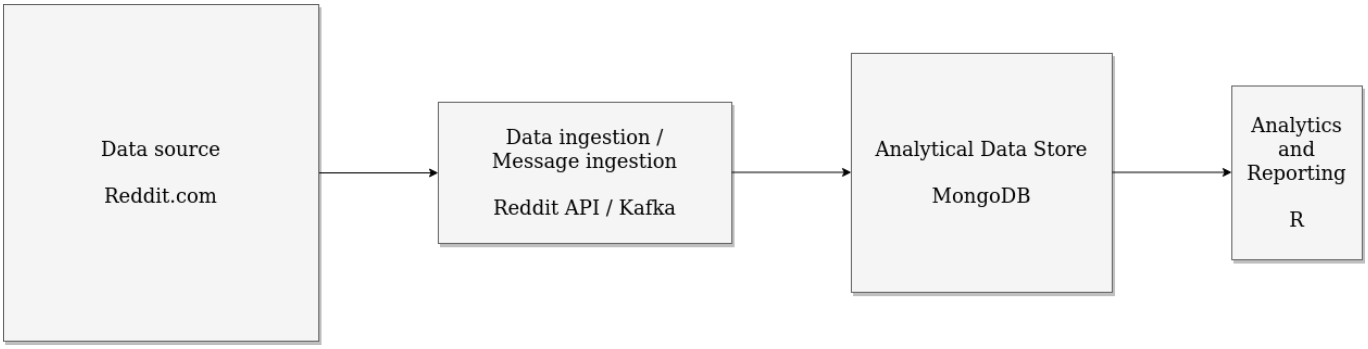
Introduzione

L'obiettivo di questo progetto è stato realizzare ed analizzare un'architettura Big Data, impiegando conoscenze apprese e tecnologie studiate durante il corso. L'architettura realizzata è la seguente:



- **Data source**: come fonte dei dati è stato scelto il forum "Reddit", esso permette agli utenti di pubblicare post relativi agli argomenti più disparati. La piattorma è strutturata in modo tale da separare i macro-argomenti nei cossiddetti "subreddit", ogni subreddit (che ha la sua pagina specifica) si contraddistingue dall'argomento che viene trattato all'interno dei post pubblicati dagli utenti.
- **Data ingestion / Message ingestion**: questa fase è stata gestita attraverso in primo luogo le API ufficiali offerte dalla piattaforma Reddit, successivamente attraverso uno scambio di messaggi Kafka è stato effettuato il salvataggio in locale dei dati recuperati. In questa fase è stato anche eseguito un processo di analisi del sentiment dei dati, dividendo quelli considerati come "postivi" da quelli "negativi".
- **Analytical Data Store**: la tecnologia scelta per memorizzare i dati recuperati è stata MongoDB, un database NoSQL Document-based. Una volta memorizzati, i dati sono subito pronti per la fase di analisi tramite query specifiche.
- **Analytics and Report**: le query di analisi sono state sviluppate utilizzando il linguaggio R, compatibile con MongoDB e molto efficiente. Oltre a ottenere i risultati delle query, R offre la possibilità di costruire in tempo reale dei grafici con i risultati ottenuti, per una loro comprensione più semplice e più immediata.

Data Source

I dati recuperati sono stati relativi ai post di alcuni subreddit, questi subreddit sono stati selezionati in base alla loro popolarità e attività: la piattaforma mette a disposizione un sistema di classifica che segnala quali sono i subreddit più in voga al momento, cioè quelli con maggior affluenza di utenti e nuovi post giornalieri, selezionando questi subreddit in poco tempo si arriva ad avere un buon numero di dati da utilizzare in un'architettura big data.

I subreddit scelti sono stati:

- r/python
- r/cscareerquestions
- r/news
- r/nba
- r/spotify
- r/jobs
- r/tennis
- r/movies
- r/offmychest
- r/depression
- r/foreveralone
- r/anger
- r/europe
- r/gaming
- r/formula1
- r/todayilearned
- r/marvelstudios
- r/healtyfood
- r/politics
- r/askreddit
- r/discordapp
- r/twitch
- r/tinder
- r/techsupport
- r/music
- r/android
- r/baseball
- r/nostupidquestions
- r/explainlikeimfive
- r/outoftheloop
- r/instagram

Questi subreddit appartengono a categorie molto diverse fra loro (news, attualità, sport, intrattenimento, salute, relazioni ecc.), questo perché può essere interessante confrontare realtà diverse e osservare come varia il comportamento e/o il sentimento generale di diverse comunità. Come fonti di dati si sono scelti sempre le pagine già filtrate per post più recenti, in modo da avere sempre nuovi post a disposizione.

