دانشکده مهندسی کامپیوتر

مبانی پردازش زبان طبیعی

گزارش پروژه

فاز ۱ - جمع‌آوری داده

مهدی امیری

فهرست مطالب

۱ لینک به ریپوی پروژه در GitHub و HuggingFace

https://github.com/AmiriShavaki/IUST-NLP-Project

https://huggingface.co/datasets/Amiri/Google-Play-Reviews-for-Sentiment-Analysis/tree/main

۲ آماده‌سازی محیط مجازی برای اجرای پروژه

شما می‌توانید به کمک conda و از طریق فایل requirements.txt که در root پروژه در اختیار شما قرار داده شده، به راحتی و فقط با اجرای دستور زیر تمامی پکیج‌های موردنیاز برای اجرای پروژه را نصب کنید: conda create –name <envname> –file requirements.txt

۳ منبع دقیق داده

۵۵ اپلیکیشن منتخب از صفحه‌ی اول گوگل‌پلی (شامل بازی‌ها و اپ‌های کاربردی) در نظر گرفته شدند. از این بین ۴۵ اپلیکیشن حداقل ۱۰۰ نظر انگلیسی ثبت شده به ازای هر امتیاز داشتند که از این ۴۵ اپ تعداد ۴۵۰۰ نظر به ازای هر امتیاز (یعنی $4500 \times 5 = 22500$ نظر در مجموع) جمع‌آوری شده است.

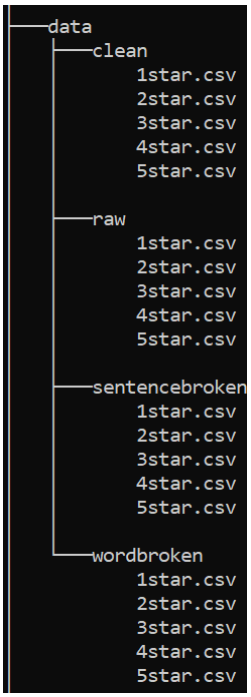
نظرات جمع‌آوری شده به ترتیب most-relevant انتخاب و جمع‌آوری شده‌اند. این نظرات معمولا طولانی‌تر و شامل موارد مربوط و قابل بررسی‌تر از نظرات بسیار کوتاه هستند.

۴ روش جمع‌آوری و ابزار مورداستفاده

برای جمع‌آوری داده‌ها از پکیج google-play-scraper استفاده شده است. به کمک این پکیج می‌توان از طریق اسم کامل یک اپلیکیشن به اطلاعات کامل آن شامل تمام یا بخشی از review های آن دست یافت. همچنین یکی از مزایای این پکیج امکان انتخاب زبان نظرات موردجمع‌آوری است که این امکان باعث می‌شود زحمت پیش‌پردازش فیلترکردن نظرات غیرانگلیسی را نداشته باشیم.

۵ فرمت داده‌ها

در کل، داده‌ها در چهار پوشه‌ی raw، clean، sentencebroken و wordbroken قرار گرفته‌اند. داده‌های هر برچسب در یک فایل csv به نام همان برچسب قرار دارد.



فایل‌های پوشه‌های raw و clean تک‌ستونی هستند و در هر ردیف متن مربوط به یک کامنت قرار دارد.

فایل‌های پوشه‌های sentencebroken و wordbroken به تعداد جملات/کلمات کامنت مربوطه ستون دارند. و در هر ردیف جملات/کلمات یک کامنت قرار گرفته است.

۶ پیش‌پردازش‌های انجام‌شده

۱.۶ روش/ابزار تفکیک جملات

برای تفکیک جملات از یکی از توکنایزرهای nltk استفاده کردم که با استفاده از یک مدل شبکه عصبی جملات متن ورودی را جدا می‌کند و در خروجی متن را با واحد جمله تحویل می‌دهد.

این توکنایزر در مسیر tokenizers/punkt/english.pickle از ریشه‌ی nltk قرار دارد.

۲.۶ روش/ابزار تفکیک توکن‌ها/کلمات

برای جدا کردن کلمات از توکنایزر مربوط به پکیج wordsegment استفاده کردم که از مدل‌های عصبی استفاده می‌کند. این پکیج همه‌ی کلمات را به حالت حروف کوچک (lowercase) تبدیل می‌کند و علائم نگارشی را حذف می‌کند. برای تحلیل‌های بعدی و ورودی دادن به مدل اصلی از آن‌جایی بزرگی و کوچکی حروف در لحن متن تاثیر گذار است از خروجی این مرحله استفاده نخواهیم کرد و صرفا برای تحلیل‌های آماری در فاز ۱ از این خروجی استفاده شده است.

۳.۶ روش/معیارهای تمیزکردن داده

۱.۳.۶ جاگذاری ایموجی‌ها

اگر بدون هیچ پیش‌پردازشی روی ایموجی‌ها، آن‌ها را به مدل‌هایی مانند Bert بدهیم، یک توکن Unknown بجای همه‌ی ایموجی‌ها قرار خواهد گرفت، از طرفی حذف‌کردن ایموجی‌ها در مرحله‌ی تمیزکاری داده‌ها منطقی به نظر نمی‌رسد چون با توجه به ماهیت تسک Sentiment analysis وجود ایموجی و نوع ایموجی به کار رفته می‌تواند سرنخ بزرگی برای تشخیص لحن نویسنده‌ی کامنت باشد. در اینجا روش به‌کار گرفته شده جاگذاری ایموجی با متن کوتاه جایگزین ایموجی است، متنی که در حد چند کلمه توصیف معنایی آن ایموجی است که به جای خود ایموجی می‌تواند به مدل، سرنخ برای تشخیص لحن نویسنده بدهد.

برای این‌کار از پکیج آماده emoji پایتون استفاده شده است که به ازای هر ایموجی یک متن کوتاه چندکلمه‌ای آماده در اختیار می‌گذارد.

۲.۳.۶ حذف کلمات پالایشی (stopwords)

منظور از کلمات پالایشی کلماتی مثل the ، and ، is ، in ، it و... است. این کلمات دربردارنده بار معنایی ارزشمندی برای تشخیص لحن نیستند و به همین دلیل با حذف آن‌ها از متن ورودی می‌توانیم حجم نویز ورودی را کم کنیم و به مدل اجازه دهیم بتواند روی کلمات کلیدی‌تر متن ورودی تمرکز کند.

برای حذف این کلمات از پکیج nltk استفاده شده است.

۳.۳.۶ بسط دادن فرم‌های کوتاه‌شده (**contractions**)

در زبان انگلیسی به کلمات و عبارت‌های کوتاه‌شده‌ای مثل can't ، don't ، won't و it's فرم contraction می‌گویند. با تبدیل این عبارت‌ها به فرم بسط یافته می‌توانیم فرآیند tokenization را بهبود بدهیم و consistency ایجاد کنیم چون فرم بسط یافته و فرم contraction به لحاظ معنایی کاملا یکسانند بهتر است در مدل با آن‌ها یک جور مواجه شویم.

به این منظور از پکیج آماده‌ی contractions استفاده شده است که مورد گفته شده را بطور کامل انجام می‌دهد.

۴.۳.۶ حذف ارقام

به نظر نمی‌رسد ارقام متن ورودی در تشخیص لحن آن بتوانند به مدل کمک کنند. پس تمام ارقام موجود در متن ورودی را حذف می‌کنیم.

۵.۳.۶ حذف کاراکترهای نامربوط (کاراکترهای غیر از ارقام، حروف، فاصله و نقطه)

برای حذف کاراکترهای غیر از ارقام، حروف، فاصله و نقطه از عبارت منظم (regex) استفاده شده است.

۶.۳.۶ حذف فاصله‌های اضافی

در نهایت بعد از اعمال همه‌ی پیش‌پردازش‌های بالا به ترتیبی که گفته شد، همه‌ی فاصله‌های اضافی به کار رفته حذف شدند تا داده‌ها تمیز و آماده مرحله tokenization باشند.

۴.۶ اندازه داده قبل/بعد تمیزکردن داده

قبل از تمیزکردن داده:

	۱ ستاره	۲ ستاره	۳ ستاره	۴ ستاره	۵ ستاره	همه برچسب‌ها
تعداد کلمات	۳۱۰۶۰۷	۳۰۰۳۴۰	۳۰۸۳۰۴	۲۹۵۹۰۷	۲۸۵۶۴۷	۱۵۰۰۸۰۵
حجم فایل	1.61MB	1.55MB	1.58MB	1.52MB	1.48MB	7.74MB

بعد از تمیزکردن داده:

	۱ ستاره	۲ ستاره	۳ ستاره	۴ ستاره	۵ ستاره	همه برچسب‌ها
تعداد کلمات	۱۸۵۴۰۹	۱۷۷۹۵۲	۱۸۱۶۳۳	۱۷۴۶۱۳	۱۷۲۳۷۵	۸۹۱۹۸۲
حجم فایل	1.11MB	1.06MB	1.08MB	1.04MB	1.03MB	5.33MB

*در اینجا معیار جداکردن کلمات، کاراکتر فاصله ($space$) بوده است.

۷ واحد و روش برچسب‌گذاری

واحد داده، یک کامنت از صفحه اپ‌ها است و برچسب‌گذاری به ازای هر کامنت انجام شده است.

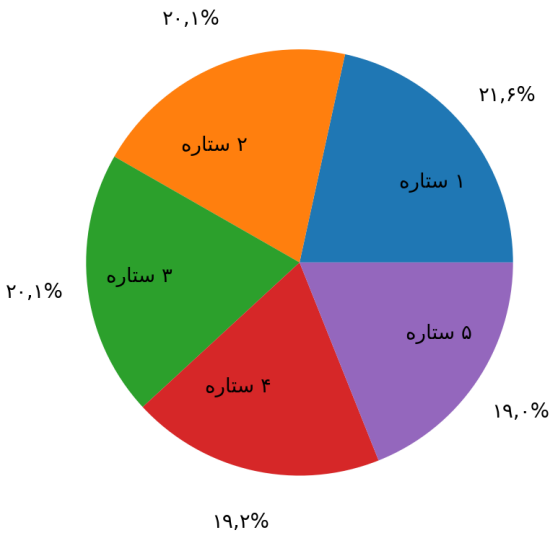
هر کامنتی که در Google Play نوشته می‌شود کاربر برحسب میزان رضایت خود از اپلیکیشن، امتیاز بین ۱ تا ۵ به اپلیکیشن می‌دهد. مبنای برچسب‌دهی ما به کامنت‌ها هم بر اساس همین امتیاز ثبت‌شده توسط خود کاربر است. برای جمع‌آوری این برچسب‌ها از خروجی پکیج google play scraper استفاده شده است. خروجی که این پکیج به ازای هر کامنت در اختیار می‌گذارد شامل متن کامنت، امتیاز(تعداد ستاره) و موارد دیگری همچون زبان کامنت و تاریخ درج آن است.

