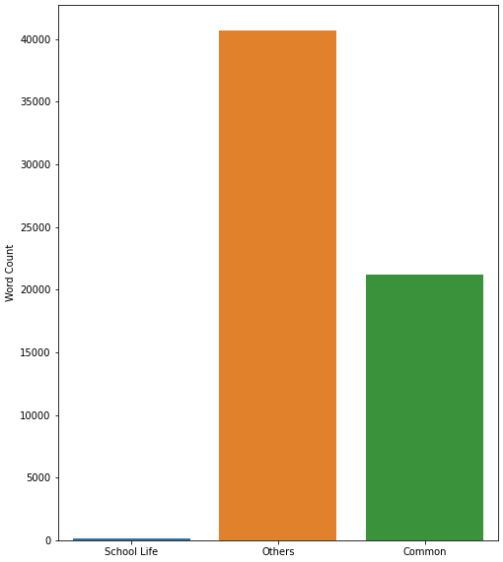
شکل ۱۳: نمودار تعداد کلمات مشترک و غیرمشترک

Romance ۶.۵.۳



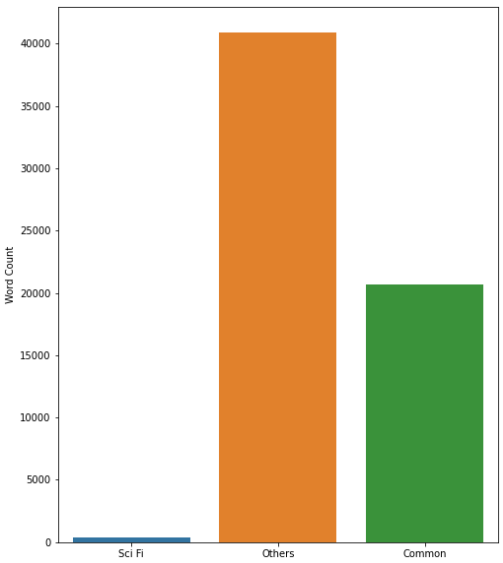
شکل ۱۴: نمودار تعداد کلمات مشترک و غیرمشترک

School Life ۷.۵.۳



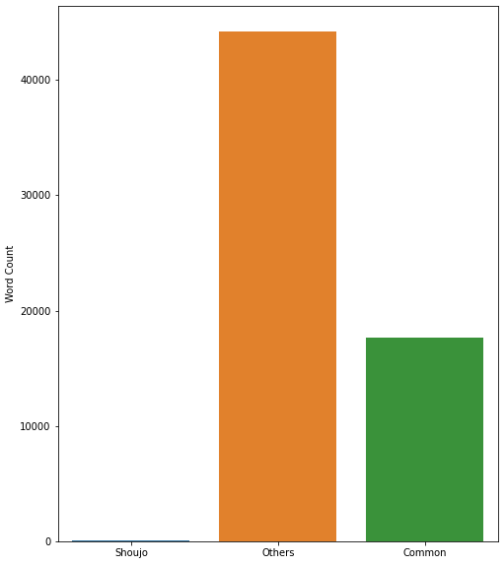
شکل ۱۵: نمودار تعداد کلمات مشترک و غیرمشترک

Sci Fi ۸.۵.۳



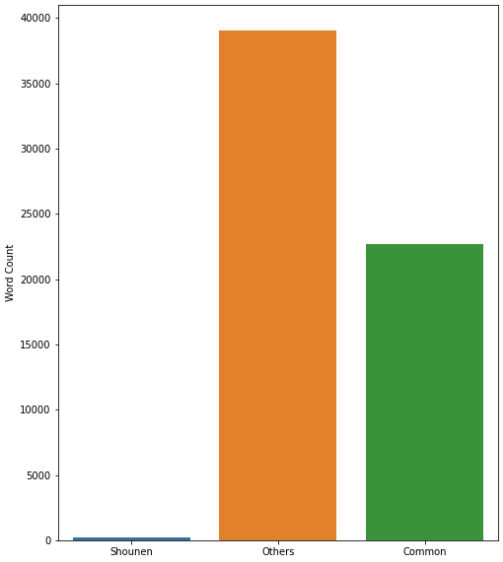
شکل ۱۶: نمودار تعداد کلمات مشترک و غیرمشترک

۹.۵.۳ Shoujo



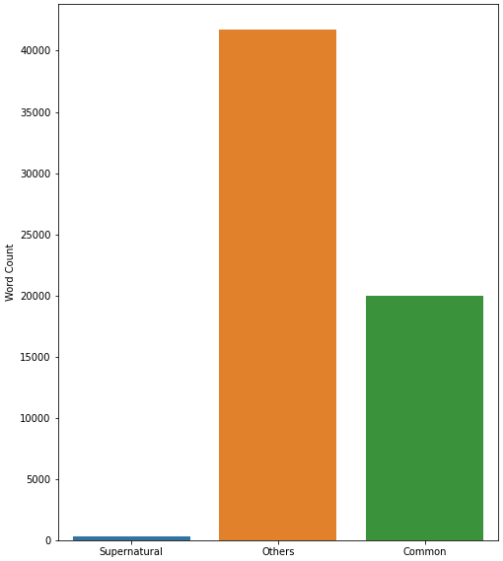
شکل ۱۷: نمودار تعداد کلمات مشترک و غیرمشترک

۱۰.۵.۳ Shounen



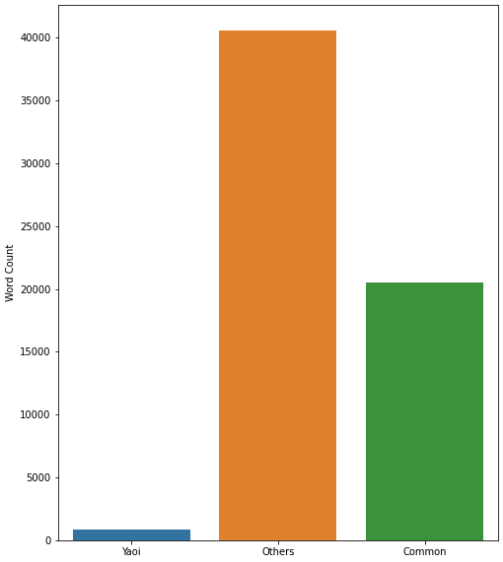
شکل ۱۸: نمودار تعداد کلمات مشترک و غیرمشترک

Supernatural ۱۱.۵.۳



شکل ۱۹: نمودار تعداد کلمات مشترک و غیرمشترک

Yaoi ۱۲.۵.۳



شکل ۲۰: نمودار تعداد کلمات مشترک و غیرمشترک

۱۳.۵.۳ تحلیل نتایج

در خلاصه‌ی انیمه‌ها یا مانگاها معمولا از یک مجموعه کلمات مشخص استفاده می‌شود. به عبارت دیگر برای شرح داستان یک انیمه یا مانگا دایره گسترده‌ای از کلمات نیاز نیست و به همین دلیل تعداد کلمات مشترک بسیار زیاد است و تعداد کلمات غیرمشترک هر برچسب بسیار اندک است. نکته قابل ذکر دیگر این است که هر برچسب با مجموعه کلمات ۱۱ برچسب دیگر مقایسه می‌شود ، از این رو تعداد کلمات غیرمشترک گروه متشکل از ۱۱ برچسب بسیار بیشتر از کلمات غیرمشترک برچسب تکی است.

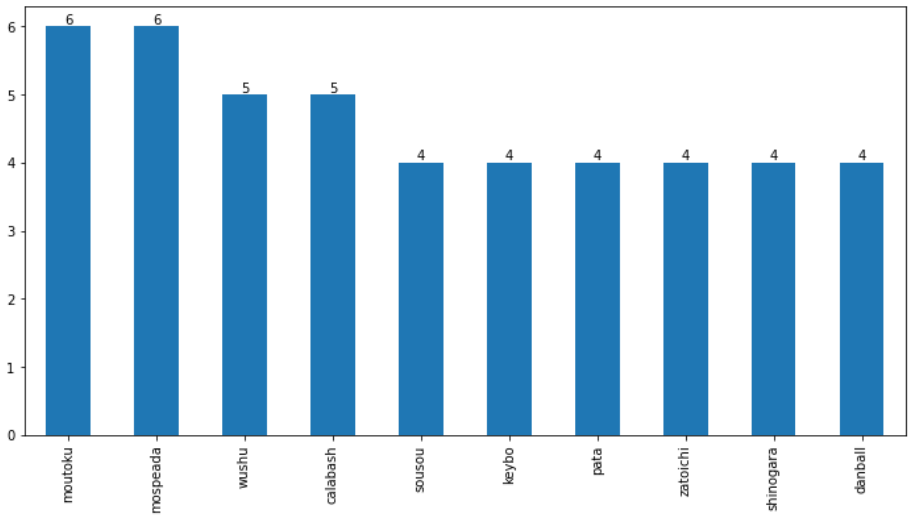
۶.۳ ۱۰ کلمه پرتکرار غیر مشترک هر برچسب

برای هر برچسب با استفاده از `collections Counter` تعداد کلمات هر برچسب به دست می‌آید سپس با استفاده از لیست کلمات مشترک که در مرحله قبل به دست آمده ، این کلمات از Counter حذف می‌شوند و فقط کلمات غیرمشترک با تعداد تکرارشان باقی می‌مانند.

کد این بخش در فایل `NLP_Project_Analysis.ipynb` قسمت `Top 10 unique words of each tag` قرار دارد.