Data Ingestion / Message Ingestion

Reddit API

Questa è la fase in cui i dati sono stati attivamente recuperati e salvati in un database locale. La prima fase è stata quella di recuperare i dati dalla piattaforma tramite le API ufficiali. Reddit infatti permette agli utenti di scaricare dal sito post tramite delle semplici richieste GET; una volta creato un account apposito e recuperato le sue credenziali necessarie all'autenticazione, è possibile inviare delle richieste ai server di

Reddit con alcune limitazioni: massimo 60 richieste al minuto, massimo 100 post per ogni richiesta, token di autenticazione valido per massimo 2 ore. Questi limiti però, almeno per questo tipo di progetto, non sono un problema, infatti con la possibilità di recuperare 6000 post al minuto non ci vuole molto per creare un database dal volume sufficiente per essere analizzato nell'ambito dei Big Data. Lo script di estrazione è stato lanciato più volte ma comunque a distanza di ore/giorni, questo perché anche nei subreddit più attivi con milioni di utenti, per avere un certo numero di nuovi post da aggiungere al database è stato opportuno aspettare che venissero effettivamente pubblicati.

Questa prima fase di estrazione dati è stata realizzata mediante un semplice script python, che consiste di tre parti principali:

1. Autenticazione
2. Estrazione dati (esempio a seguire)

```
pythonSub = requests.get("https://oauth.reddit.com/r/python/new",
headers=headers, params={'limit':'100'})
```

3. Salvataggio deli dati recuperati in file JSON, filtrando solo le informazioni utili agli scopi di analisi prefissati e invio sul canale kafka (questo script si comporta da producer, il consumer sarà lo script di sentiment)

```
for response in responseData:
        for post in response.json()["data"]["children"]:
            dataPost = {
                'subreddit': post['data']['subreddit'],
                'user': post['data']['author'],
                'title': post['data']['title'],
                'selftext': post['data']['selftext'],
                'score': post['data']['score'],
                'time': post['data']['created_utc'],
                'commentsCount': post['data']['num_comments']
            }
            producer.send("reddit-posts-dev", dataPost)
```

Sentiment analysis

Prima di salvare i dati su un database in locale, è stata effettuata un'analisi del sentiment per dividere i post negativi da quelli positivi. Ciò è stato fatto perché può essere utile monitorare l'andamento del sentiment generale e confrontare i vari subreddit da un punto di vista del coinvolgimento degli utenti.

Lo script di sentiment è stato realizzato sempre in python, con l'utilizzo della libreria TextBlob, che offre un metodo basato su machine learning per processare in modo semplice un testo e assegnare uno score in base al sentiment registrato. L'analisi del sentiment è stata eseguita sul campo "title" dei post estratti, che è stato ritenuto il più adatto per estrapolare l'umore dell'autore.

Questo microservizio inizialmente consuma i messaggi dal canale kafka per ottenere accesso ai post estratti, successivamente esegue la sentiment analysis e infine sulla base del risultato ottenuto invia come producer un nuovo messaggio sul canale kafka su due topic diversi: "reddit-positive-dev" se lo score ottenuto è maggiore o uguale a 0, "reddit-negative-dev" altrimenti. Questa suddivisione di topic verrà utilizzata dal microservizio che si occupa di salvare i dati in un DB locale, effettivamente si avranno due database distinti, uno per i post considerati positivi, uno per quelli considerati negativi.

```
if __name__ == '__main__':

    consumer = KafkaConsumer("reddit-posts-dev",
auto_offset_reset='earliest',
    enable_auto_commit=True, group_id=None,
                             value_deserializer=lambda x:
loads(x.decode('utf-8')))

    producer = KafkaProducer(value_serializer=lambda v:
json.dumps(v).encode('utf-8'))

    for message in consumer:

        message = message.value
        postText = message["title"] + message["selftext"]

        # converting to blob and sentiment analysis of the post
        blob = TextBlob(postText)
        sentiment = blob.sentiment.polarity

        message["scoreSentiment"] = sentiment

        # sending posts to other kafka microservices
        if sentiment >= 0:
            producer.send("reddit-positive-dev", message)
        else:
            producer.send("reddit-negative-dev", message)
```

Analytical Data Store

MongoDB

La tecnologia scelta per memorizzare i dati estratti è stata MongoDB, un sistema NoSQL document-based. Questa scelta è stata motivata principalmente da due fattori: in primo luogo le API ufficiali della piattaforma Reddit potrebbero cambiare, aggiornarsi e rendere disponibili nuovi tipi di informazioni, un database non relazionale come Mongo permette di aggiungere nuove colonne semplicemente e in modo scalabile, cosa che un database relazionale classico non può offrire.

Non è stato scelto un Key-Value database in quanto solitamente sistemi di questo tipo vengono utilizzati in casi in cui sono presenti dati relativi ad informazioni personali degli utenti e riguardo le loro sessioni all'interno delle piattaforme interessate, inoltre sono sconsigliati quando si eseguono query by data, cosa che invece fa parte degli obiettivi del progetto.