۸ آمار داده به تفکیک برچسب

۱.۸ تعداد کامنت‌ها

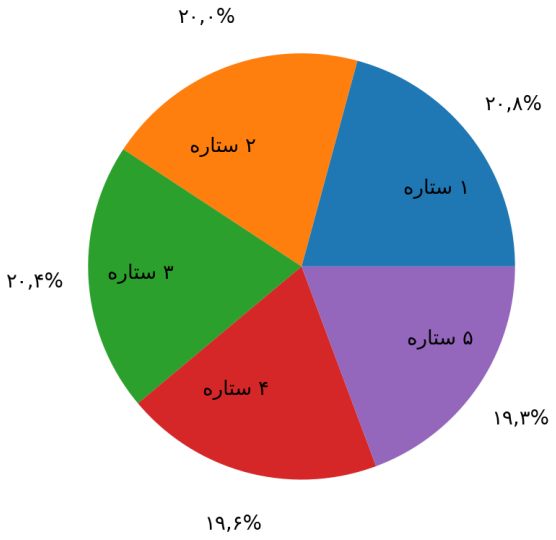
	۱ ستاره	۲ ستاره	۳ ستاره	۴ ستاره	۵ ستاره	همه برچسب‌ها
تعداد کامنت جمع‌آوری‌شده	۴۵۰۰	۴۵۰۰	۴۵۰۰	۴۵۰۰	۴۵۰۰	۲۲۵۰۰

۲.۸ تعداد جملات



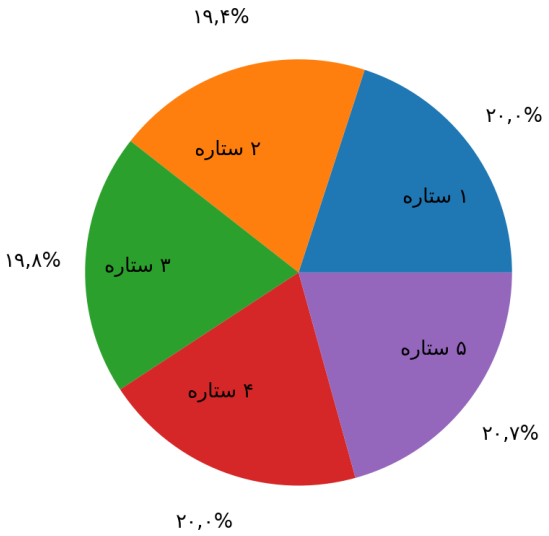
	۱ ستاره	۲ ستاره	۳ ستاره	۴ ستاره	۵ ستاره	همه برچسب‌ها
تعداد جملات	۲۰۷۹۱	۱۹۴۱۰	۱۹۳۹۹	۱۸۴۷۷	۱۸۲۶۸	۹۶۳۴۵

۳.۸ تعداد کلمات



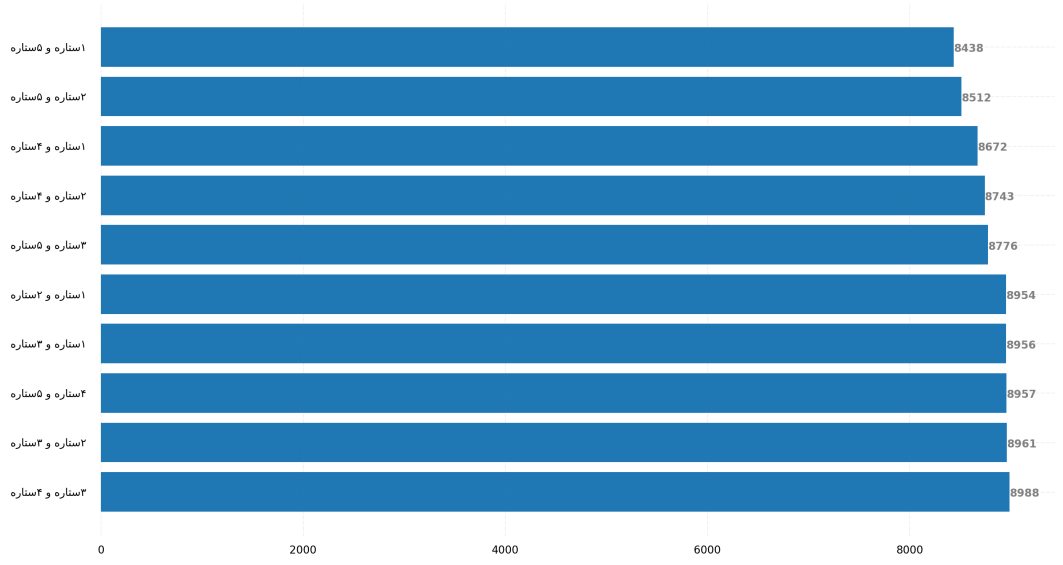
	۱ ستاره	۲ ستاره	۳ ستاره	۴ ستاره	۵ ستاره	همه برچسب‌ها
تعداد کلمات	۱۸۵۴۰۹	۱۷۷۹۵۲	۱۸۱۶۳۳	۱۷۴۶۱۳	۱۷۲۳۷۵	۸۹۱۹۸۲

۴.۸ تعداد کلمات منحصربفرد



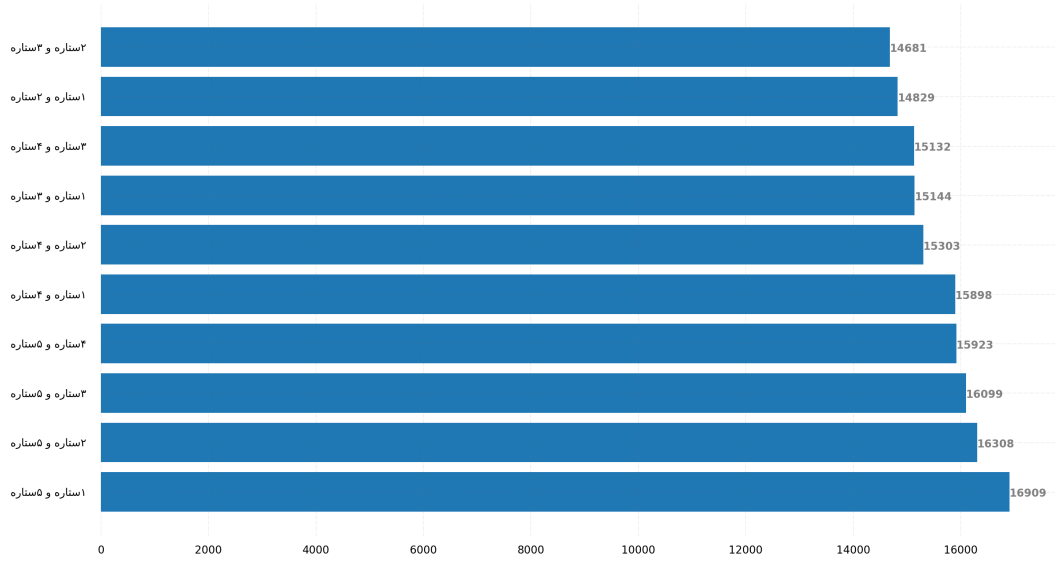
	۱ ستاره	۲ ستاره	۳ ستاره	۴ ستاره	۵ ستاره
تعداد کلمات منحصر به فرد	۱۶۵۹۵	۱۶۱۴۲	۱۶۴۶۱	۱۶۶۴۷	۱۷۱۹۰

۵.۸ تعداد کلمات منحصربفرد مشترک بین جفت برچسب‌ها



	۱ستاره و ۲ستاره	۱ستاره و ۳ستاره	۱ستاره و ۴ستاره	۱ستاره و ۵ستاره	۲ستاره و ۳ستاره	۲ستاره و ۴ستاره	۲ستاره و ۵ستاره	۳ستاره و ۴ستاره	۳ستاره و ۵ستاره	۴ستاره و ۵ستاره
کلمات منحصربفرد مشترک	۸۹۵۴	۸۹۵۶	۸۶۷۲	۸۴۳۸	۸۹۶۱	۸۷۴۳	۸۵۱۲	۸۹۸۸	۸۷۷۶	۸۹۵۷

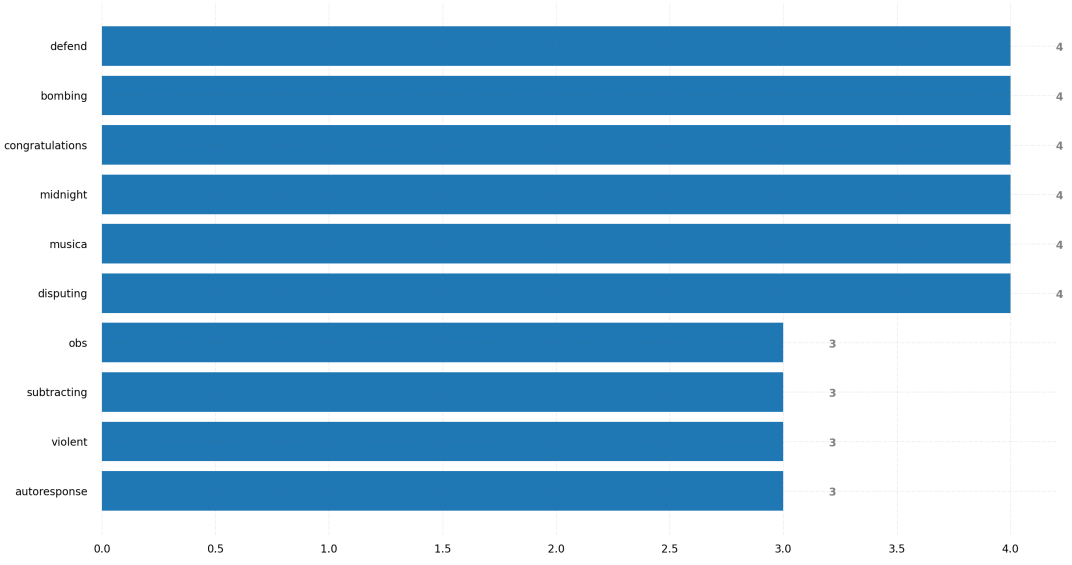
۶.۸ تعداد کلمات منحصربفرد غیرمشترک بین جفت برچسب‌ها



	۱ستاره و ۲ستاره	۱ستاره و ۳ستاره	۱ستاره و ۴ستاره	۱ستاره و ۵ستاره	۲ستاره و ۳ستاره	۲ستاره و ۴ستاره	۲ستاره و ۵ستاره	۳ستاره و ۴ستاره	۳ستاره و ۵ستاره	۴ستاره و ۵ستاره
کلمات منحصربفرد غیرمشترک	۱۴۸۲۹	۱۵۱۴۴	۱۵۸۹۸	۱۶۹۰۹	۱۴۶۸۱	۱۵۳۰۳	۱۶۳۰۸	۱۵۱۳۲	۱۶۰۹۹	۱۵۹۲۳