نتایج به دست آمده در ادامه به تفکیک ژانر آورده شده‌اند:

۱.۶.۳ Action

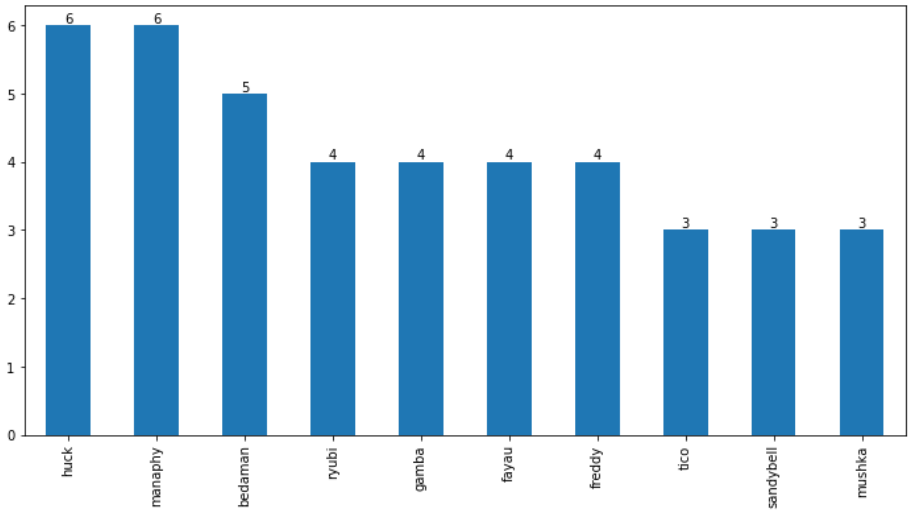


شکل ۲۱: نمودار تعداد کلمات غیرمشترک

	Word	Count
0	moutoku	6
1	mospeada	6
2	wushu	5
3	calabash	5
4	sousou	4
5	keybo	4
6	pata	4
7	zatoichi	4
8	shinogara	4
9	danball	4

شکل ۲۲: جدول تعداد کلمات غیرمشترک

۲.۶.۳ Adventure

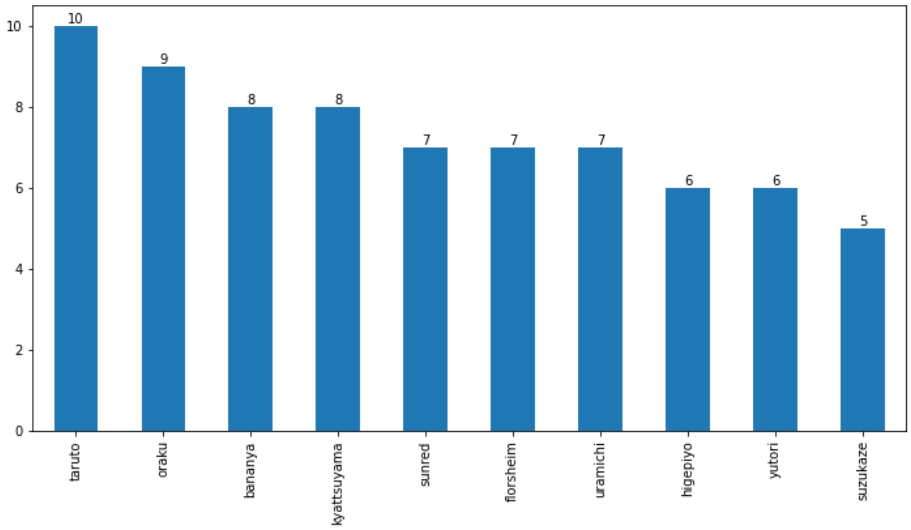


شکل ۲۳: نمودار تعداد کلمات غیرمشترک

	Word	Count
0	huck	6
1	manaphy	6
2	bedaman	5
3	ryubi	4
4	gamba	4
5	fayau	4
6	freddy	4
7	tico	3
8	sandybell	3
9	mushka	3

شکل ۲۴: جدول تعداد کلمات غیرمشترک

۳.۶.۳ Comedy

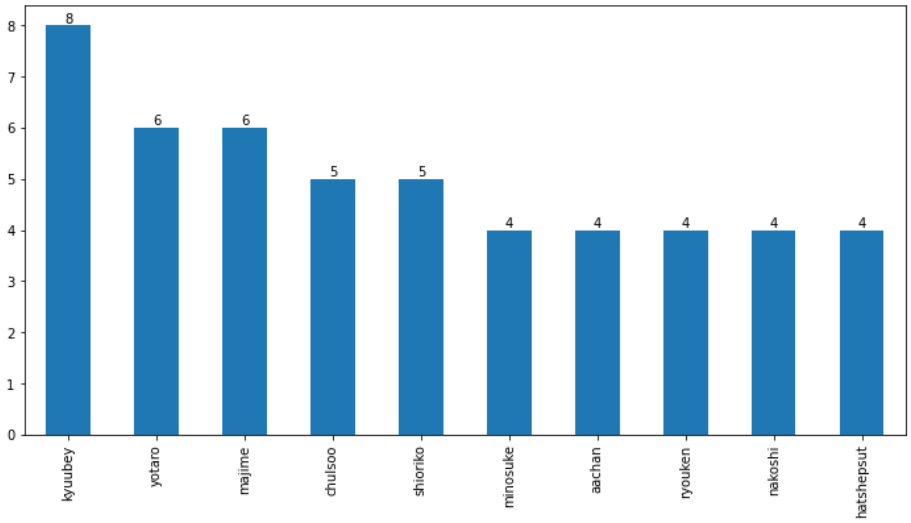


شکل ۲۵: نمودار تعداد کلمات غیرمشترک

	Word	Count
0	taruto	10
1	oraku	9
2	bananya	8
3	kyattsuyama	8
4	sunred	7
5	florsheim	7
6	uramichi	7
7	higepiyo	6
8	yutori	6
9	suzukaze	5

شکل ۲۶: جدول تعداد کلمات غیرمشترک

۴.۶.۳ Drama

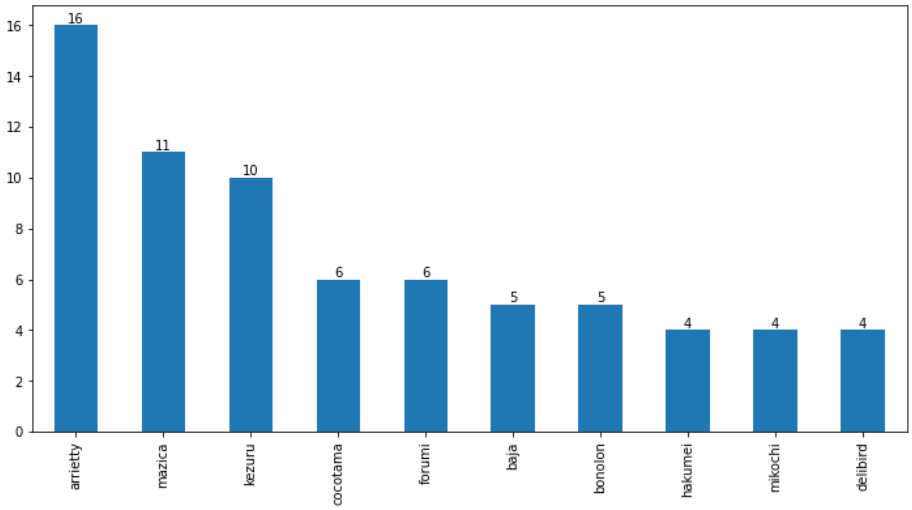


شکل ۲۷: نمودار تعداد کلمات غیرمشترک

	Word	Count
0	kyuubey	8
1	yotaro	6
2	majime	6
3	chulsoo	5
4	shioriko	5
5	minosuke	4
6	aachan	4
7	ryouken	4
8	nakoshi	4
9	hatshepsut	4

شکل ۲۸: جدول تعداد کلمات غیرمشترک

۵.۶.۳ Fantasy

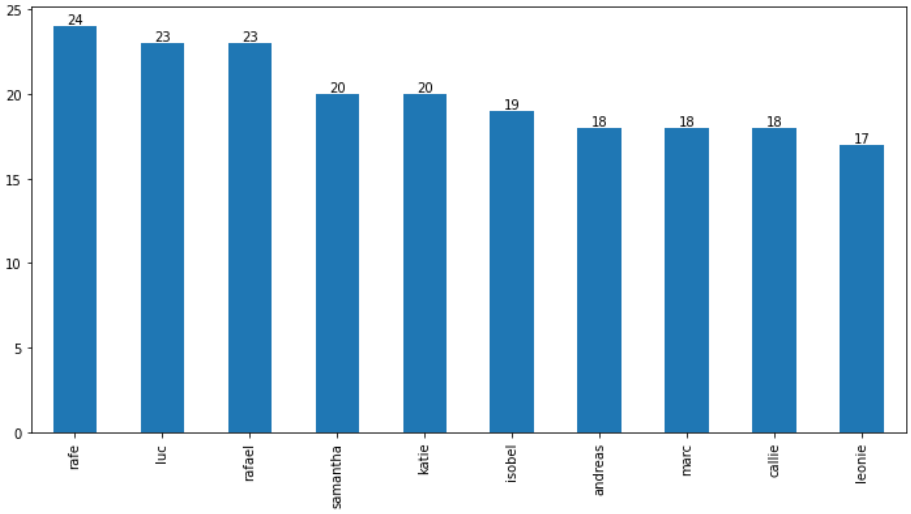


شکل ۲۹: نمودار تعداد کلمات غیرمشترک

	Word	Count
0	arrietty	16
1	mazica	11
2	kezuru	10
3	cocotama	6
4	forumi	6
5	baja	5
6	bonolon	5
7	hakumei	4
8	mikochi	4
9	delibird	4

شکل ۳۰: جدول تعداد کلمات غیرمشترک

۶.۶.۳ Romance

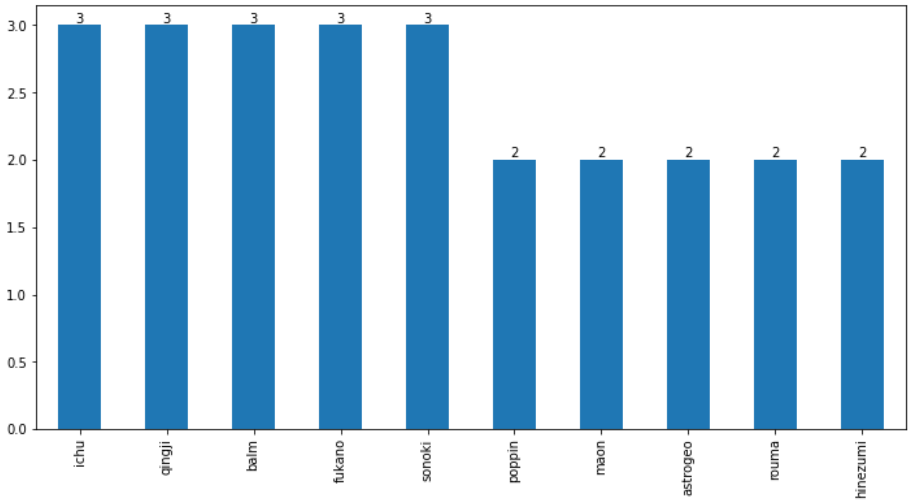


شکل ۳۱: نمودار تعداد کلمات غیرمشترک

	Word	Count
0	rafe	24
1	luc	23
2	rafael	23
3	samantha	20
4	katie	20
5	isobel	19
6	andreas	18
7	marc	18
8	callie	18
9	leonie	17