Database document-based, così come quelli column-family (per esempio Cassandra), sono consigliati nel caso in cui si debba interagire con piattaforme web per eseguire operazioni di analisi (entrambi i sistemi sono anche compatibili con R, tool molto usato per l'analisi). La decisione finale è stata indirizzata verso MongoDB prettamente per ragioni di maggiore dimestichezza nel suo utilizzo.

Struttura del database

Come anticipato nella sezione relativa alla data ingestion (sotto-sezione Reddit API), i campi dei post recuperati che sono stati effettivamente estratti sono i seguenti:

Campo	Descrizione
_id	Titolo del post utilizzato come id del record
_class	Tipo di oggetto salvato, corrisponde al tipo della classe Java associata
commentsCount	Numero di commenti presenti sul post
score	Score totale del post, differenza tra upvote e downvote dati dagli utenti
selftext	Corpo del testo del post
subreddit	Subreddit in cui è stato postato il post
time	Data del post
user	Autore del post
scoreSentiment	Score registrato dalla funzione di sentiment (TextBlob)

Si nota che come id è stato scelto il titolo, questo perché tra tutti i parametri è quello che ha maggiori probabilità di essere univoco ed inoltre la piattaforma Reddit impedisce agli autori dei post di modificarlo. In questo modo, nel caso in cui ci fosse poca attività in un subreddit, nel caso gli script di ingestion ripescassero di nuovo un post già salvato, esso verrebbe scartato garantendo l'assenza di duplicati nel database.

Microservizi Java

Sono stati svilupatti due microservizi (questa volta in Java e non più Python) che si occupano di memorizzare i dati nel database locale. Sono due microservizi distinti in quanto si occupano di raccogliere i due diversi risultati ottenuti dallo script di sentimenti e inviati sul canale Kafka su due topic diversi: "reddit-positive-dev" e "reddit-negative-dev". Il primo microservizio raccoglie tutti i post considerati come positivi e li memorizza in un primo database, mentre il secondo microservizio tutti quei post considerati come negativi e li memorizza in un secondo database distinto.

Ciascun microservizio si interfaccia al sistema MongoDB appoggiandosi allo strumento offerto dalla libreria SpringBoot. Il supporto nativo di SpringBoot per i database di Mongo è stata una delle motivazioni per cui è stato scelto di utilizzare Java come linguaggio di programmazione di questi due microservizi.

A seguire il codice utilizzato in questi due microservizi, a titolo di esempio verrà mostrato quello relativo ai post "positivi":

1. **Model**

```
@Getter @Setter
@NoArgsConstructor @AllArgsConstructor
@Builder @ToString
@Document(collection = "positive_reddit_posts")
public class RedditPost {

    private String id;

    private String subreddit;

    private String user;

    @Id
    private String title;

    private String selftext;

    private Float score;

    private Date time;

    private Float commentsCount;

    private Float scoreSentiment;

}
```

Viene definita una classe Java, "RedditPost", sarà utilizzata per tutti gli oggetti che verrano creati a partire dai messaggi ricevuti dal canale Kafka e che poi verranno memorizzati nel database. Tramite l'annotazione di SpringBoot **@Document** si segnala già che la classe sarà utilizzata per oggetti creati appositamente per essere salvati in un database Mongo.

2. **Messaging**

```
@Slf4j
@Service
public class KafkaPositiveRedditListener {

    @Autowired
    private RedditPostService redditPostService;
```

```
    private static final String TOPIC_NAME = "reddit-positive-dev";

    @KafkaListener(topics = TOPIC_NAME, groupId = "group-id")
    public void consumeMessage(@Payload String post) {
        JSONObject jsonObj = new JSONObject(post);

        RedditPost r = RedditPost.builder()
                .title(jsonObj.getString("title"))
                .user(jsonObj.getString("user"))
                .commentsCount(jsonObj.getFloat("commentsCount"))
                .selftext(jsonObj.getString("selftext"))
                .subreddit(jsonObj.getString("subreddit"))
                .score(jsonObj.getFloat("score"))
                .time(new
java.util.Date((long)jsonObj.getFloat("time")*1000))
                .scoreSentiment(jsonObj.getFloat("scoreSentiment"))
                .build();

        redditPostService.savePost(r);
        log.info(r.toString());
    }
```

Si prende dal canale Kafka tutti i messaggi inviati sul topic selezionato (in questo caso "reddit-positive-dev"). Ciascun messaggio viene elaborato come oggetto JSON, che viene deserializzato in un oggetto Java di tipo RedditPost, i cui attributi saranno proprio i campi selezionati precedentemente durante la fase di ingestion. A questo punto ogni oggetto creato viene salvato all'interno di una repository (se ne occuperà Service).