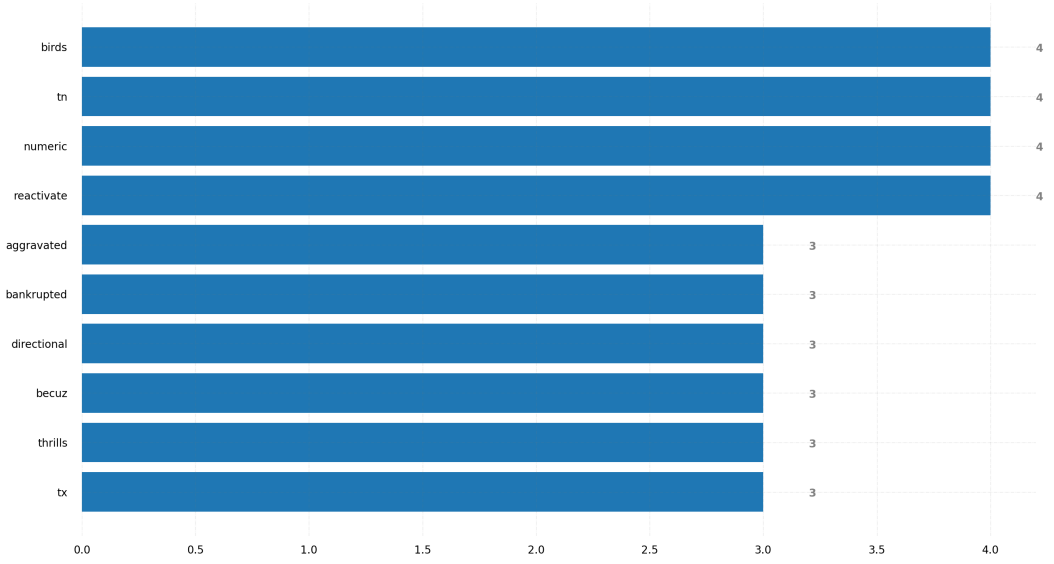
۷.۸ ۱۰ کلمه پرتکرار غیرمشترک هر برچسب

۱.۷.۸ ۱ ستاره



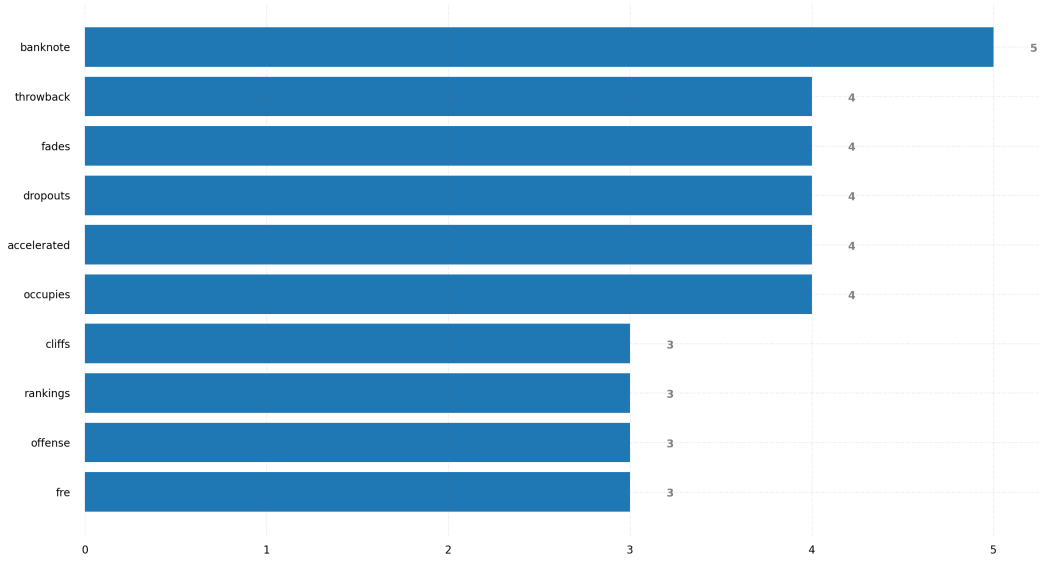
defend	bombing	congratulations	midnight	musica	disputing	obs	subtracting	violent	autoresponse
4	4	4	4	4	4	3	3	3	3

۲.۷.۸ ۲ ستاره



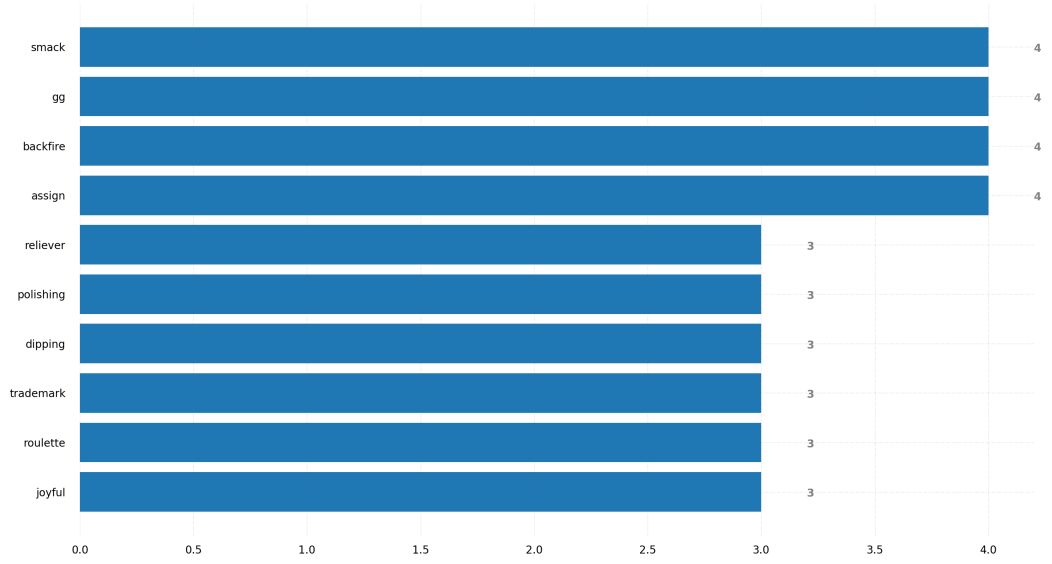
birds	tn	numeric	reactivate	aggravated	bankrupted	directional	becuz	thrills	tx
4	4	4	4	3	3	3	3	3	3

۳.۷.۸ ۳ ستاره



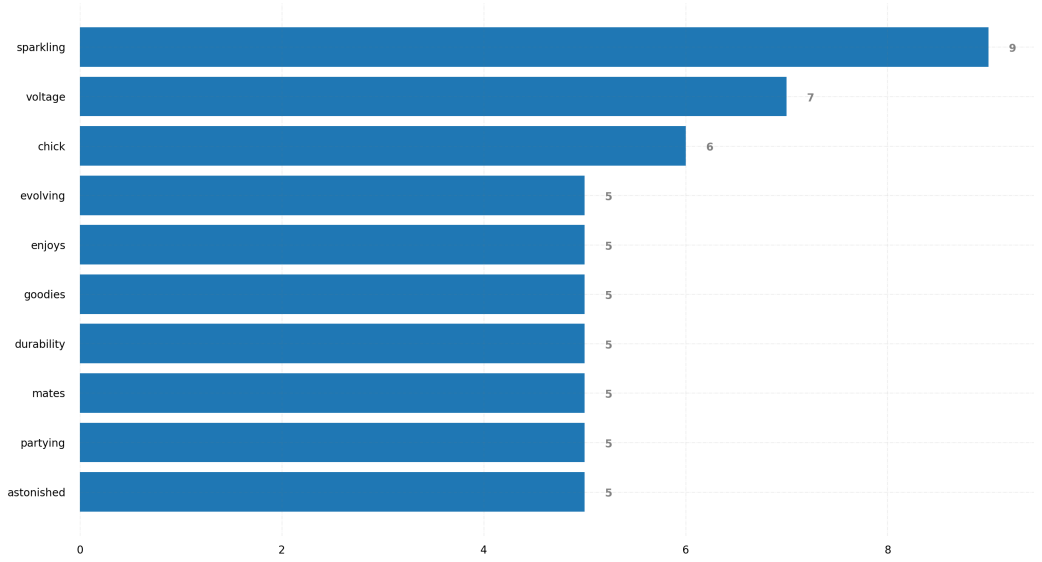
banknote	throwback	fades	dropouts	accelerated	occupies	cliffs	rankings	offense	fre
5	4	4	4	4	4	3	3	3	3

۴.۷.۸ ۴ ستاره



smack	gg	backfire	assign	reliever	polishing	dipping	trademark	roulette	joyful
4	4	4	4	3	3	3	3	3	3

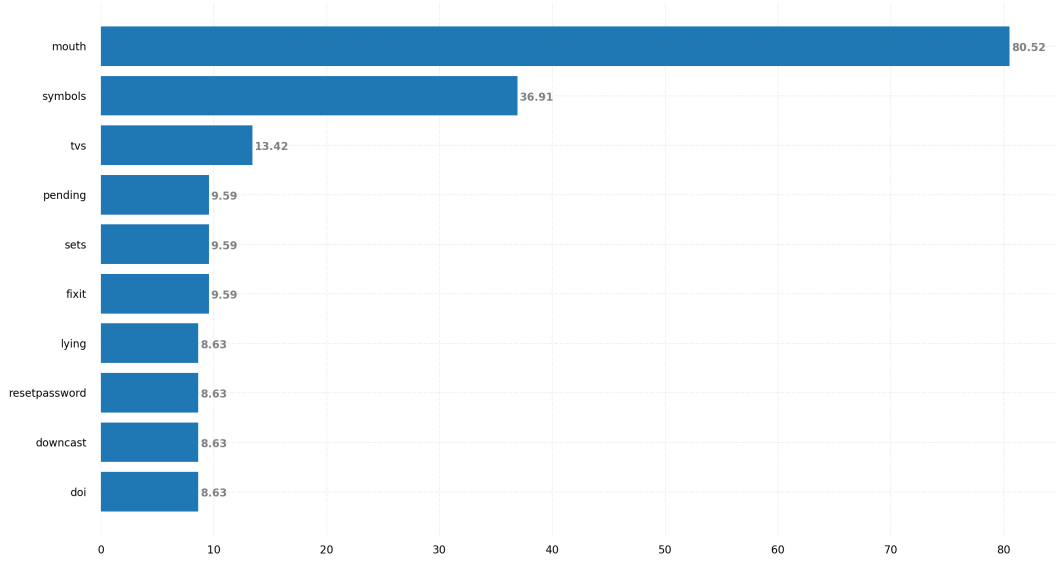
۵.۷.۸ ۵ ستاره



sparkling	voltage	chick	evolving	enjoys	goodies	durability	mates	partying	astonished
9	7	6	5	5	5	5	5	5	5