شکل ۳۲: جدول تعداد کلمات غیرمشترک

School Life ۷.۶.۳

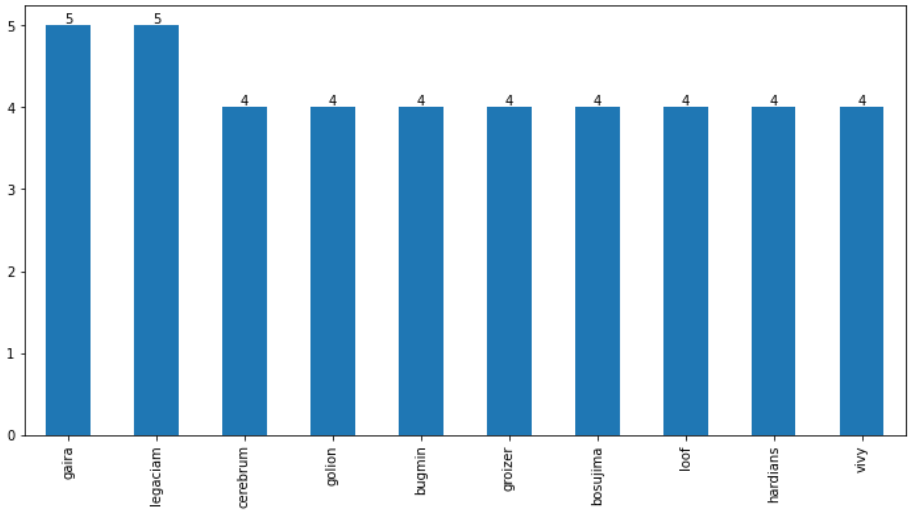


شکل ۳۳: نمودار تعداد کلمات غیرمشترک

	Word	Count
0	ichu	3
1	qingji	3
2	balm	3
3	fukano	3
4	sonoki	3
5	poppin	2
6	maon	2
7	astrogeo	2
8	rouma	2
9	hinezumi	2

شکل ۳۴: جدول تعداد کلمات غیرمشترک

Sci Fi ۸.۶.۳

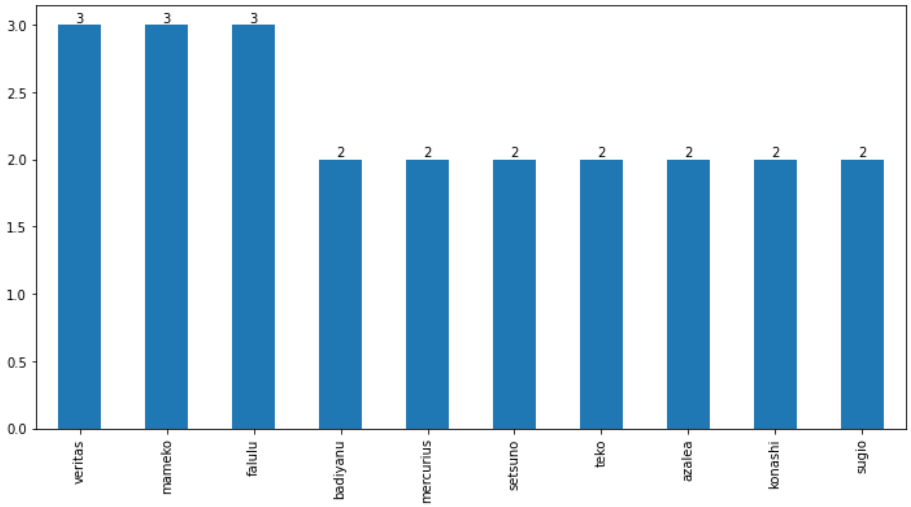


شکل ۳۵: نمودار تعداد کلمات غیرمشترک

	Word	Count
0	gaira	5
1	legaciam	5
2	cerebrum	4
3	golion	4
4	bugmin	4
5	groizer	4
6	bosujima	4
7	loof	4
8	hardians	4
9	vivy	4

شکل ۳۶: جدول تعداد کلمات غیرمشترک

۹.۶.۳ Shoujo

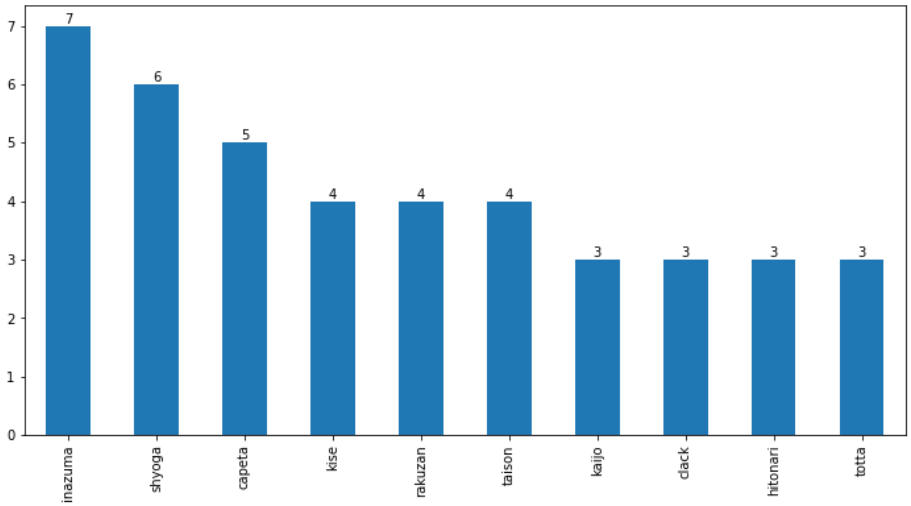


شکل ۳۷: نمودار تعداد کلمات غیرمشترک

	Word	Count
0	veritas	3
1	mameko	3
2	falulu	3
3	badiyanu	2
4	mercurius	2
5	setsuno	2
6	teko	2
7	azalea	2
8	konashi	2
9	sugio	2

شکل ۳۸: جدول تعداد کلمات غیرمشترک

۱۰.۶.۳ Shounen

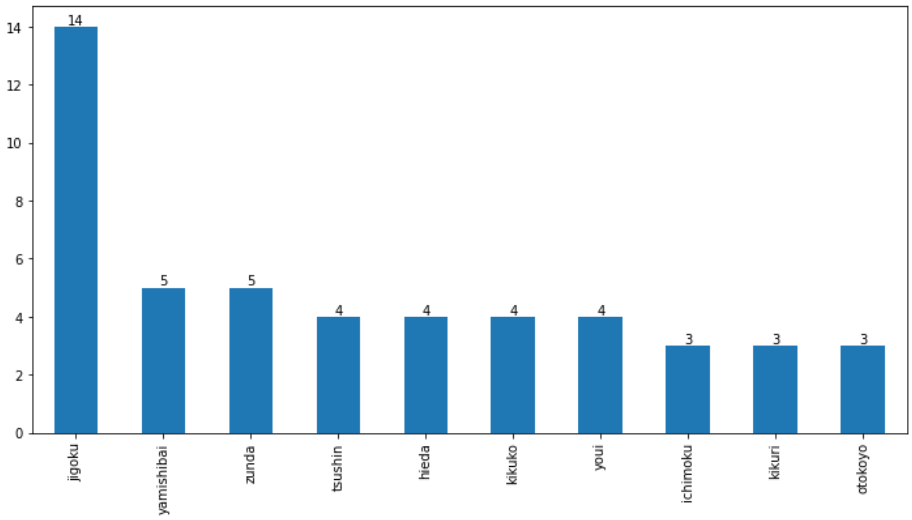


شکل ۳۹: نمودار تعداد کلمات غیرمشترک

	Word	Count
0	inazuma	7
1	shyoga	6
2	capeta	5
3	kise	4
4	rakuzan	4
5	taison	4
6	kaijo	3
7	clack	3
8	hitonari	3
9	totta	3

شکل ۴۰: جدول تعداد کلمات غیرمشترک

۱۱.۶.۳ Supernatural

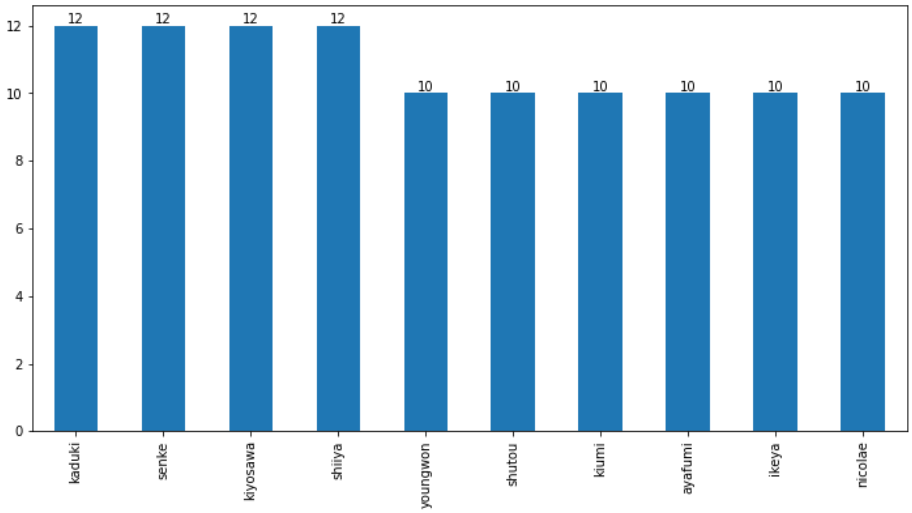


شکل ۴۱: نمودار تعداد کلمات غیرمشترک

	Word	Count
0	jigoku	14
1	yamishibai	5
2	zunda	5
3	tsushin	4
4	hieda	4
5	kikuko	4
6	youi	4
7	ichimoku	3
8	kikuri	3
9	otokoyo	3

شکل ۴۲: جدول تعداد کلمات غیرمشترک

Yaoi ۱۲.۶.۳



شکل ۴۳: نمودار تعداد کلمات غیرمشترک

	Word	Count
0	kaduki	12
1	senke	12
2	kiyosawa	12
3	shiiya	12
4	youngwon	10
5	shutou	10
6	kiumi	10
7	ayafumi	10
8	ikeya	10
9	nicolae	10

شکل ۴۴: جدول تعداد کلمات غیرمشترک

۱۳.۶.۳ تحلیل نتایج

کلمات غیرمشترک معمولا اسامی خاص هستند و از آن جایی که شرح داستان یک انیمه یا مانگا دایره لغات گسترده‌ای ندارد ، تعداد تکرار کلمات غیرمشترک بسیار کم است.