3. **Service**

```
@Service
public class RedditPostService {

    @Autowired
    private RedditItPostRepository redditItPostRepository;

    public void savePost(RedditPost redditPost){
        redditItPostRepository.save(redditPost);
    }

    public List<RedditPost> retrieveAllPost(){
        return redditItPostRepository.findAll();
    }
}
```

La classe Service viene utilizzata come di consueto in un'architettura SpringBoot per salvare oggetti all'interno di repository e quindi database, sono quindi presenti il metodo "save(RedditPost)" e "retrieveAllPost()", il primo si occupa del salvataggio degli oggetti (sovrascrivendo quello di default in

modo che accetti oggetti di tipo RedditPost) mentre il secondo può essere utile per future operazioni di analisi.

Analytics and Report

Per la fase di analisi dei dati estratti, si è ipotizzato di dover confrontare i vari subreddit fra di loro, cercando di osservare differenze per quanto riguarda la popolarità di certi argomenti, quali possano essere quelli attrivuibili ad umori più positivi e quali più negativi, osservando inoltre se la risposta da parte di ciascuna community rispecchia quei determinati umori.

La tecnologia scelta per questa fase è stato il linguaggio R, questa decisione è stata presa sulla base del gran numero di strumenti offerti dal linguaggio, sia dal punto di vista delle query realizzabili sia per il fatto che, tramite l'IDE ufficiale RStudio è possibile visualizzare in tempo reale grafici di semplice interpretazione riguardo le query eseguite. Inoltre, R offre compatibilità completa con l'infrastruttura MongoDB, garantendo alta efficienza.