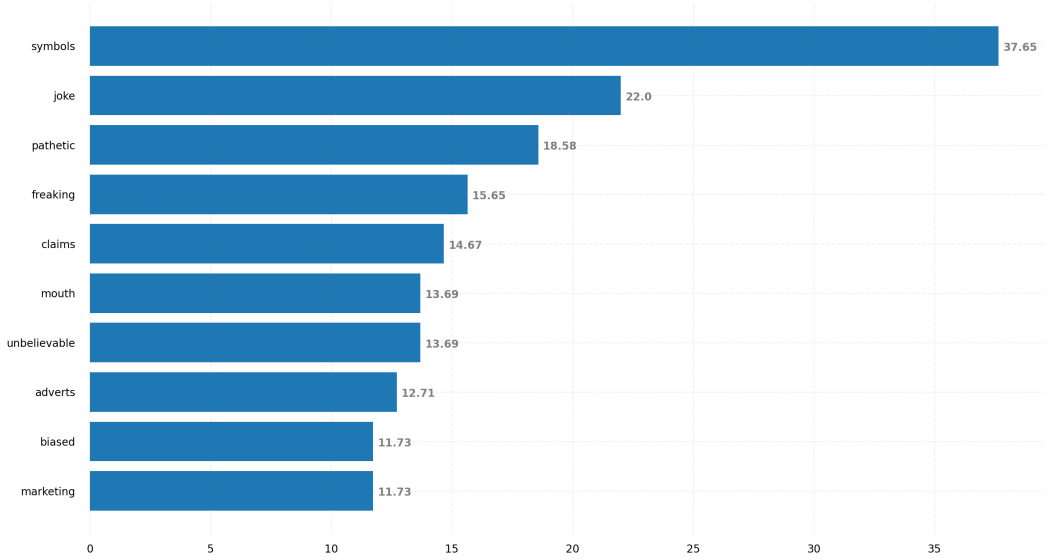
۸.۸ ۱۰ کلمه مشترک برتر هر برچسب نسبت به برچسب‌های دیگر بر اساس معیار **Relative Normalized Frequency**

۱.۸.۸ ۱ ستاره به ۲ ستاره



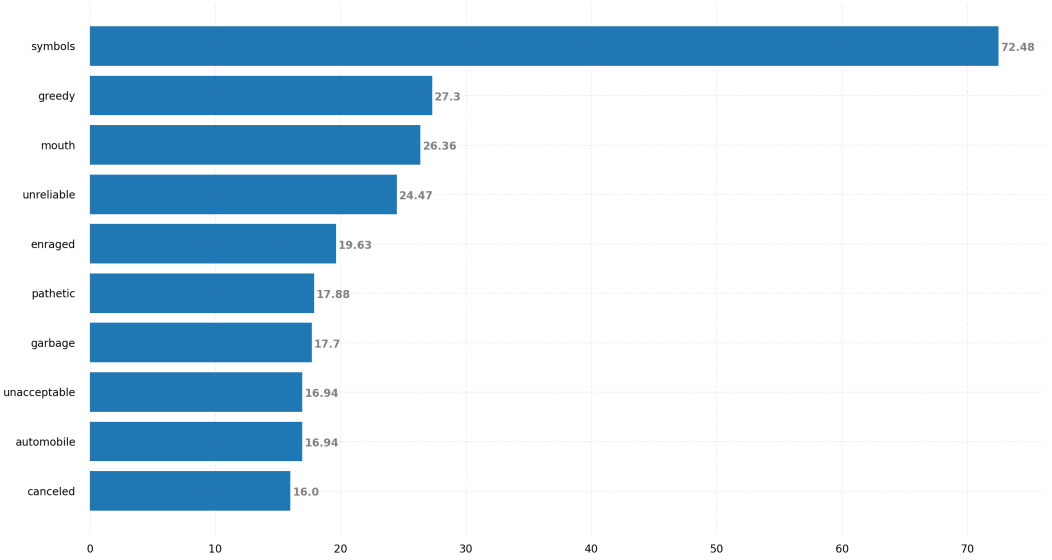
mouth	symbols	tvs	pending	sets	fixit	lying	resetpassword	downcast	doi
80.5	36.9	13.4	9.6	9.6	9.6	8.6	8.6	8.6	8.6

۲.۸.۸ ۱ ستاره به ۳ ستاره



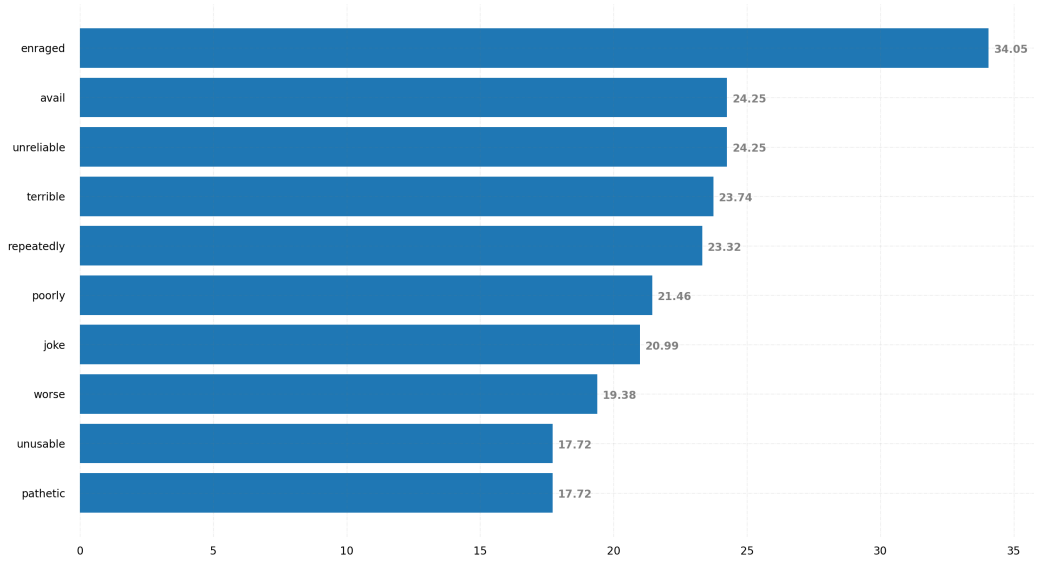
symbols	joke	pathetic	freaking	claims	mouth	unbelievable	adverts	biased	marketing
37.6	22.0	18.6	15.6	14.7	13.7	13.7	12.7	11.7	11.7

۳.۸.۸ ۱ ستاره به ۴ ستاره



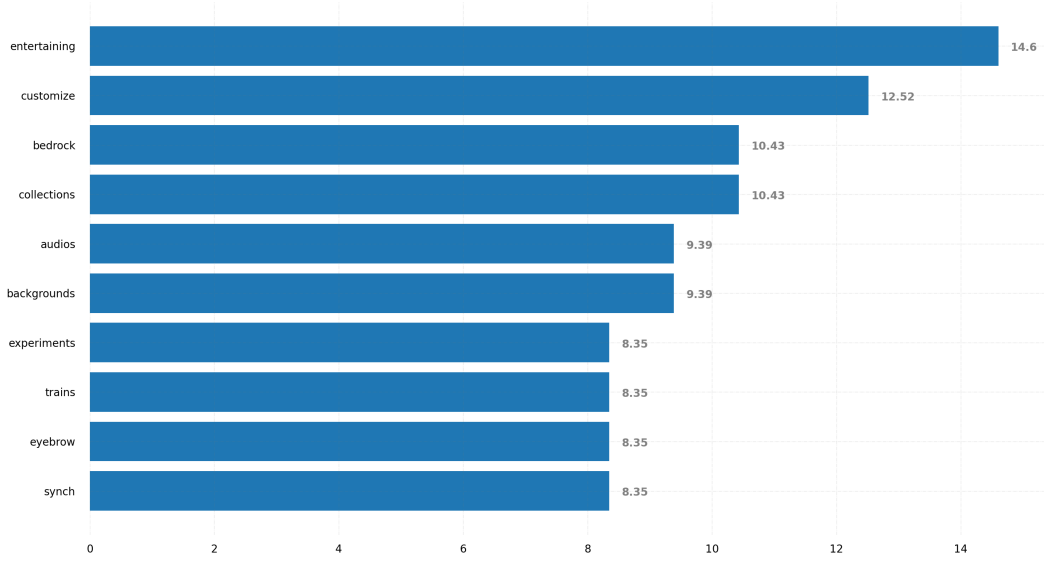
symbols	greedy	mouth	unreliable	enraged	pathetic	garbage	unacceptable	automobile	canceled
72.5	27.3	26.4	24.5	19.6	17.9	17.7	16.9	16.9	16.0

۴.۸.۸ ۱ ستاره به ۵ ستاره



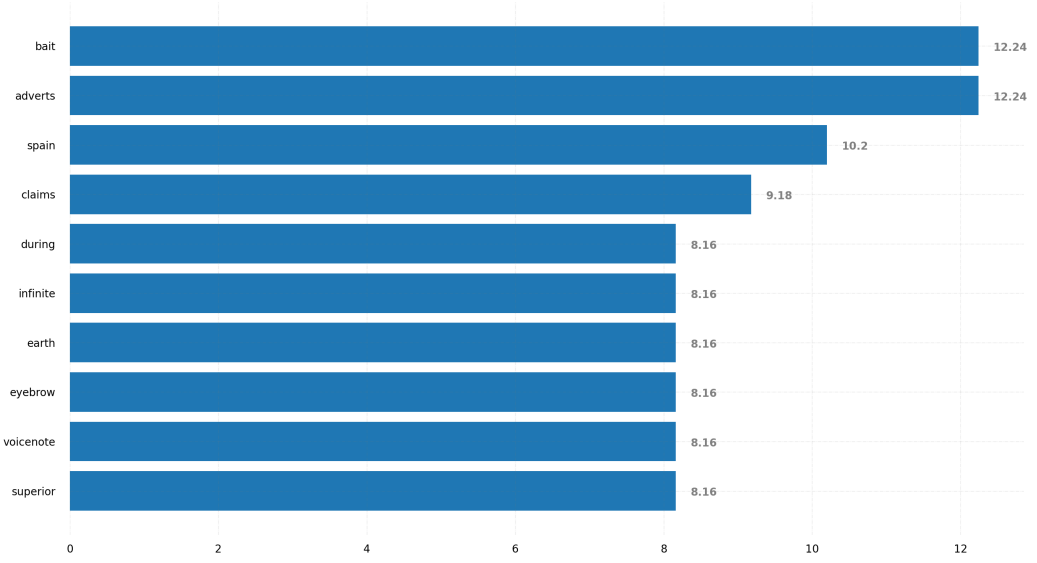
enraged	avail	unreliable	terrible	repeatedly	poorly	joke	worse	unusable	pathetic
34.1	24.3	24.3	23.7	23.3	21.5	21.0	19.4	17.7	17.7