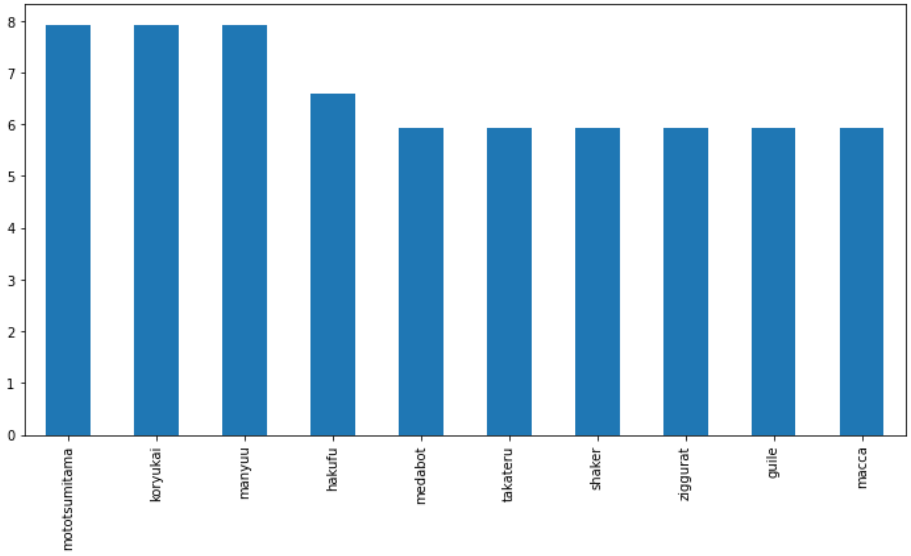
۷.۳ ۱۰ کلمه مشترک برتر هر برچسب نسبت به برچسب‌های دیگر بر اساس Relative Normalized Frequency

با کمک کلمات مشترک به دست آمده در قسمت‌های قبلی ، در این بخش با استفاده از `collections Counter` تعداد تکرار کلمات هر برچسب و گروه متشکل از ۱۱ برچسب دیگر شمرده می‌شود. سپس تعداد به دست آمده در فرمول آورده شده قرار داده می‌شود و Relative Normalized Frequency همه کلمات محاسبه می‌شود. ۱۰ کلمه که بیشترین مقدار را دارند انتخاب می‌شوند.

کد این بخش در فایل `NLP_Project_Analysis.ipynb` قسمت `Top 10 common words for each tag` قرار دارد.

نتایج به دست آمده در ادامه به تفکیک ژانر آورده شده‌اند:

Action ۱.۷.۳

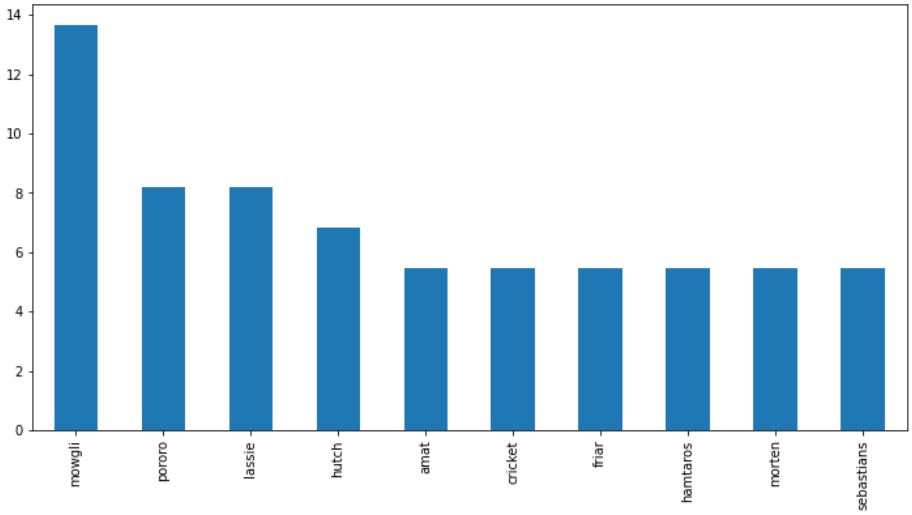


شکل ۴۵: نمودار کلمات مشترک برتر

	Word	Relative Normalized Frequency
0	mototsumitama	7.9203402938901775
1	koryukai	7.9203402938901775
2	manyuu	7.9203402938901775
3	hakufu	6.600283578241815
4	medabot	5.940255220417633
5	takateru	5.940255220417633
6	shaker	5.940255220417633
7	ziggurat	5.940255220417633
8	guile	5.940255220417633
9	macca	5.940255220417633

شکل ۴۶: جدول کلمات مشترک برتر

Adventure ۲.۷.۳

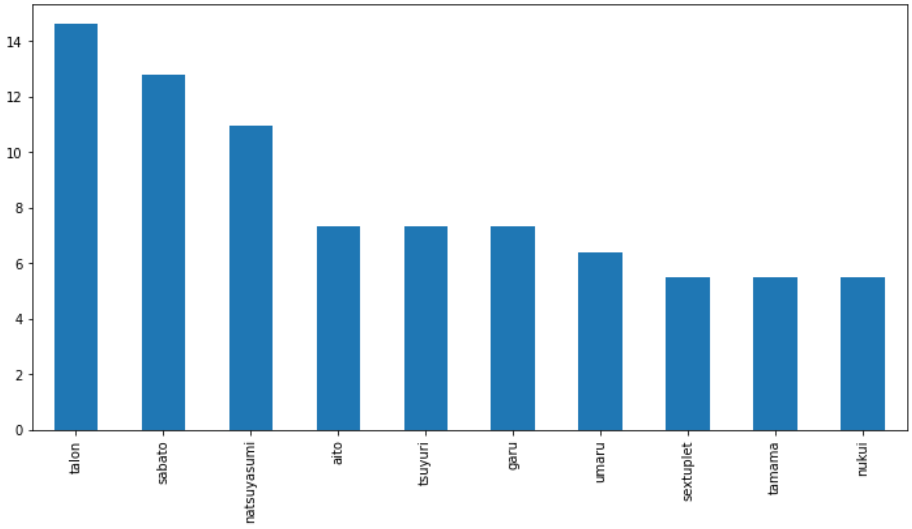


شکل ۴۷: نمودار کلمات مشترک برتر

	Word	Relative Normalized Frequency
0	mowgli	13.666253650765555
1	pororo	8.199752190459332
2	lassie	8.199752190459332
3	hutch	6.833126825382777
4	amat	5.4665014603062225
5	cricket	5.4665014603062225
6	friar	5.4665014603062225
7	hamtaros	5.4665014603062225
8	morten	5.4665014603062225
9	sebastians	5.4665014603062225

شکل ۴۸: جدول کلمات مشترک برتر

Comedy ۳.۷.۳

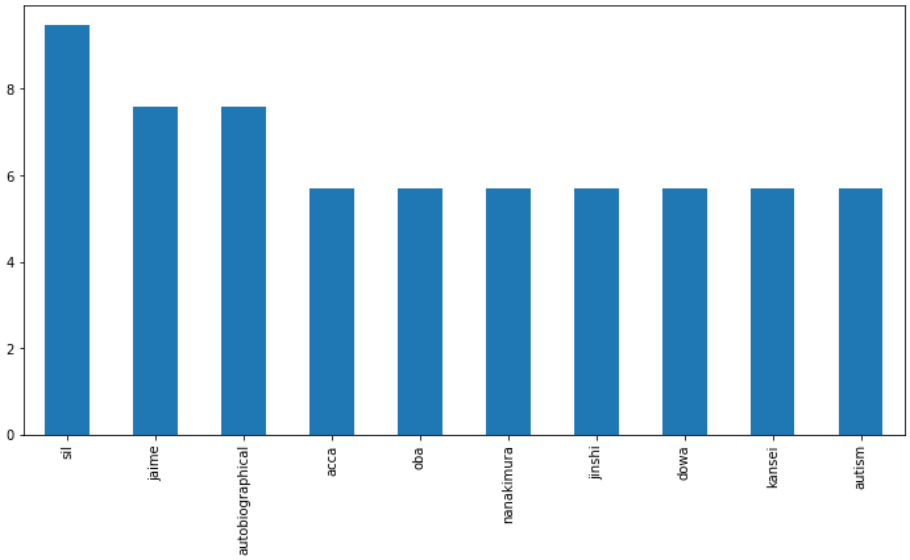


شکل ۴۹: نمودار کلمات مشترک برتر

	Word	Relative Normalized Frequency
0	talon	14.620822082208221
1	sabato	12.793219321932193
2	natsuyasumi	10.965616561656166
3	aito	7.3104110411041106
4	tsuyuri	7.3104110411041106
5	garu	7.3104110411041106
6	umaru	6.396609660966097
7	sextuplet	5.482808280828083
8	tamama	5.482808280828083
9	nukui	5.482808280828083

شکل ۵۰: جدول کلمات مشترک برتر

Drama ۴.۷.۳

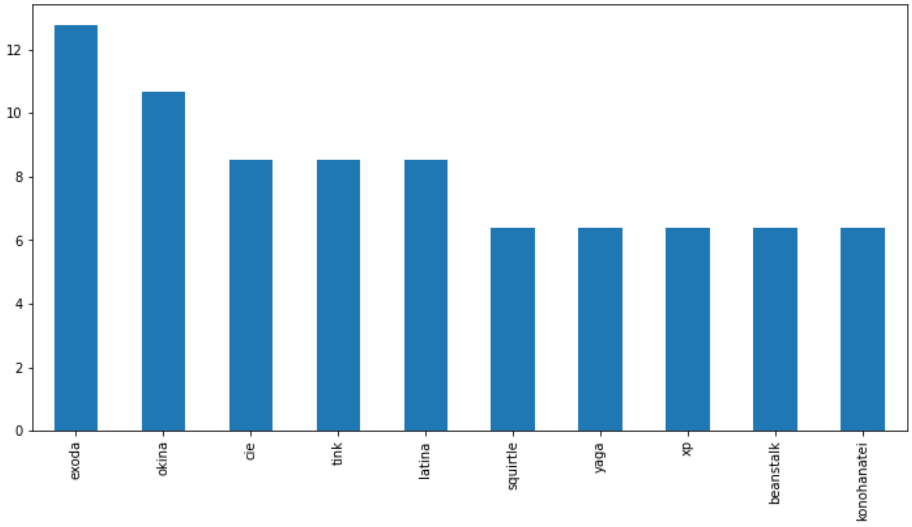


شکل ۵۱: نمودار کلمات مشترک برتر

	Word	Relative Normalized Frequency
0	sil	9.478408915028634
1	jaime	7.582727132022907
2	autobiographical	7.582727132022907
3	acca	5.687045349017179
4	oba	5.687045349017179
5	nanakimura	5.687045349017179
6	jinshi	5.687045349017179
7	dowa	5.687045349017179
8	kansei	5.687045349017179
9	autism	5.687045349017179