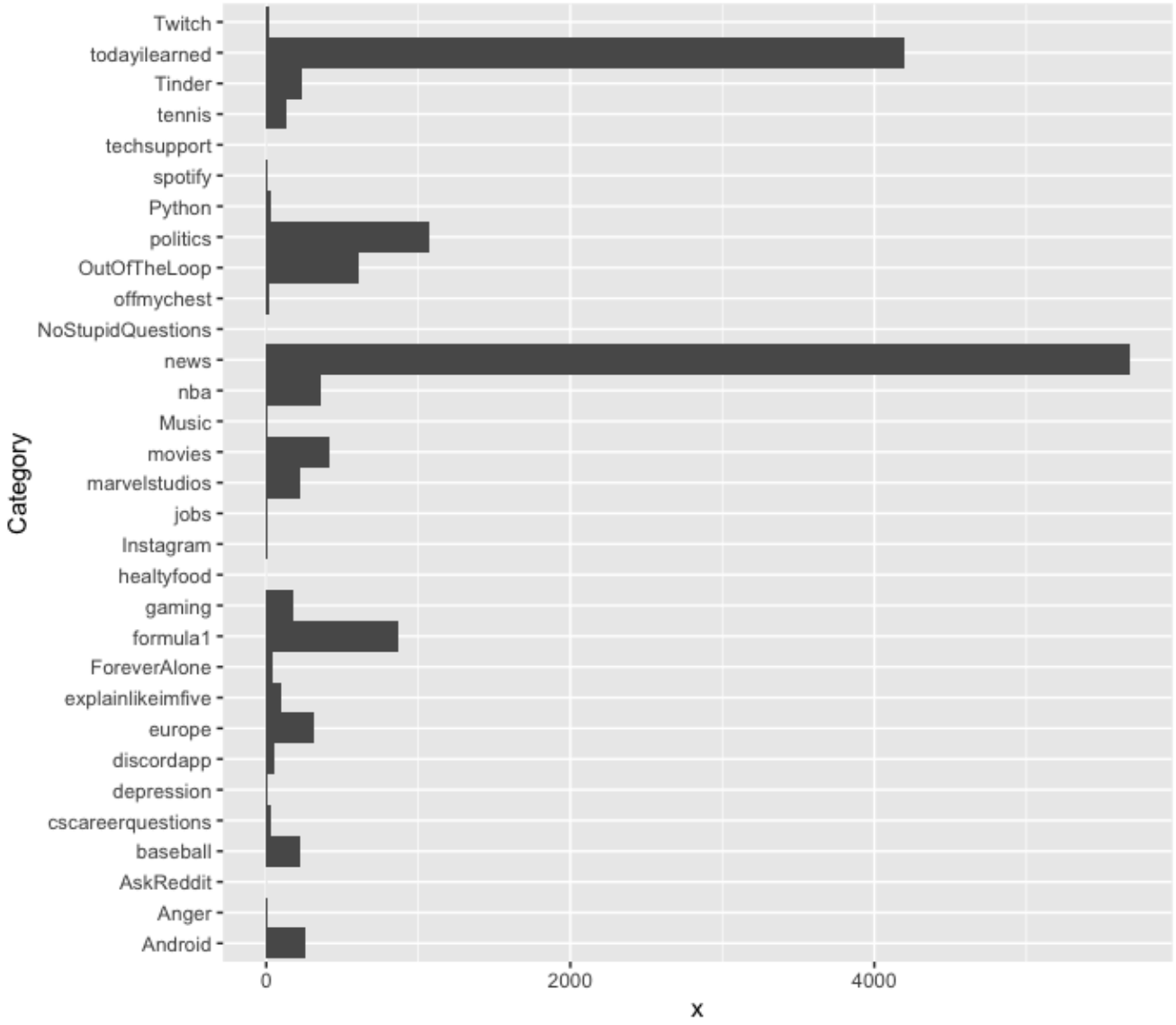
Analisi

1. subredditByScore/Comments

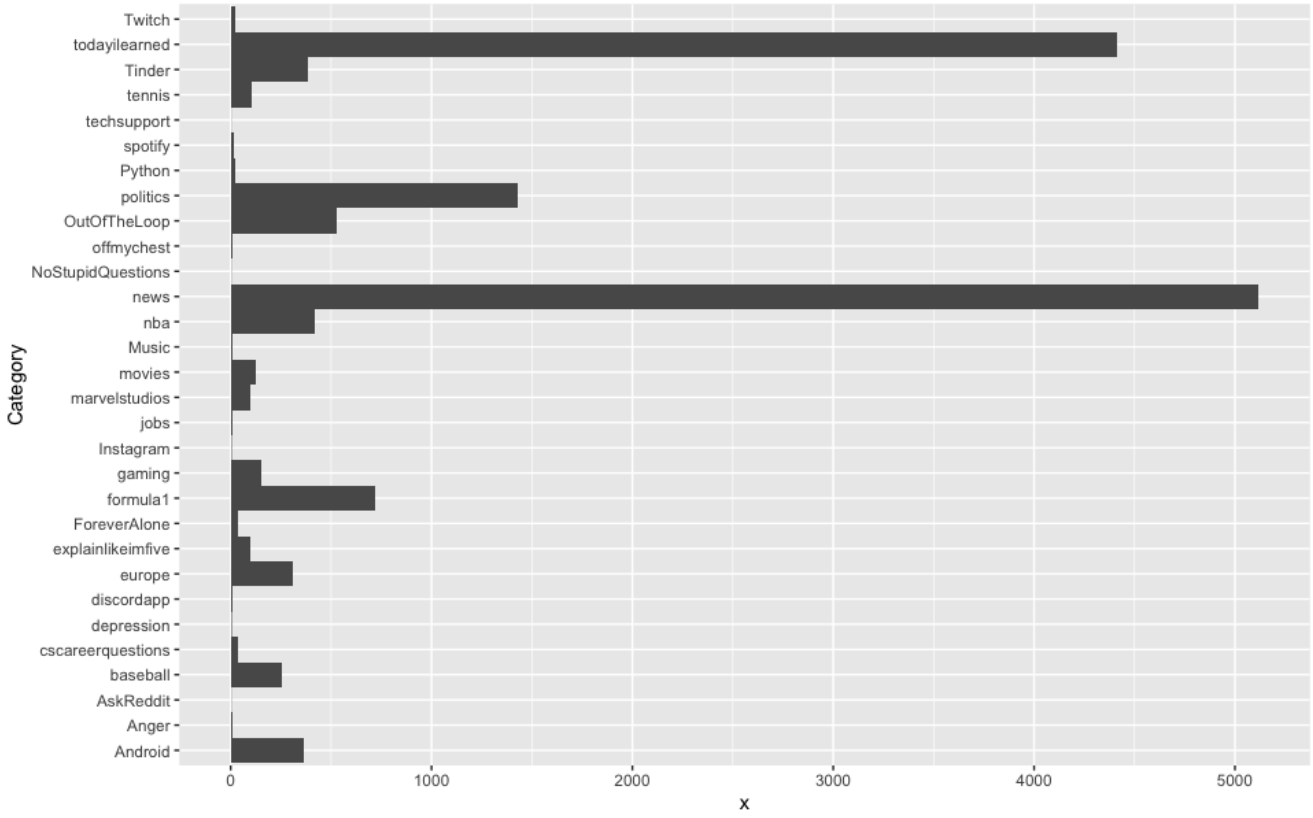
La prima analisi effettuata è stata quella di confrontare l'approvazione dei post positivi e quella dei negativi, osservando quali subdreddit fossero più sbilanciati in un senso in un database, anche in relazione alla loro controparte nell'altro DB.
Per fare ciò è stata realizzata una query che eseguisse una media di Score e numero di commenti dei post, raggruppati per subreddit di appartenenza. Questo perché se un post ha uno Score alto e/o un alto numero di commenti, è probabile che gli altri utenti della community condividano opinioni e umori degli autori dei post.

```
subredditByScore = aggregate(data$score, by=list(Category=data$subreddit),
FUN=mean)

ggplot(subredditByScore, aes(x=x, y=Category)) + geom_bar(stat='identity',
width=1)
```