۵.۸.۸ ۲ ستاره به ۱ ستاره



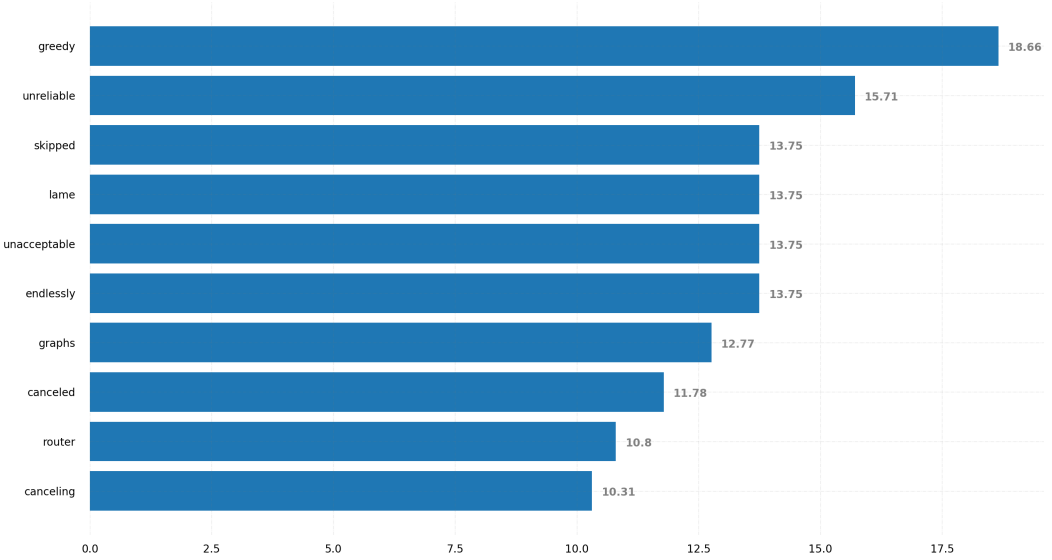
entertaining	customize	bedrock	collections	audios	backgrounds	experiments	trains	eyebrow	synch
14.6	12.5	10.4	10.4	9.4	9.4	8.3	8.3	8.3	8.3

۶.۸.۸ ۲ ستاره به ۳ ستاره



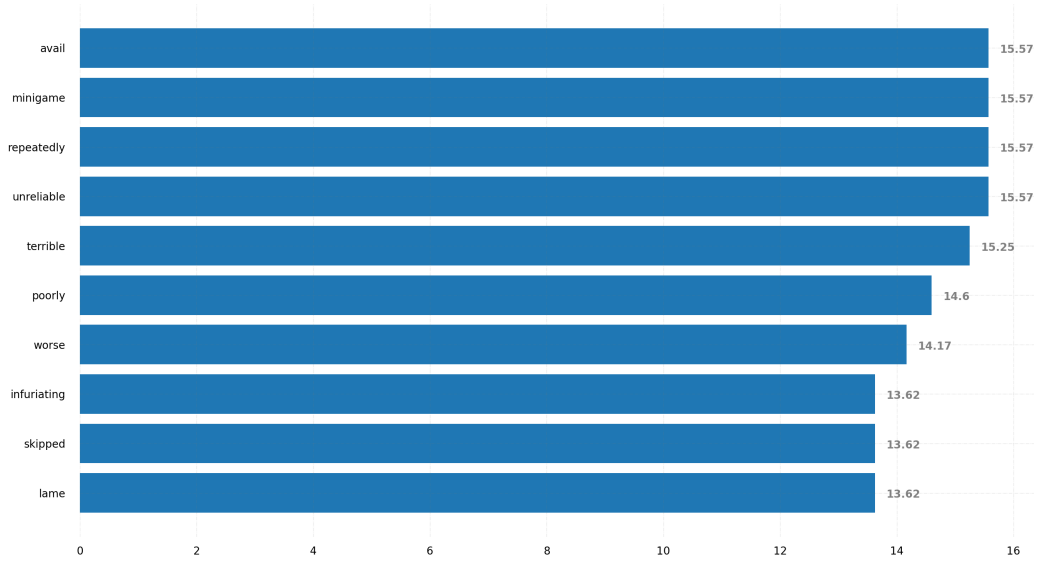
bait	adverts	spain	claims	during	infinite	earth	eyebrow	voicenote	superior
12.2	12.2	10.2	9.2	8.2	8.2	8.2	8.2	8.2	8.2

۷.۸.۸ ۲ ستاره به ۴ ستاره



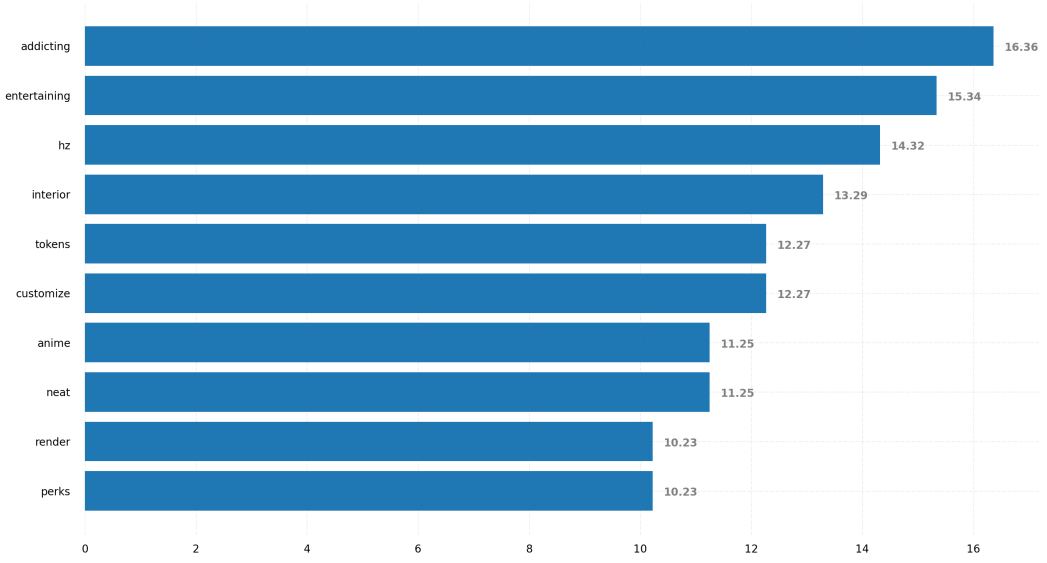
greedy	unreliable	skipped	lame	unacceptable	endlessly	graphs	canceled	router	canceling
18.7	15.7	13.7	13.7	13.7	13.7	12.8	11.8	10.8	10.3

۸.۸.۸ ۲ ستاره به ۵ ستاره



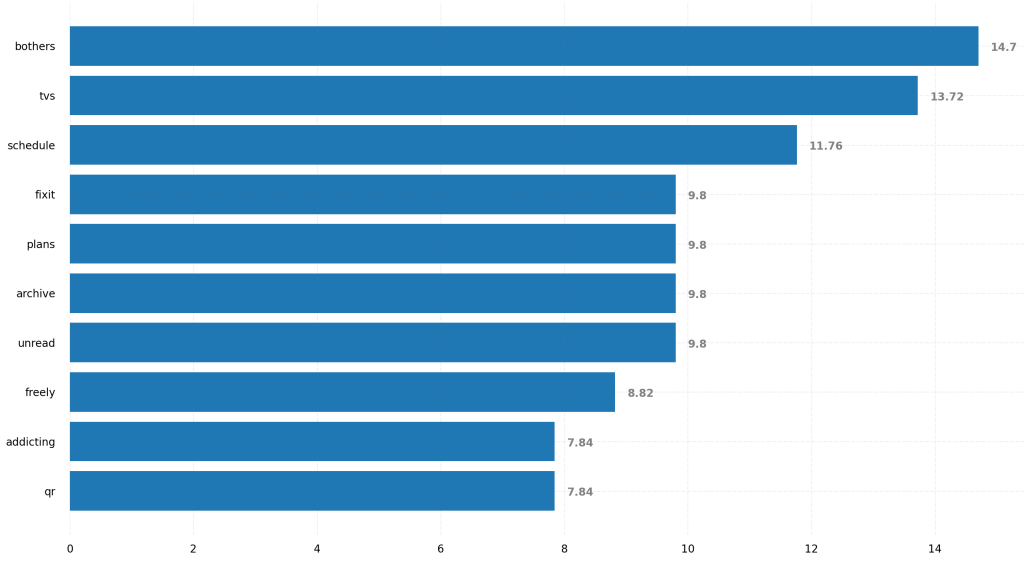
avail	minigame	repeatedly	unreliable	terrible	poorly	worse	infuriating	skipped	lame
15.6	15.6	15.6	15.6	15.2	14.6	14.2	13.6	13.6	13.6

۹.۸.۸ ۳ ستاره به ۱ ستاره



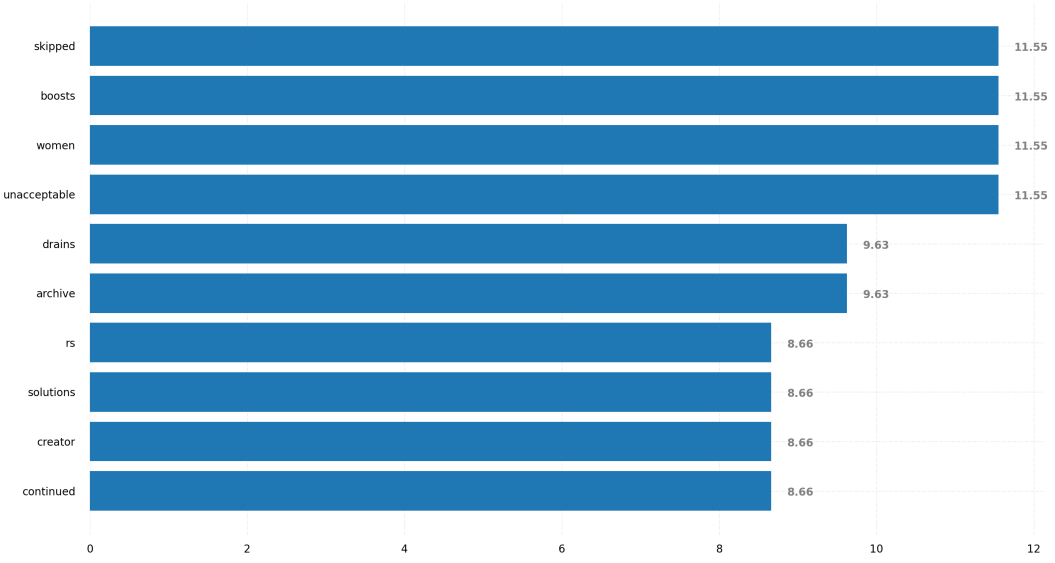
addicting	entertaining	hz	interior	tokens	customize	anime	neat	render	perks
16.4	15.3	14.3	13.3	12.3	12.3	11.2	11.2	10.2	10.2