شکل ۵۲: جدول کلمات مشترک برتر

Fantasy ۵.۷.۳

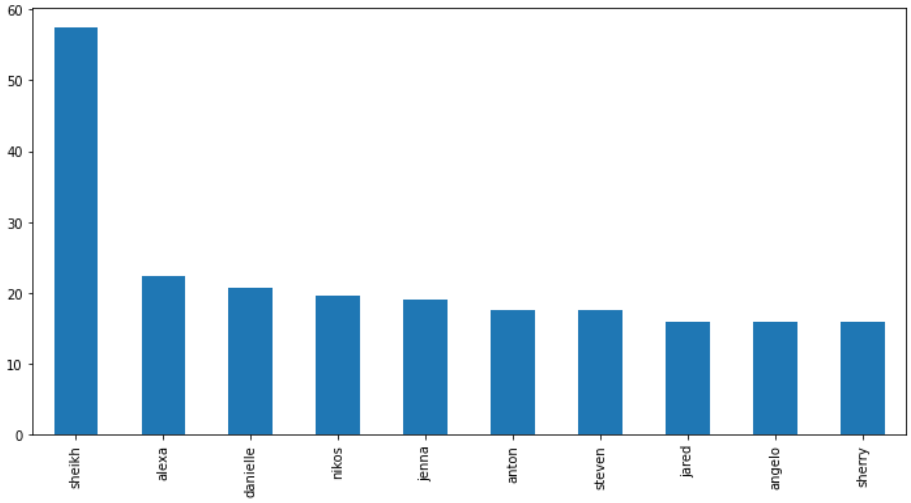


شکل ۵۳: نمودار کلمات مشترک برتر

	Word	Relative Normalized Frequency
0	exoda	12.781640299750208
1	okina	10.651366916458507
2	cie	8.521093533166807
3	tink	8.521093533166807
4	latina	8.521093533166805
5	squirtle	6.390820149875104
6	yaga	6.390820149875104
7	xp	6.390820149875104
8	beanstalk	6.390820149875104
9	konohanatei	6.390820149875104

شکل ۵۴: جدول کلمات مشترک برتر

Romance ۶.۷.۳

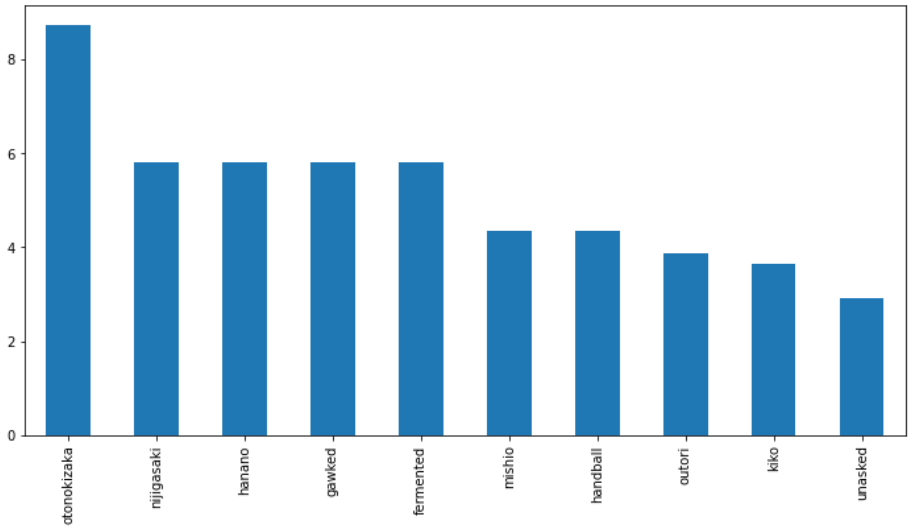


شکل ۵۵: نمودار کلمات مشترک برتر

	Word	Relative Normalized Frequency
0	sheikh	57.37896620925313
1	alexa	22.31404241470955
2	danielle	20.720182242230297
3	nikos	19.657608793910796
4	jenna	19.126322069751044
5	anton	17.53246189727179
6	steven	17.53246189727179
7	jared	15.938601724792537
8	angelo	15.938601724792537
9	sherry	15.938601724792537

شکل ۵۶: جدول کلمات مشترک برتر

School Life ۷.۷.۳

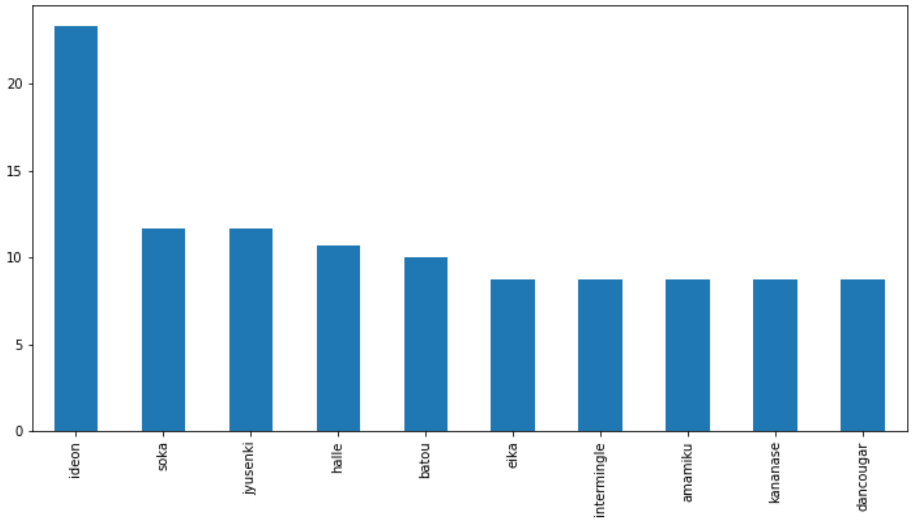


شکل ۵۷: نمودار کلمات مشترک برتر

	Word	Relative Normalized Frequency
0	otonokizaka	8.717265680056379
1	nijigasaki	5.811510453370919
2	hanano	5.811510453370919
3	gawked	5.811510453370919
4	fermented	5.811510453370919
5	mishio	4.358632840028189
6	handball	4.358632840028189
7	outori	3.874340302247279
8	kiko	3.6321940333568246
9	unasked	2.9057552266854594

شکل ۵۸: جدول کلمات مشترک برتر

Sci Fi ۸.۷.۳

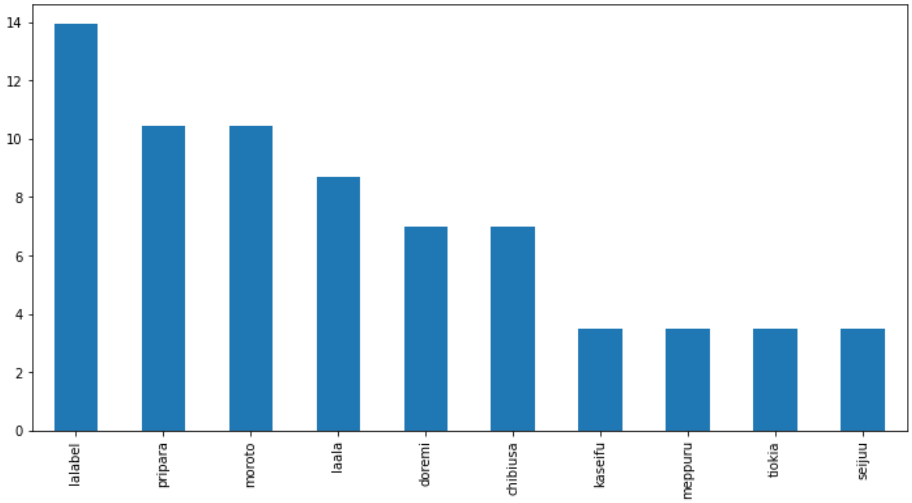


شکل ۵۹: نمودار کلمات مشترک برتر

	Word	Relative Normalized Frequency
0	ideon	23.35015174506829
1	soka	11.675075872534144
2	jyusenki	11.675075872534144
3	halle	10.702152883156296
4	batou	10.007207890743551
5	eika	8.756306904400608
6	intermingle	8.756306904400608
7	amamiku	8.756306904400608
8	kananase	8.756306904400608
9	dancougar	8.756306904400608

شکل ۶۰: جدول کلمات مشترک برتر

۹.۷.۳ Shoujo

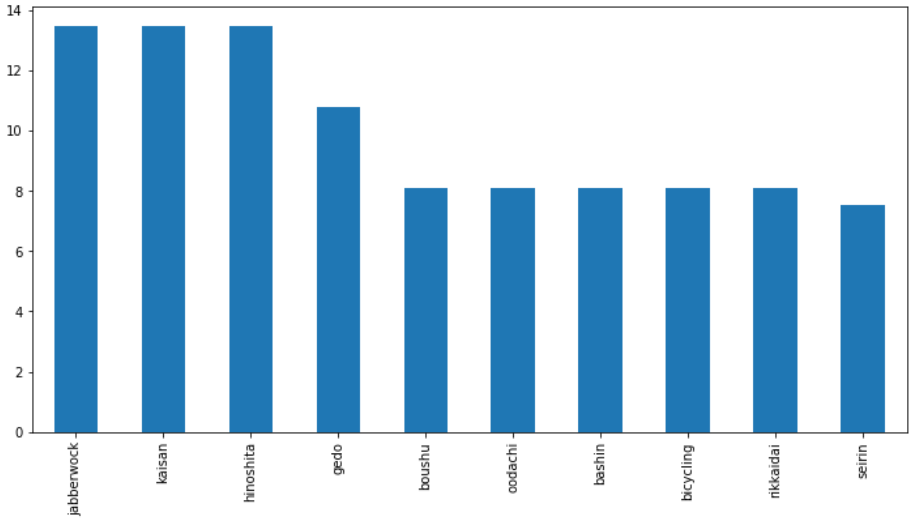


شکل ۶۱: نمودار کلمات مشترک برتر

	Word	Relative Normalized Frequency
0	lalabel	13.935600090069803
1	pripara	10.451700067552354
2	moroto	10.451700067552354
3	laala	8.709750056293627
4	doremi	6.967800045034902
5	chibiusa	6.967800045034902
6	kaseifu	3.4839000225174517
7	meppuru	3.4839000225174517
8	tiokia	3.4839000225174517
9	seijuu	3.4839000225174517