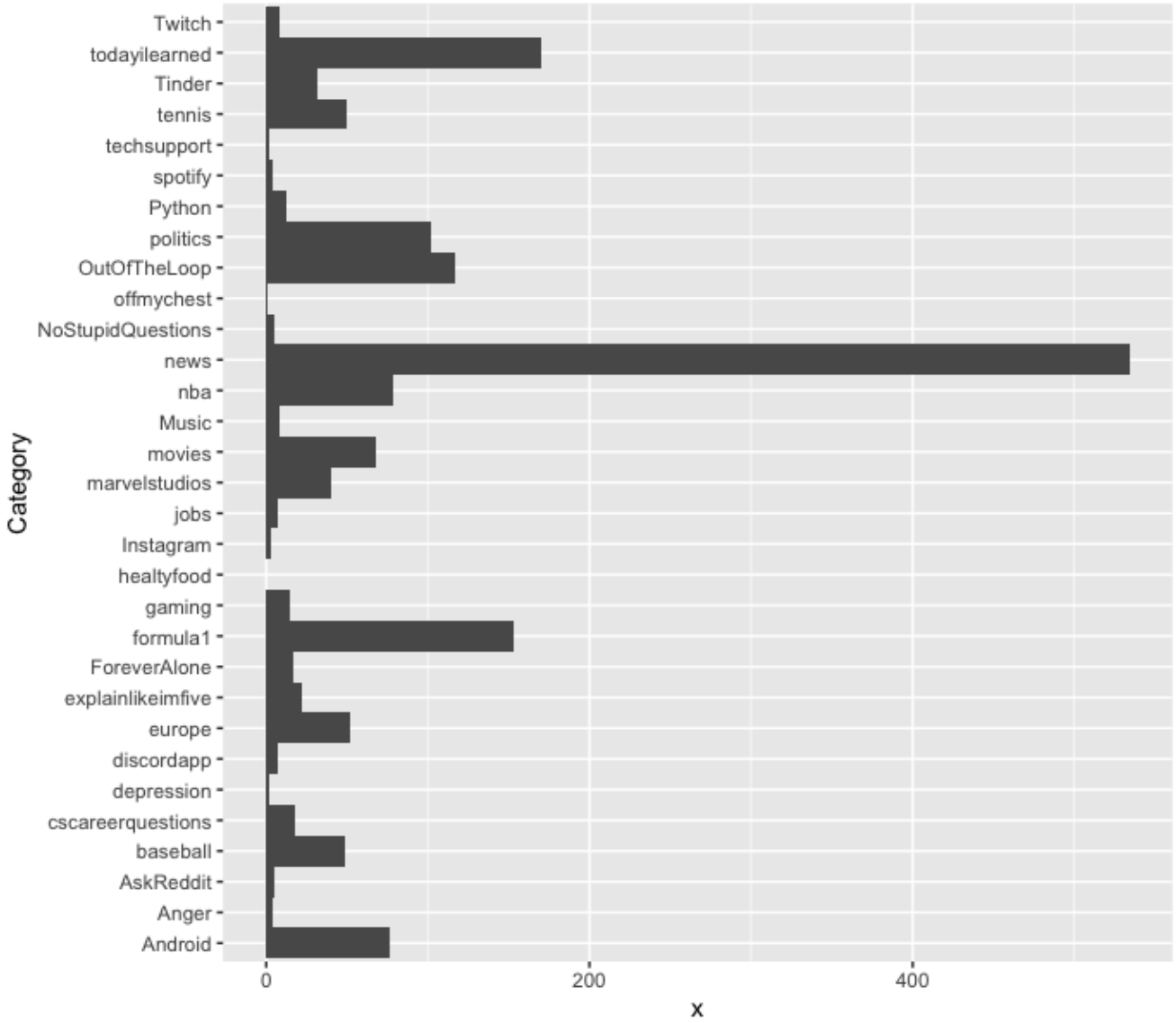
Positive subredditByScore



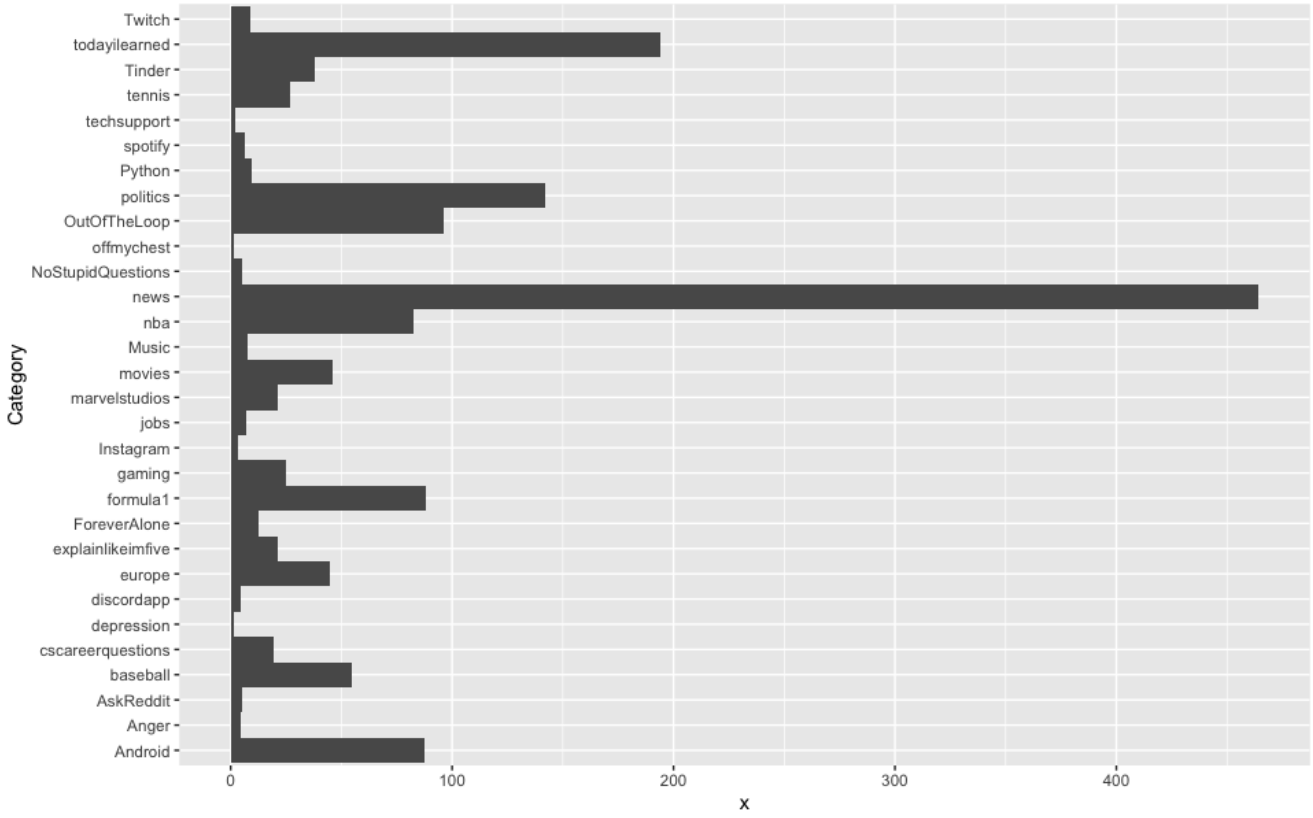
Negative subredditByScore

```
subredditByComments = aggregate(data$commentsCount,
by=list(Category=data$subreddit), FUN=mean)

ggplot(subredditByComments, aes(x=x, y=Category)) +
geom_bar(stat='identity', width=1)
```



Positive subredditByComments



Negative subredditByComments

2. maxScorePost/minScorePost

La seconda analisi mira a estrarre quale post ha ottenuto uno score maggiore/minore, per ogni subreddit. Ciò può essere utile per capire che tipologie di post sono più in voga all'interno di ogni community e che argomenti vengono trattati all'interno di questi post, o viceversa quali sono gli argomenti da non trattare.

```
group <- as.data.table(data)

maxScorePost <- group[group[, .I[which.max(score)], by=subreddit]$V1]
```