۱۰.۸.۸ ۳ ستاره به ۲ ستاره



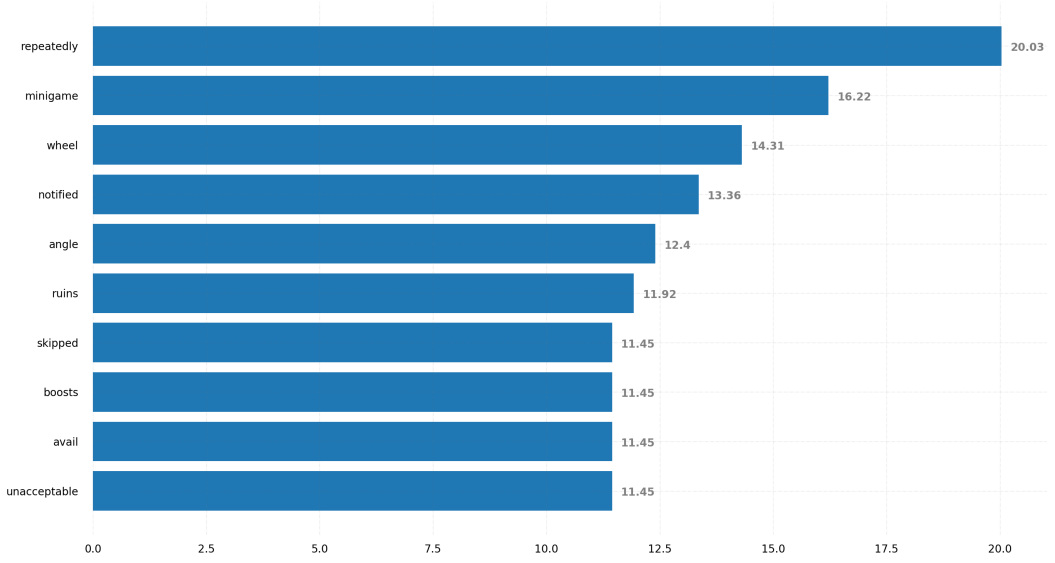
bothers	tvs	schedule	fixit	plans	archive	unread	freely	addicting	qr
14.7	13.7	11.8	9.8	9.8	9.8	9.8	8.8	7.8	7.8

۱۱.۸.۸ ۳ ستاره به ۴ ستاره



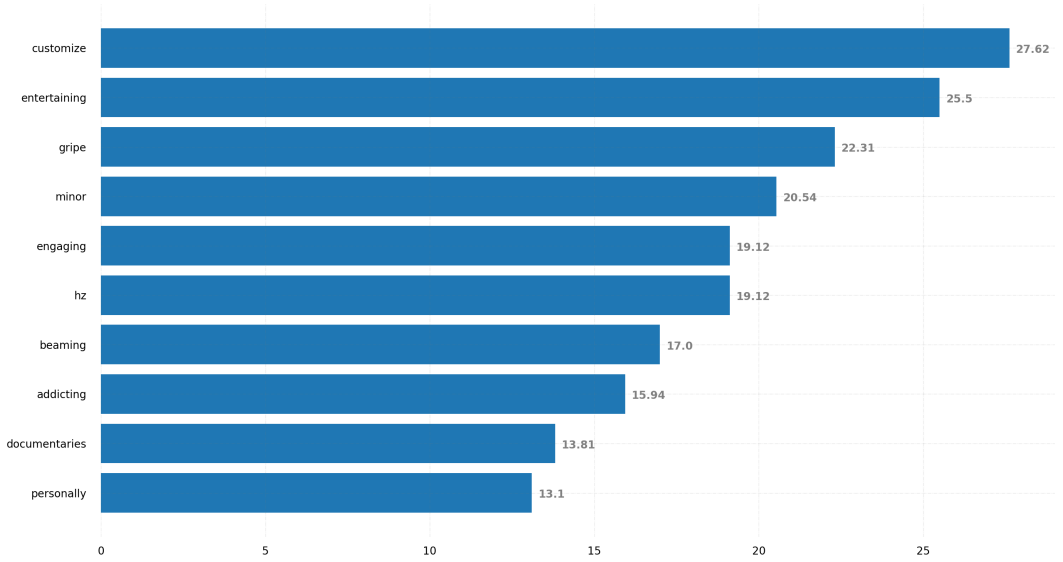
skipped	boosts	women	unacceptable	drains	archive	rs	solutions	creator	continued
11.6	11.6	11.6	11.6	9.6	9.6	8.7	8.7	8.7	8.7

۱۲.۸.۸ ۳ ستاره به ۵ ستاره



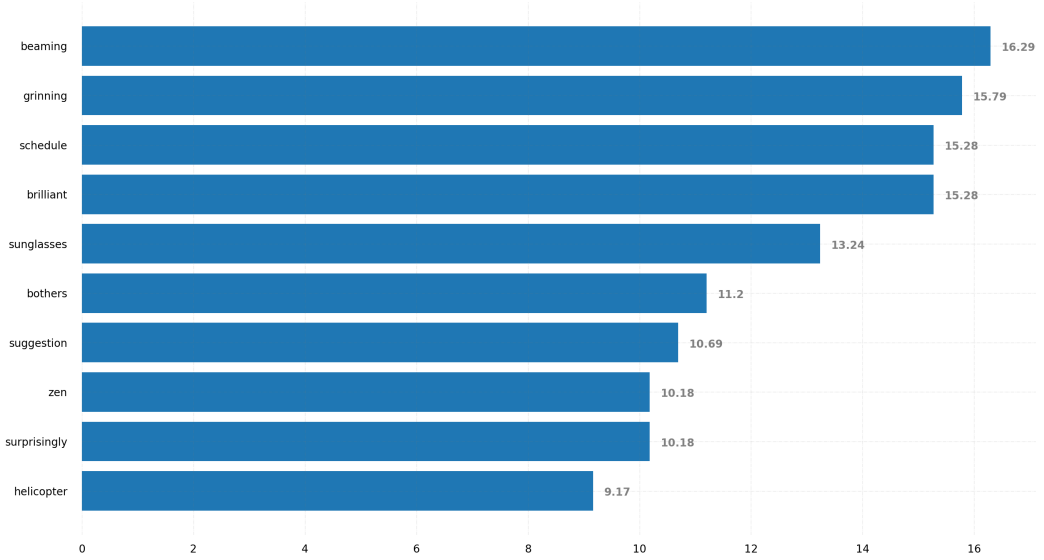
repeatedly	minigame	wheel	notified	angle	ruins	skipped	boosts	avail	unacceptable
20.0	16.2	14.3	13.4	12.4	11.9	11.4	11.4	11.4	11.4

۱۳.۸.۸ ۴ ستاره به ۱ ستاره



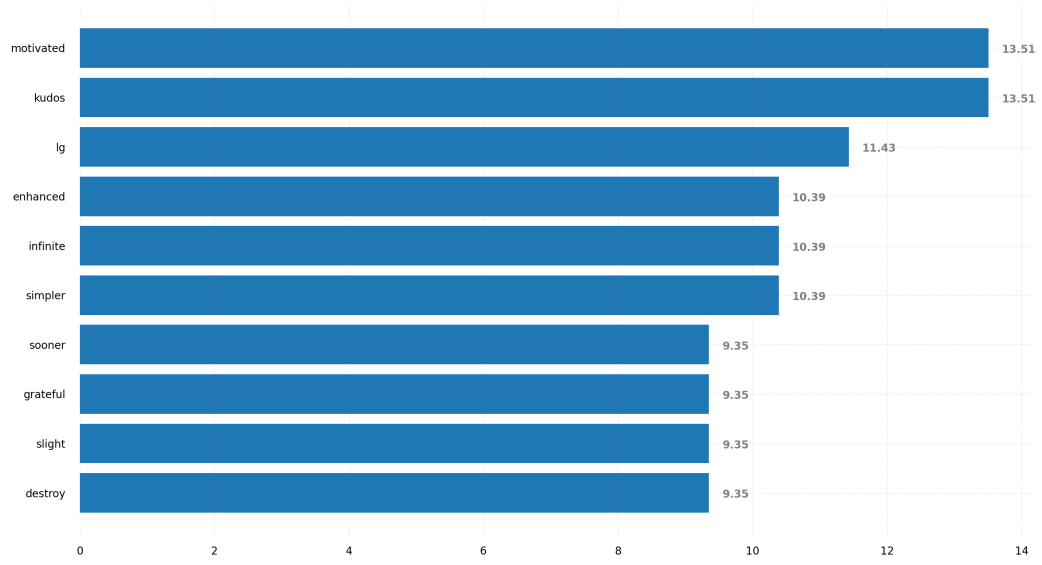
customize	entertaining	gripe	minor	engaging	hz	beaming	addicting	documentaries	personally
27.6	25.5	22.3	20.5	19.1	19.1	17.0	15.9	13.8	13.1

۱۴.۸.۸ ۴ ستاره به ۲ ستاره



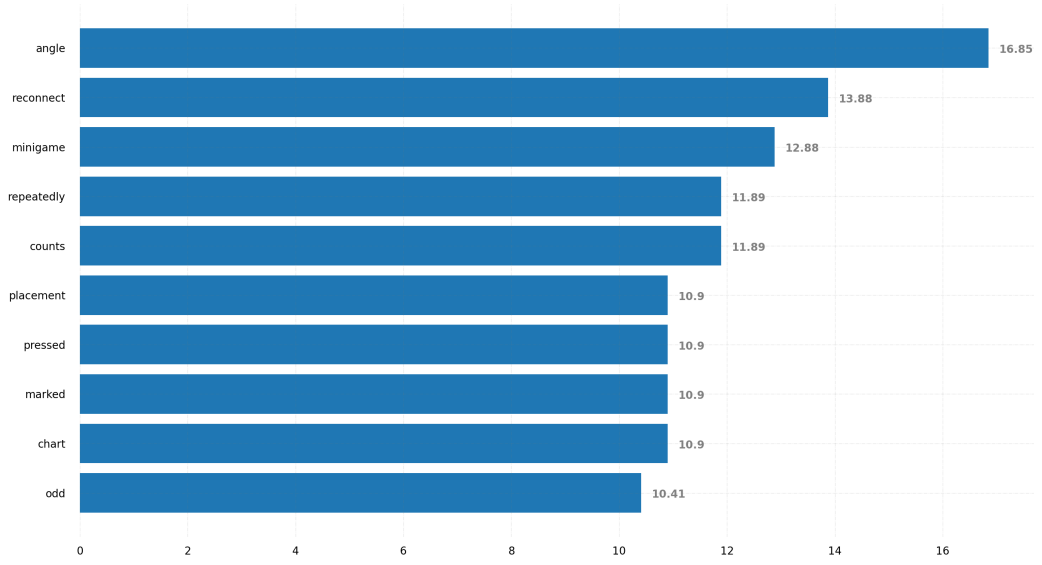
beaming	grinning	schedule	brilliant	sunglasses	bothers	suggestion	zen	surprisingly	helicopter
16.3	15.8	15.3	15.3	13.2	11.2	10.7	10.2	10.2	9.2