شکل ۶۲: جدول کلمات مشترک برتر

Shounen ۱۰.۷.۳

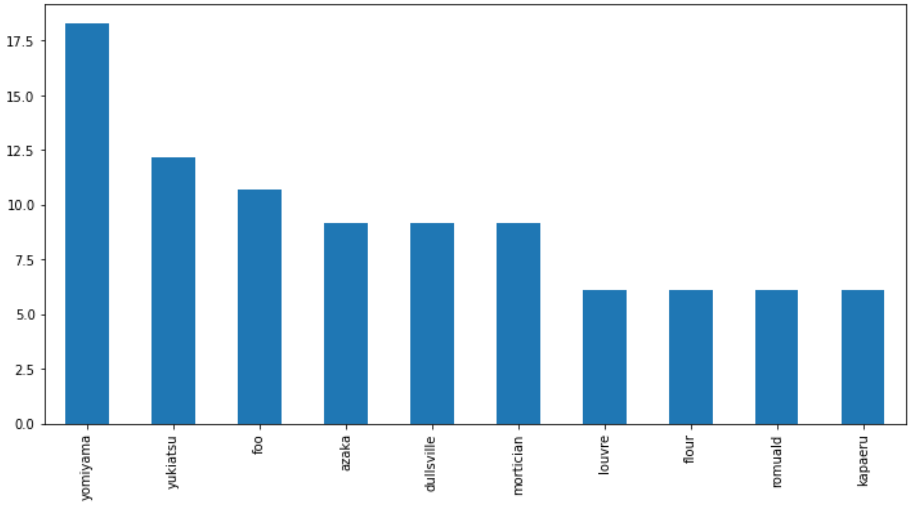


شکل ۶۳: نمودار کلمات مشترک برتر

	Word	Relative Normalized Frequency
0	jabberwock	13.46725995724818
1	kaisan	13.46725995724818
2	hinoshita	13.46725995724818
3	gedo	10.773807965798543
4	boushu	8.080355974348908
5	oodachi	8.080355974348908
6	bashin	8.080355974348908
7	bicycling	8.080355974348908
8	rikkaidai	8.080355974348908
9	seirin	7.541665576058981

شکل ۶۴: جدول کلمات مشترک برتر

۱۱.۷.۳ Supernatural

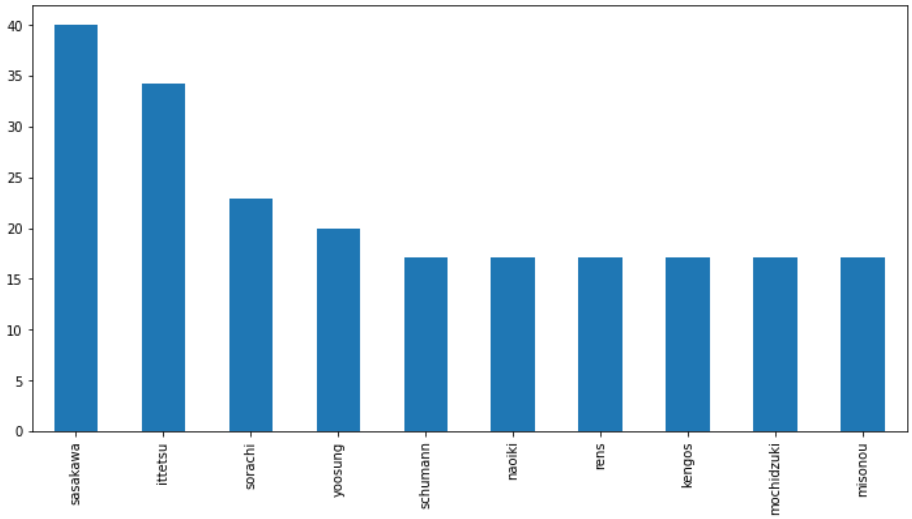


شکل ۶۵: نمودار کلمات مشترک برتر

	Word	Relative Normalized Frequency
0	yomiyama	18.286603438055717
1	yukiatsu	12.191068958703813
2	foo	10.667185338865837
3	azaka	9.143301719027859
4	dullsville	9.143301719027859
5	mortician	9.143301719027859
6	louvre	6.095534479351906
7	flour	6.095534479351906
8	romuald	6.095534479351906
9	kapaeru	6.095534479351906

شکل ۶۶: جدول کلمات مشترک برتر

Yaoi ۱۲.۷.۳



شکل ۶۷: نمودار کلمات مشترک برتر

	Word	Relative Normalized Frequency
0	sasakawa	39.99074679876624
1	ittetsu	34.277782970371064
2	sorachi	22.85185531358071
3	yoosung	19.99537339938312
4	schumann	17.138891485185532
5	naoiki	17.138891485185532
6	rens	17.138891485185532
7	kengos	17.138891485185532
8	mochidzuki	17.138891485185532
9	misonou	17.138891485185532

شکل ۶۸: جدول کلمات مشترک برتر

۸.۳ ۱۰ کلمه برتر هر برچسب بر اساس TF-IDF

در این قسمت از جملات تفکیک شده هر برچسب استفاده می‌شود. داکیومنت مربوط به ۱۲ برچسب به عنوان ورودی به `CountVectorizer` داده می‌شود و با کمک `TfidfTransformer` مقادیر TF-IDF برای همه

کلمات همه داکیومنت‌ها حساب می‌شود. هر سطر از آرایه خروجی بیانگر مقادیر TF-IDF برای داکیومنت یک برچسب است.

کد این بخش در فایل `NLP_Project_Analysis.ipynb` قسمت `TF-IDF` قرار دارد.

نتایج به دست آمده در ادامه به تفکیک ژانر آورده شده‌اند:

۱.۸.۳ Action

	tfidf
world	0.35514917085941217
num	0.28493164180671554
one	0.25900036300193774
life	0.17737931977605564
new	0.15191667854003887
school	0.13684217007822524
however	0.11637832957565959
girl	0.1160659045298189
day	0.11536294817667732
years	0.11426946051623489

شکل ۶۹: جدول TFIDF

۲.۸.۳ Adventure

	tfidf
world	0.4288628102609746
one	0.24907694880773318
num	0.24417490247481502
life	0.1772685944174186
new	0.14070197852862373
young	0.1209613054582236
however	0.11990140354840347
day	0.11513184495421283
time	0.11234960244093496
find	0.1111572127923873

شکل ۷۰: جدول TFIDF

Comedy ۳.۸.۳

	tfidf
school	0.3195698320189088
one	0.27676634352182494
num	0.2105349101372839
girl	0.20306329528128403
life	0.1965414619747757
world	0.18520740215084372
day	0.1695676659692169
love	0.1642488892920645
high	0.16152618218352216
new	0.14411352044284456

شکل ۷۱: جدول TFIDF

Drama ۴.۸.۳

	tfidf
one	0.292904622583894
school	0.2288972830217864
num	0.2207102977289587
life	0.2152297538552476
love	0.20372737782400205
world	0.16604018106280338
day	0.16461929931776717
girl	0.1566352971313732
time	0.13207434125289005
man	0.1282176622306489

شکل ۷۲: جدول TFIDF

۵.۸.۳ Fantasy

	tfidf
world	0.4006381032461344
one	0.2728505010966556
life	0.22028069840126846
num	0.18068803253477406
girl	0.15274999273754
day	0.14663338047424024
new	0.12456398217287488
however	0.11737282991737381
young	0.11290935610361452
time	0.11241341456875238

شکل ۷۳: جدول TFIDF

۶.۸.۳ Romance

	tfidf
one	0.2934998974894924
love	0.2779649312707669
school	0.25665058773852745
life	0.1962214185484256
day	0.17883705158937563
girl	0.1775424710711485
num	0.1563668325944334
man	0.14332855737514594
high	0.1340353186550155
time	0.13153862765557747

شکل ۷۴: جدول TFIDF

School Life ۷.۸.۳

	tfidf
school	0.5451188191932779
high	0.24259979620934083
one	0.24006662564651238
love	0.21229917909243118
girl	0.1985616002709384
day	0.16767640533183759
student	0.16173319747289389
life	0.15218509304377123
new	0.11711042371230027
num	0.11613612734198163

شکل ۷۵: جدول TFIDF

Sci Fi ۸.۸.۳

	tfidf
num	0.3599735969921317
world	0.2842279410822365
one	0.2417754813400303
earth	0.20644689326689297
new	0.17228138628258322
life	0.14756591314499745
girl	0.1235773656879289
time	0.11965196701313587
years	0.11747118997158418
planet	0.10700346017213608

شکل ۷۶: جدول TFIDF

۹.۸.۳ Shoujo

	tfidf
school	0.37144864918376636
one	0.2766571671694065
love	0.2654452714472779
girl	0.22132573373552522
num	0.18623395621561628
day	0.1817200761196944
high	0.17094500750362276
life	0.14575464438767155
boy	0.13046569567567803
new	0.11415748371621827

شکل ۷۷: جدول TFIDF

۱۰.۸.۳ Shounen

	tfidf
school	0.27443448440222495
one	0.273489090016981
num	0.2441818640744206
world	0.24377669505217323
girl	0.17719391906285 39
new	0.1686853695956589
life	0.16260783426194825
day	0.14532062264606005
high	0.14369994655707052
however	0.11776912913323825

شکل ۷۸: جدول TFIDF

۱۱.۸.۳ Supernatural

	tfidf
one	0.30601240722070355
school	0.2252889738214608
world	0.21761564365043012
num	0.1947491197407587
girl	0.19398178672365562
life	0.1915263210689258
day	0.17495192789949954
new	0.11862968444413434
time	0.11648115199624574
high	0.11586728558256329

شکل ۷۹: جدول TFIDF

۱۲.۸.۳ Yaoi

	tfidf
one	0.32595827659609633
love	0.29289184048243455
day	0.19827477235195676
school	0.1943117537915179
man	0.1815557878001053
life	0.15195699292682757
two	0.13375187641481157
time	0.1315226784745647
like	0.12235819805354983
student	0.11889055681316583

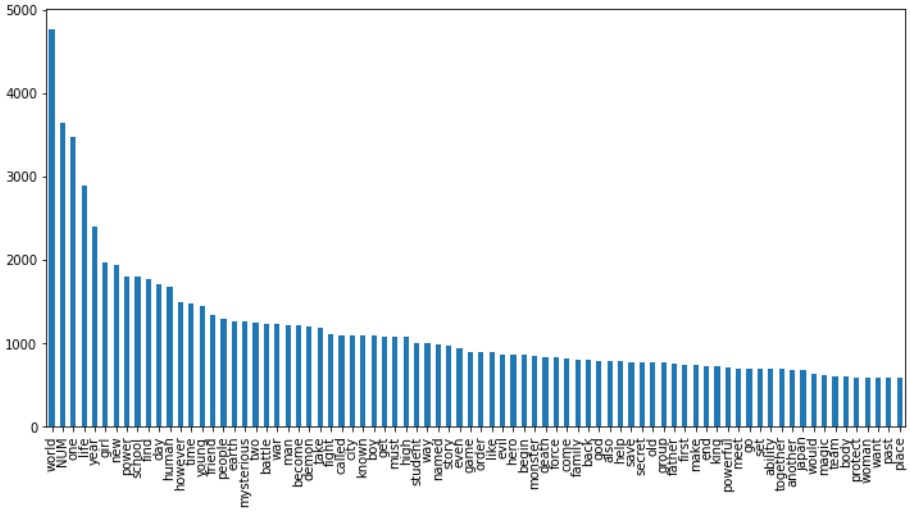
شکل ۸۰: جدول TFIDF

۹.۳ هیستوگرام تعداد تکرار هر کلمه منحصر به فرد به ترتیب از فرکانس بالا به پایین

برای هر برچسب با استفاده از یک `collections Counter` می‌توان تعداد کلمات را به دست آورد.
کد این بخش در فایل `NLP_Project_Analysis.ipynb` قسمت `Word frequency histogram` قرار دارد.
از آن جایی که تعداد کلمات منحصر به فرد هر برچسب زیاد است ، هیستوگرام برای ۸۰ کلمه پرتکرار رسم شده است.

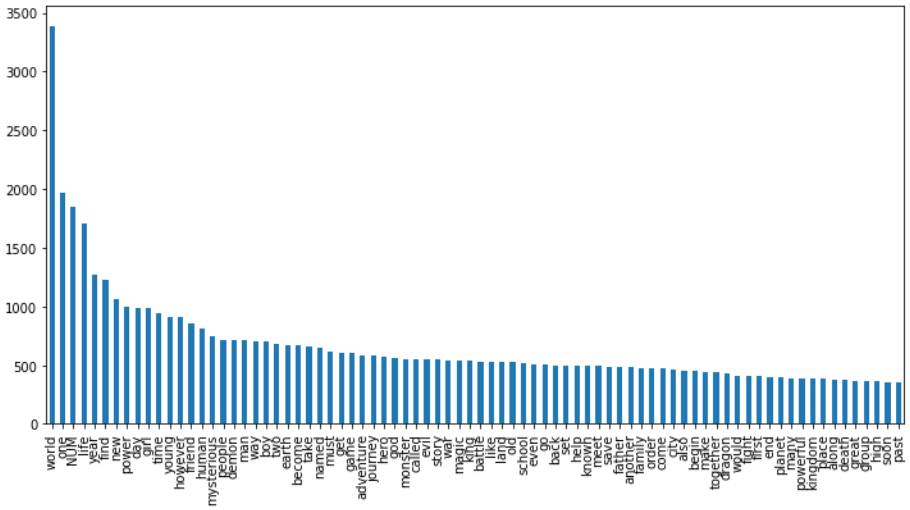
نتایج به دست آمده در ادامه به تفکیک ژانر آورده شده‌اند:

Action ۱.۹.۳



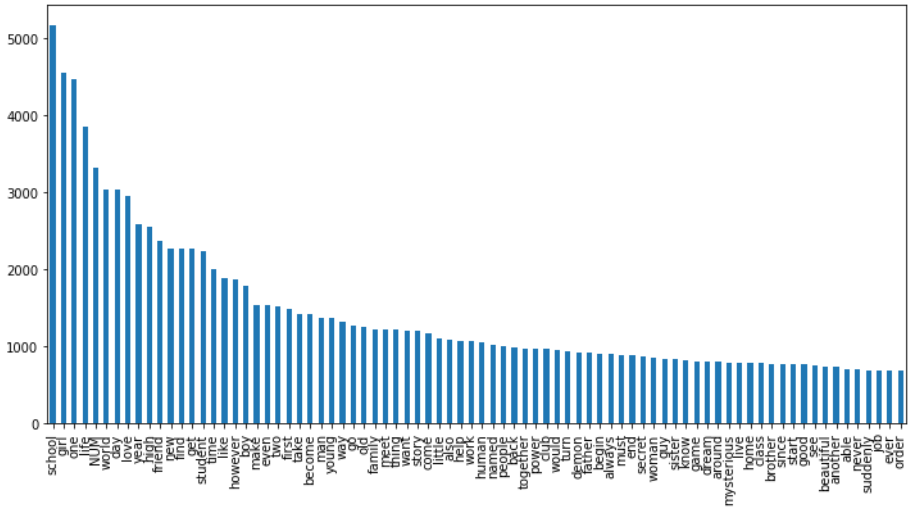
شکل ۸۱: هیستوگرام تعداد تکرار کلمات منحصر به فرد

Adventure ۲.۹.۳



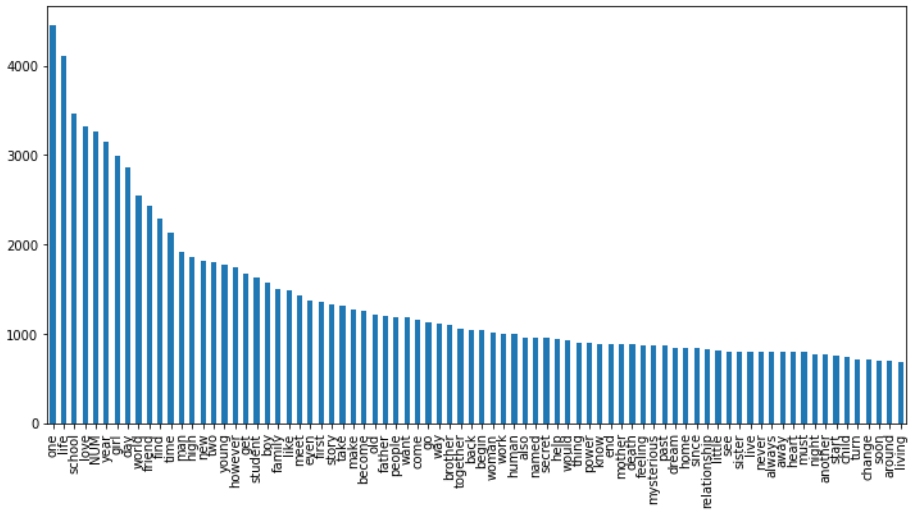
شکل ۸۲: هیستوگرام تعداد تکرار کلمات منحصر به فرد

Comedy ۳.۹.۳



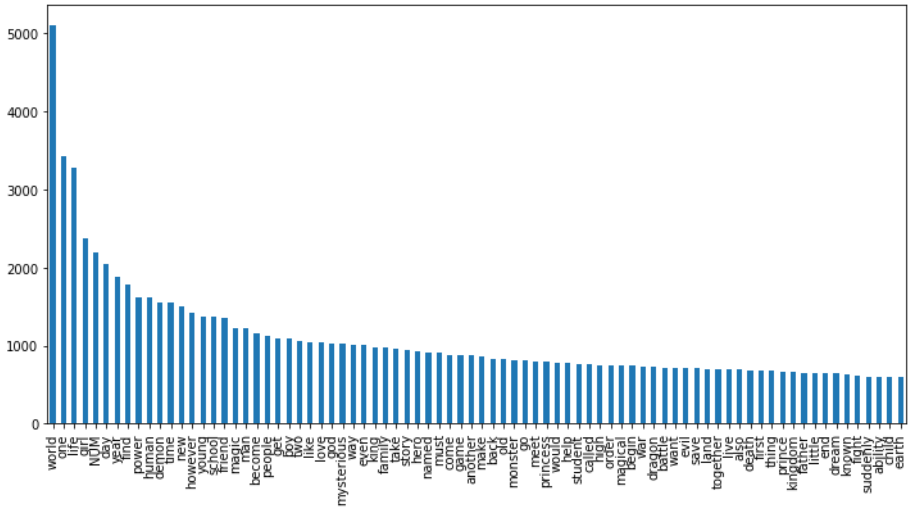
شکل ۸۳: هیستوگرام تعداد تکرار کلمات منحصر به فرد

Drama ۴.۹.۳



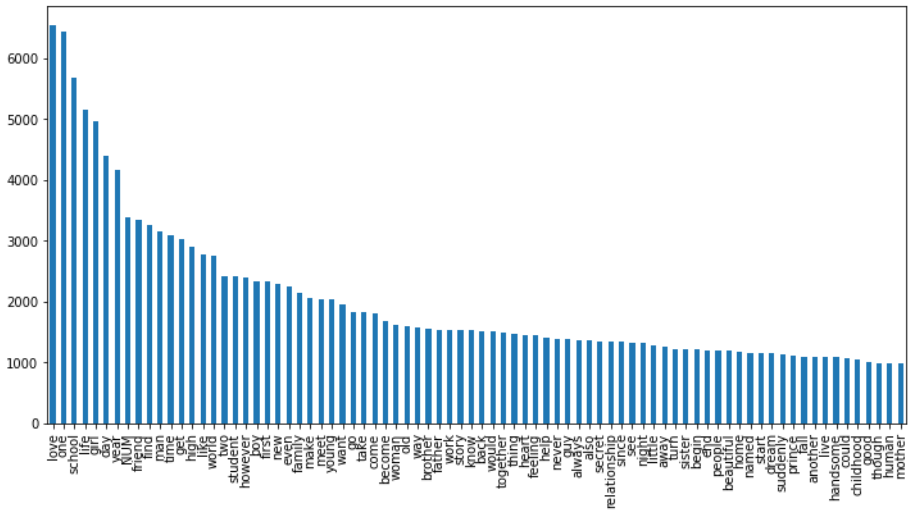
شکل ۸۴: هیستوگرام تعداد تکرار کلمات منحصر به فرد

Fantasy ۵.۹.۳



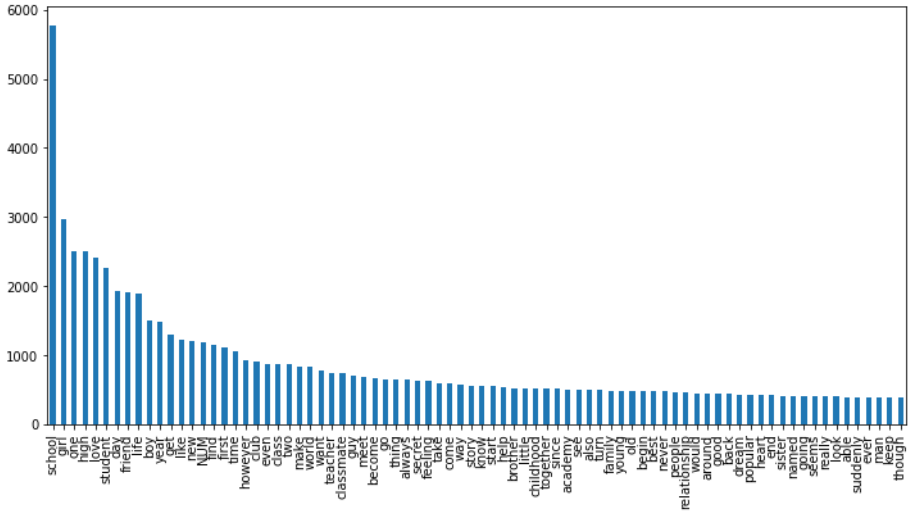
شکل ۸۵: هیستوگرام تعداد تکرار کلمات منحصر به فرد