۱۵.۸.۸ ۴ ستاره به ۳ ستاره



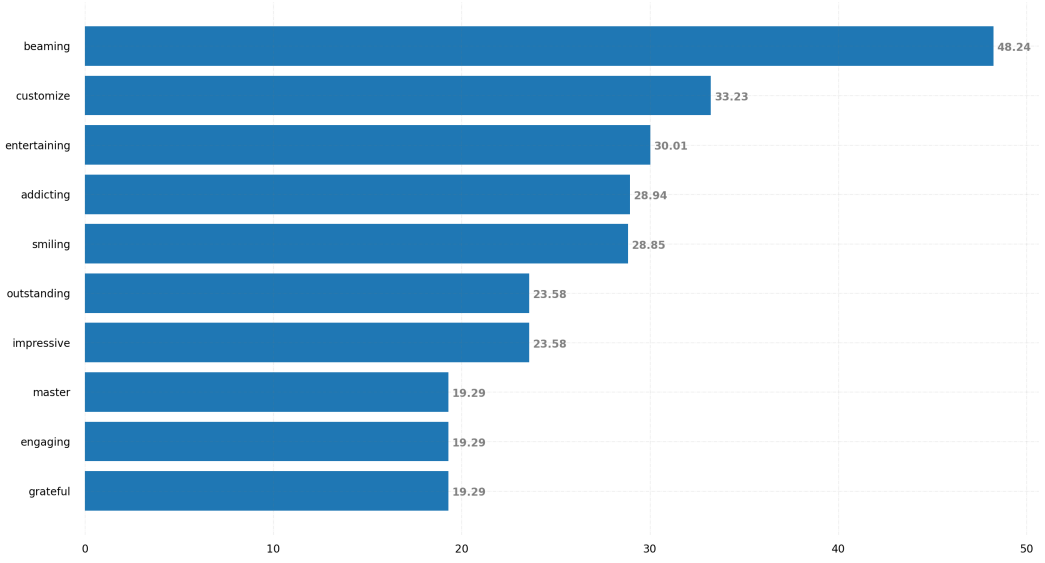
motivated	kudos	lg	enhanced	infinite	simpler	sooner	grateful	slight	destroy
13.5	13.5	11.4	10.4	10.4	10.4	9.4	9.4	9.4	9.4

۱۶.۸.۸ ۴ ستاره به ۵ ستاره



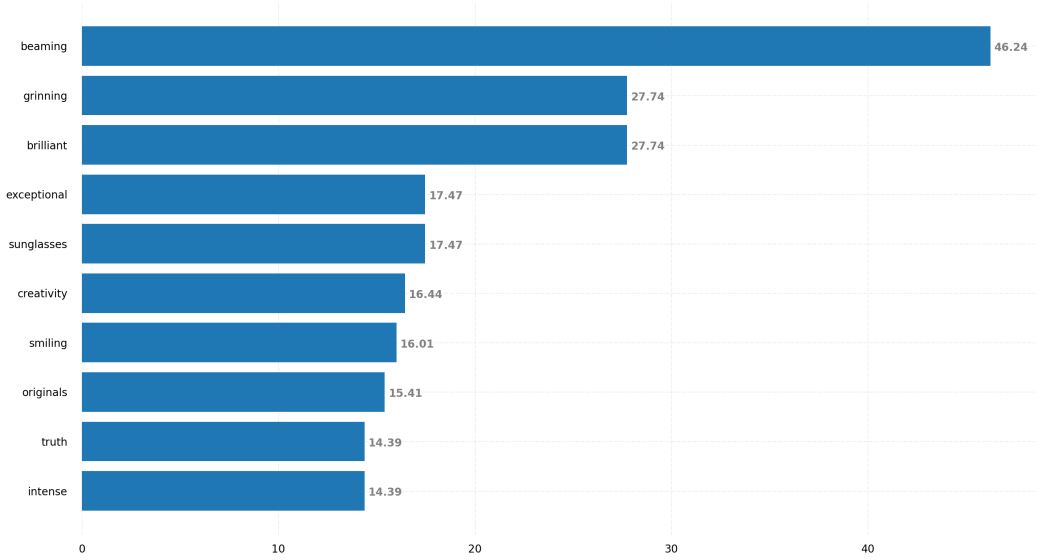
angle	reconnect	minigame	repeatedly	counts	placement	pressed	marked	chart	odd
16.8	13.9	12.9	11.9	11.9	10.9	10.9	10.9	10.9	10.4

۱۷.۸.۸ ۵ ستاره به ۱ ستاره



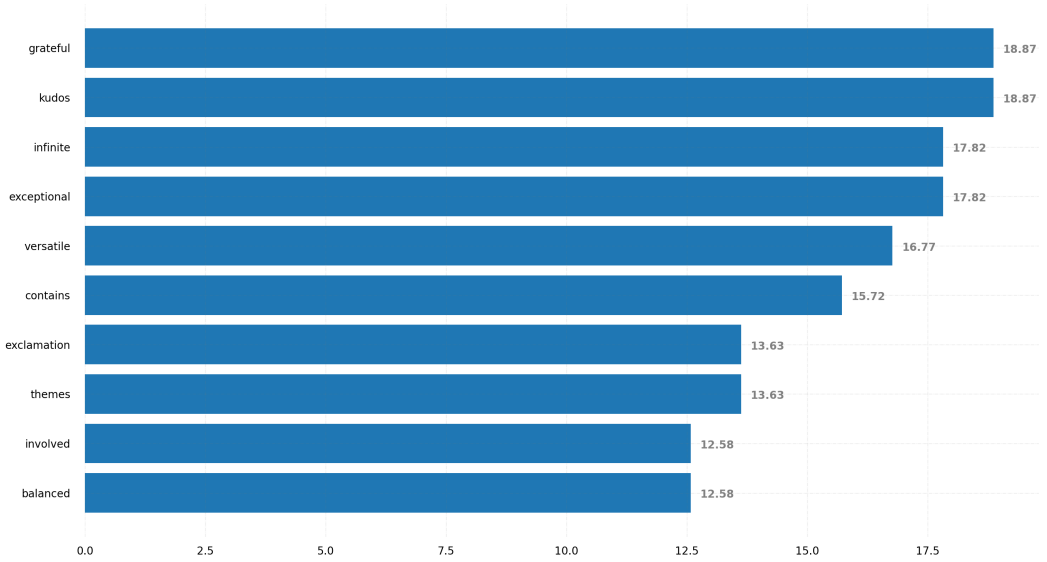
beaming	customize	entertaining	addicting	smiling	outstanding	impressive	master	engaging	grateful
48.2	33.2	30.0	28.9	28.8	23.6	23.6	19.3	19.3	19.3

۱۸.۸.۸ ۵ ستاره به ۲ ستاره



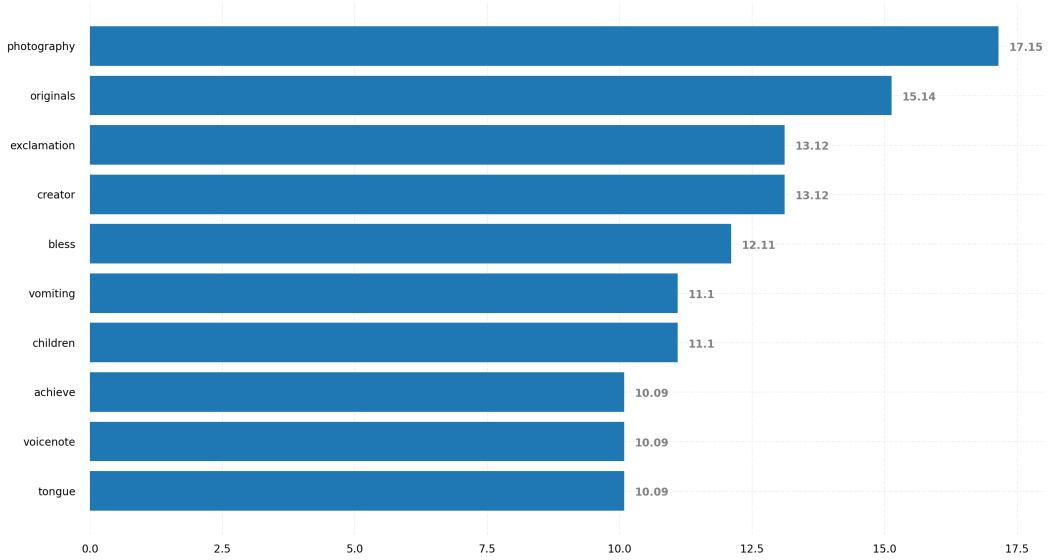
beaming	grinning	brilliant	exceptional	sunglasses	creativity	smiling	originals	truth	intense
46.2	27.7	27.7	17.5	17.5	16.4	16.0	15.4	14.4	14.4

۱۹.۸.۸ ۵ ستاره به ۳ ستاره



grateful	kudos	infinite	exceptional	versatile	contains	exclamation	themes	involved	balanced
18.9	18.9	17.8	17.8	16.8	15.7	13.6	13.6	12.6	12.6