Romance ۶.۹.۳



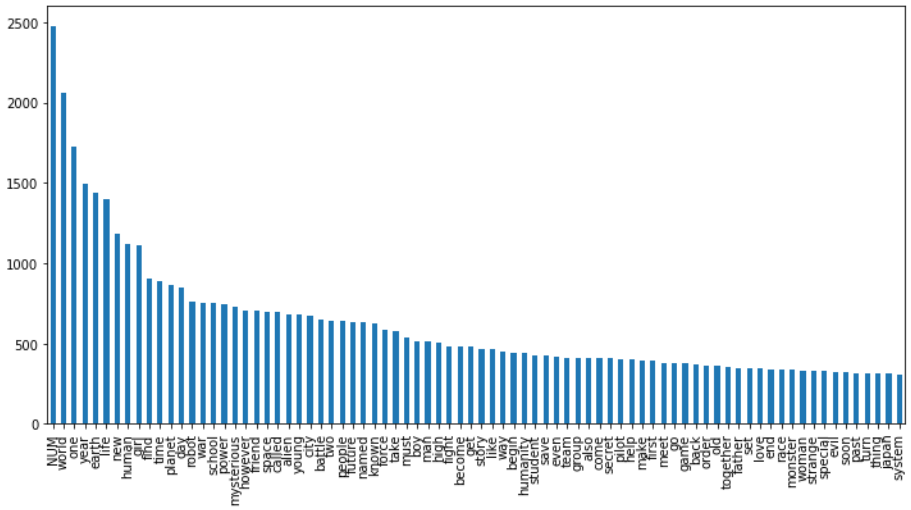
شکل ۸۶: هیستوگرام تعداد تکرار کلمات منحصر به فرد

School Life ۷.۹.۳



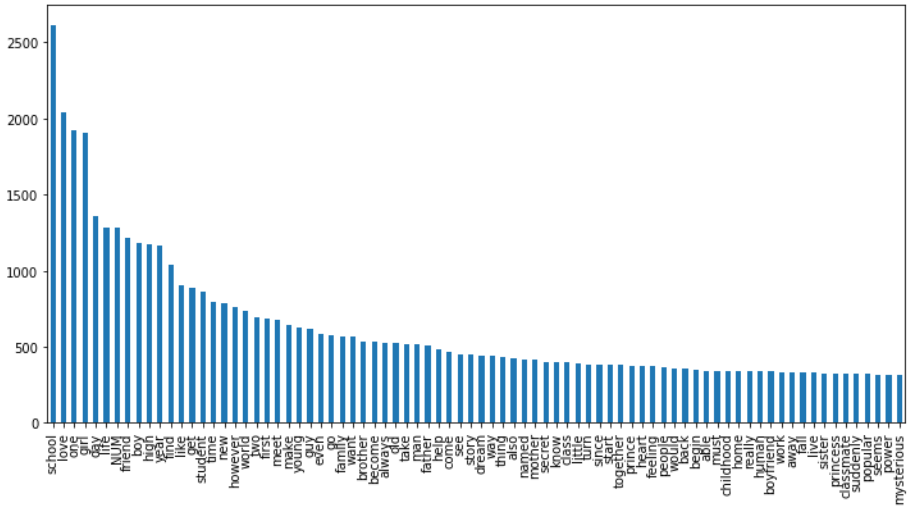
شکل ۸۷: هیستوگرام تعداد تکرار کلمات منحصر به فرد

Sci Fi ۸.۹.۳



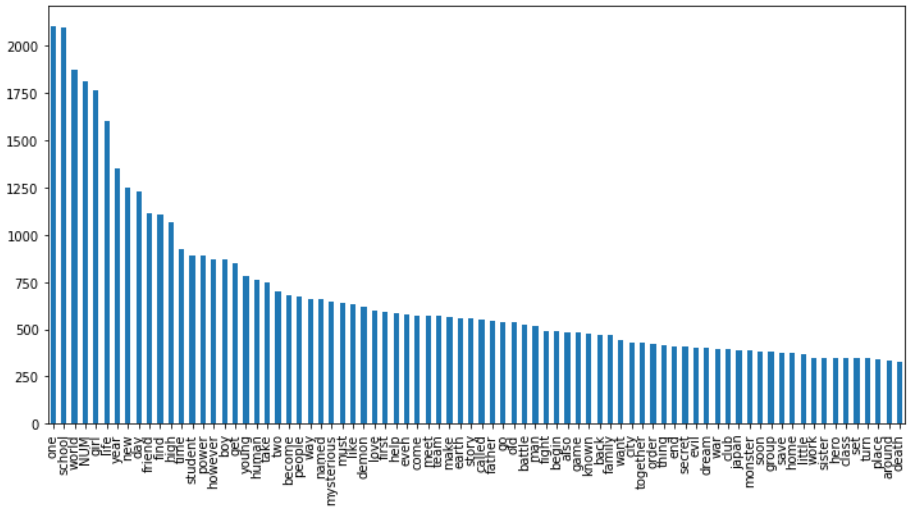
شکل ۸۸: هیستوگرام تعداد تکرار کلمات منحصر به فرد

Shoujo ۹.۹.۳



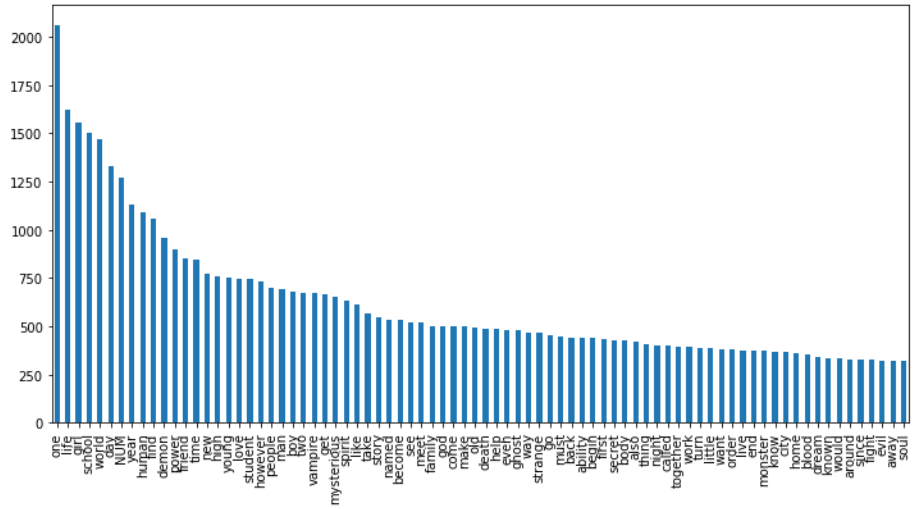
شکل ۸۹: هیستوگرام تعداد تکرار کلمات منحصر به فرد

Shounen ۱۰.۹.۳



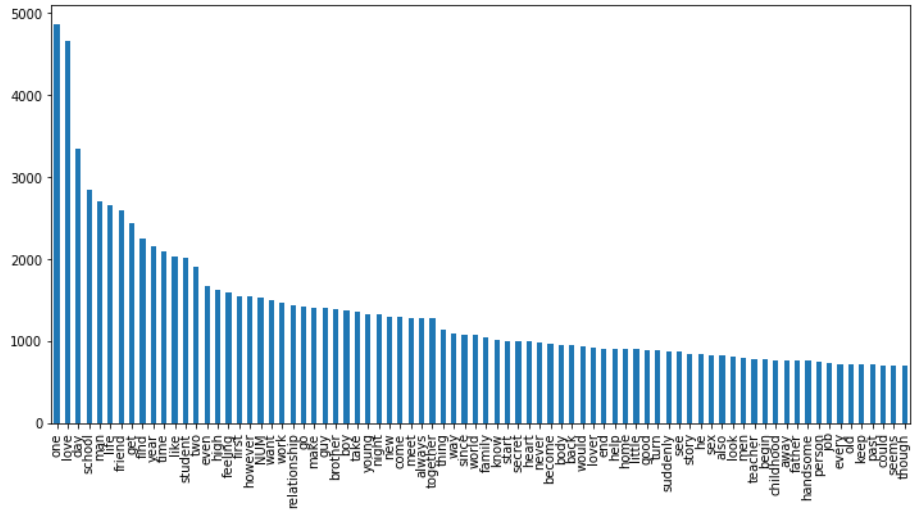
شکل ۹۰: هیستوگرام تعداد تکرار کلمات منحصر به فرد

۱۱.۹.۳ Supernatural



شکل ۹۱: هیستوگرام تعداد تکرار کلمات منحصر به فرد

۱۲.۹.۳ Yaoi



شکل ۹۲: هیستوگرام تعداد تکرار کلمات منحصر به فرد

۴ جمع بندی

در فرآیند جمع آوررری داده ، به منظور گردآوری داده بیشتر تلاش شد تا از وبسایت MyAnimeList نیز داده استخراج شود. اما API خود وبسایت اطلاعات خلاصه انیمه را استخراج نمی‌کرد. علاوه بر آن استخراج داده با کمک اسکریپت پایتون نیز به دلیل بالا بودن تعداد درخواست‌ها ، موجب مسدود شدن IP می‌شد. (وبسایت MyAnimeList هنگامی که تعداد request زیادی دریافت کند، IP مربوط را مسدود می‌کند.)

همچنین به این دلیل که کلمات غیرمشترک هر برچسب تعداد اندکی دارد و بیشتر کلمات برچسب‌ها مشترک هستند؛ ممکن است تشخیص ژانرها با مشکل مواجه شود. البته این نکته قابل ذکر است که پیش بینی ژانر بر اساس خلاصه در واقع multi-label classification است؛ یعنی یک انیمه یا مانگا فقط به یک ژانر خاص تعلق ندارد، بلکه می‌توان چندین ژانر داشته باشد و کم بودن کلمات غیرمشترک در این مسئله طبیعی است.

در ادامه پیاده‌سازی این پروژه، اگر داده جمع آوری شده کافی و مناسب نبود ، مجدد به گردآوری داده بیشتر پرداخته می‌شود.