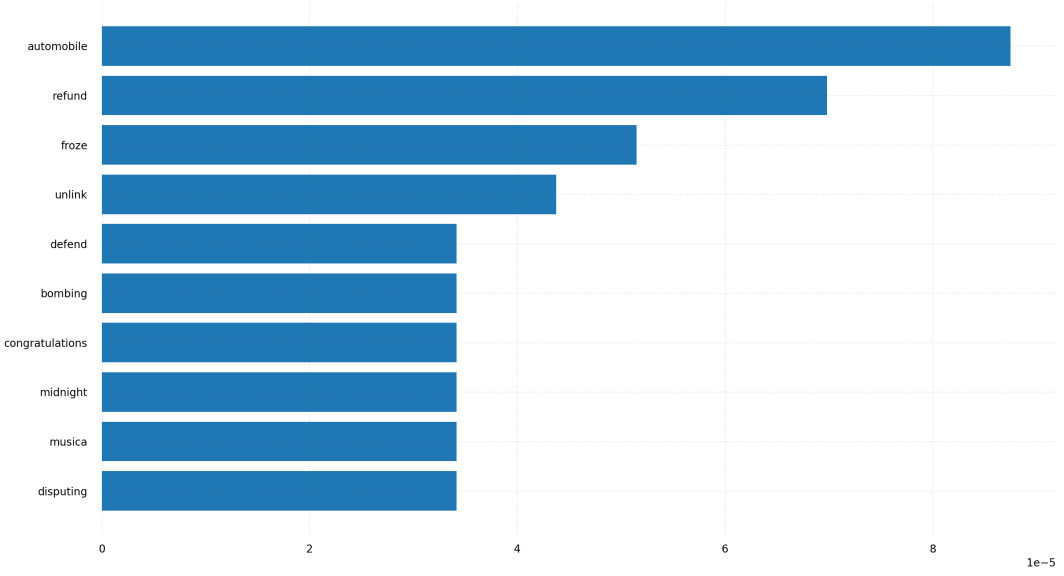
۲۰.۸.۸ ۵ ستاره به ۴ ستاره



photography	originals	exclamation	creator	bless	vomiting	children	achieve	voicenote	tongue
17.2	15.1	13.1	13.1	12.1	11.1	11.1	10.1	10.1	10.1

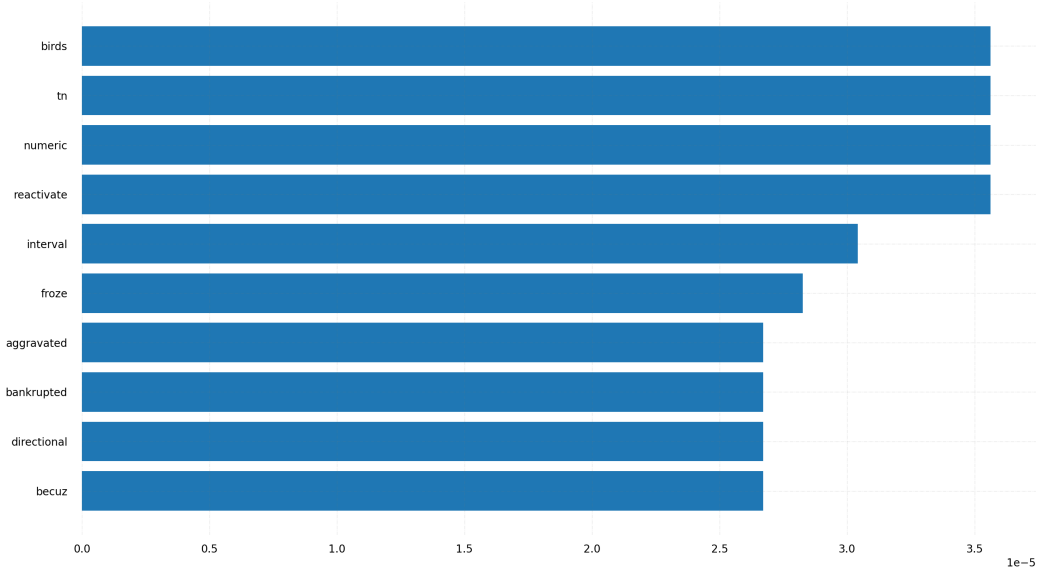
۹.۸ ۱۰ کلمه برتر هر برچسب بر اساس معیار TF-IDF

۱.۹.۸ ۱ ستاره



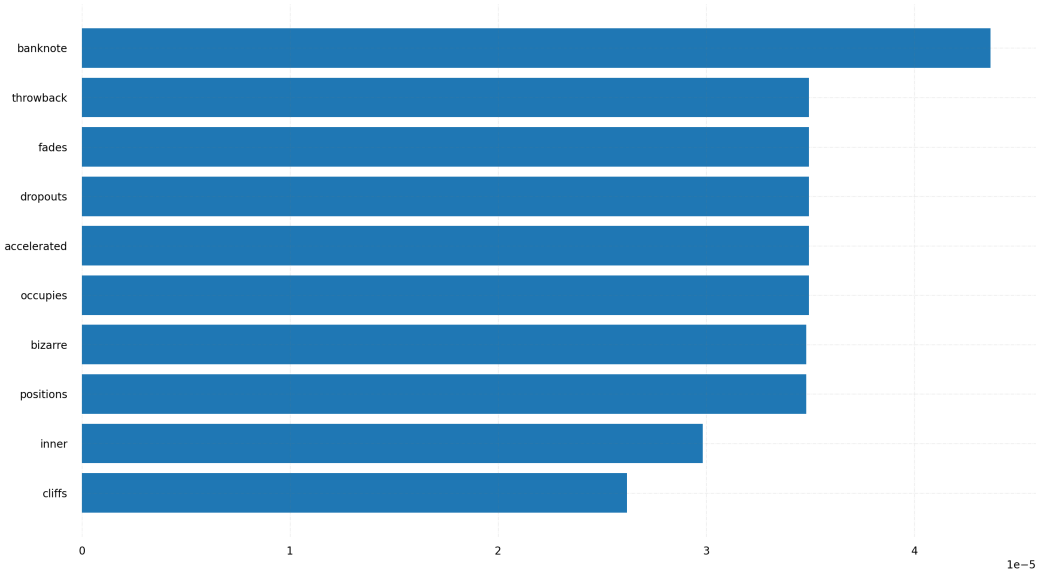
	automobile	refund	froze	unlink	defend	bombing	congratulations	midnight	musica	disputing
TF-IDF	0.000087	0.000070	0.000051	0.000044	0.000034	0.000034	0.000034	0.000034	0.000034	0.000034

۲.۹.۸ ۲ ستاره



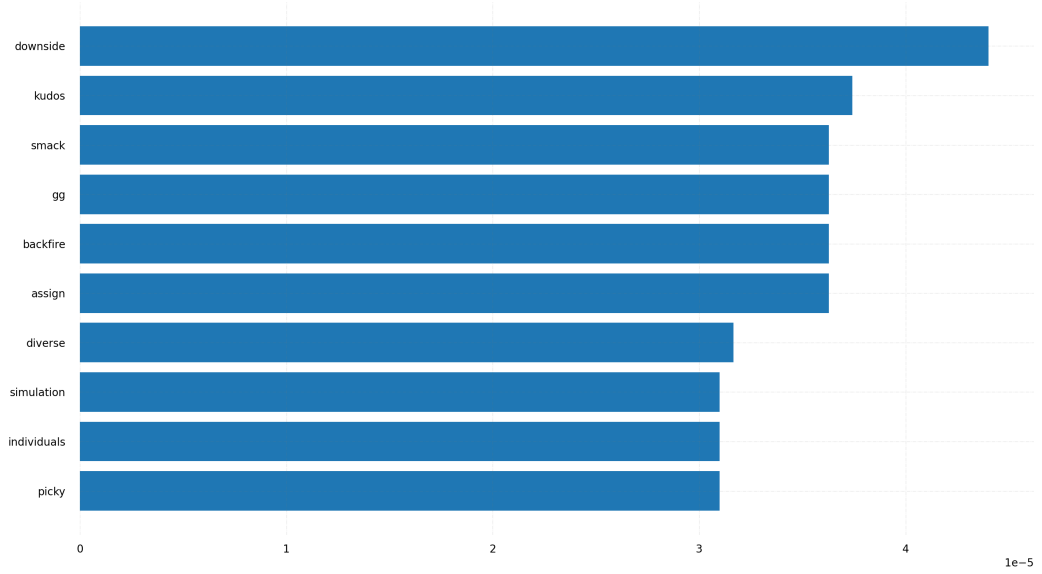
	birds	tn	numeric	reactivate	interval	froze	aggravated	bankrupted	directional	becuz
TF-IDF	0.000036	0.000036	0.000036	0.000036	0.000030	0.000028	0.000027	0.000027	0.000027	0.000027

۳.۹.۸ ۳ ستاره



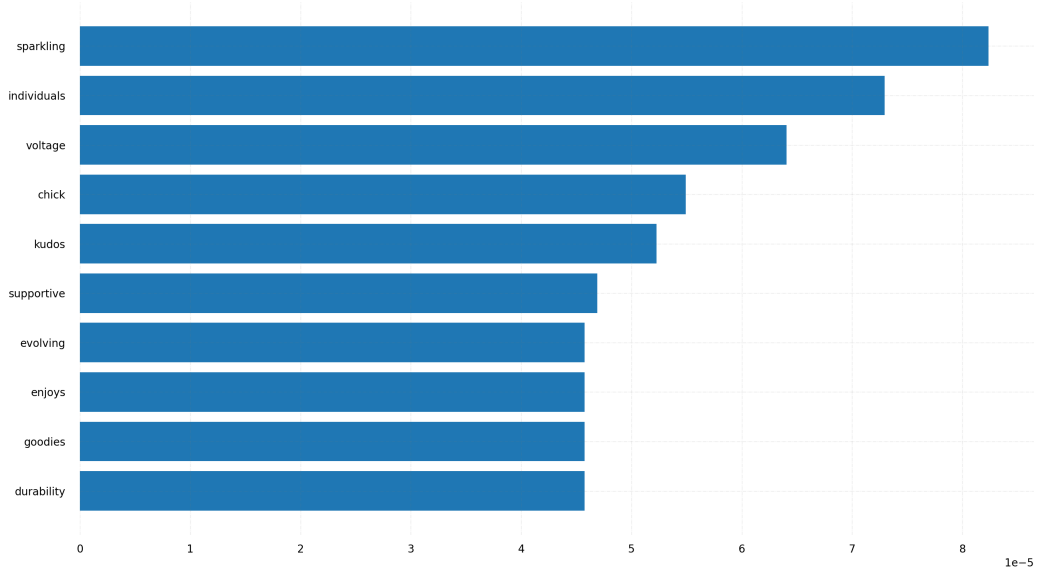
	banknote	throwback	fades	dropouts	accelerated	occupies	bizarre	positions	inner	cliffs
TF-IDF	0.000044	0.000035	0.000035	0.000035	0.000035	0.000035	0.000035	0.000035	0.000030	0.000026

۴.۹.۸ ۴ ستاره



	downside	kudos	smack	gg	backfire	assign	diverse	simulation	individuals	picky
TF-IDF	0.000044	0.000037	0.000036	0.000036	0.000036	0.000036	0.000032	0.000031	0.000031	0.000031

۵.۹.۸ ۵ ستاره



	sparkling	individuals	voltage	chick	kudos	supportive	evolving	enjoys	goodies	durability
TF-IDF	0.000082	0.000073	0.000064	0.000055	0.000052	0.000047	0.000046	0.000046	0.000046	0.000046

۱۰.۸ هیستوگرام تعداد تکرار هر کلمه منحصربفرد