	X_id	X_class	commentsCount	score	scoreSentiment	selftext	subreddit	time	user
1	Am I the only person who's never had a bad experien…	com.example.redditpositive.model.RedditPost	0	312	0.000000000	I see so many people complaining about glitchy exper…	spotify		
2	I'm 33 and can't keep a job longer than a year	com.example.redditpositive.model.RedditPost	1155	1084	0.000000000		jobs	2022-07-16T20:22:24.000Z	UncleNoGood102
3	Wimbledon truly is an exhibition this year	com.example.redditpositive.model.RedditPost	0	5294	0.000000000		tennis		
4	'Reign of Fire' Was a Star-Studded Dragon Epic Done …	com.example.redditpositive.model.RedditPost	1908	40113	0.221428573		movies	2022-07-07T23:42:56.000Z	Rebel_Saint
5	My boyfriend butt dialed me	com.example.redditpositive.model.RedditPost	0	8012	0.000000000	My boyfriend was at the gym with his brothers when …	offmychest		
6	About the Sub, Common Misconceptions, and an In-d…	com.example.redditpositive.model.RedditPost	0	410	0.000000000	Hello everyone and welcome to r/foreveralone ! Forev…	ForeverAlone		
7	Our most-broken and least-understood rules is "help…	com.example.redditpositive.model.RedditPost	0	2364	0.000000000	We understand that most people who reply immediat…	depression		
8	It's been years since this happened and I'm still speec…	com.example.redditpositive.model.RedditPost	661	29196	0.000000000		Tinder	2022-07-12T02:03:44.000Z	Tcrizzlez
9	Python is the 2nd most demanded programming lang…	com.example.redditpositive.model.RedditPost	132	800	0.250000000		Python	2022-07-07T13:41:20.000Z	__dacia__
10	Scrum and "Agile" has ruined every job I've ever work…	com.example.redditpositive.model.RedditPost	640	3958	0.045757577	The story is always the same. 1) Join a company as a …	cscareerquestions	2022-07-22T13:20:00.000Z	edgeless779
11	Pregnant Texas woman driving in HOV lane told polic…	com.example.redditpositive.model.RedditPost	5275	122574	0.333333343		news	2022-07-08T23:49:20.000Z	PeliPal
12	The prime minister of Finland, Sanna Marin, is leading…	com.example.redditpositive.model.RedditPost	1281	26986	0.000000000		europe	2022-07-03T11:16:16.000Z	squirrel-bear
13	Lindsey Graham and Rudy Giuliani subpoenaed in Geo…	com.example.redditpositive.model.RedditPost	2238	71866	0.000000000		politics	2022-07-05T18:16:32.000Z	alicen_chains
14	[Charania] Kevin Durant has requested a trade out of …	com.example.redditpositive.model.RedditPost	0	35312	0.000000000		nba		
15	Get help now	com.example.redditpositive.model.RedditPost	11	26	0.014285714	To anyone in here who has anger issues and hasn't s…	Anger	2022-07-18T02:57:04.000Z	ShakespearesNads
16	Open world, technically	com.example.redditpositive.model.RedditPost	1429	68538	0.000000000		gaming	2022-07-14T11:56:48.000Z	Alzward
17	Daniel Ricciardo on his instagram	com.example.redditpositive.model.RedditPost	1406	27375	0.000000000		formula1	2022-07-13T09:59:28.000Z	Clorox43xD
18	TIL Michael Jackson wore white tape on his fingers so …	com.example.redditpositive.model.RedditPost	1932	76369	0.000000000		todayilearned	2022-07-17T12:03:12.000Z	GregJamesDahlen
19	I've seen this argument a whole lot on here…	com.example.redditpositive.model.RedditPost	1437	14642	0.200000003		marvelstudios	2022-07-20T02:50:40.000Z	Mr_Skeletal
20	Six Healthy Dinner Ideas to Kick Start Your Weight Los…	com.example.redditpositive.model.RedditPost	0	1	0.279837072	https://preview.redd.it/n19vxmw2nac…	healtyfood	2022-01-17T18:46:24.000Z	Used_Assignment9425
21	What is your favorite action scene from a movie or TV…	com.example.redditpositive.model.RedditPost	17	7	0.300000012		AskReddit	2022-07-01T17:21:04.000Z	kleptomania156
22	Is this against the tos?	com.example.redditpositive.model.RedditPost	140	2994	0.000000000		discordapp	2022-07-17T09:08:16.000Z	MarkTheBoy_YT
23	Twitch has banned trans streamer Keffals for openly t…	com.example.redditpositive.model.RedditPost	79	3808	0.000000000		Twitch	2022-07-21T18:12:16.000Z	handdrawnmoustache
24	Gonna be swapping my Motherboard, any potential is…	com.example.redditpositive.model.RedditPost	10	20	0.072499998	I woke up on my birthday and the first thing I saw wa…	techsupport	2022-07-17T03:22:40.000Z	vulcanfury12
25	Why is no one going against Ticketmaster?	com.example.redditpositive.model.RedditPost	515	1650	0.007291667	Edit about Phish debate: They aren't popular in Europ…	Music	2022-07-01T23:15:12.000Z	Paranoid_Android101
26	[UPDATE] NewPipe, an open source YouTube player g…	com.example.redditpositive.model.RedditPost	317	2222	0.200000003		Android	2022-07-06T17:23:12.000Z	Lawsonator85
27	Astros catcher, Martín Maldonado, sneaks up behind …	com.example.redditpositive.model.RedditPost	622	8498	0.050000001		baseball	2022-07-03T20:09:36.000Z	hanSoes
28	"brainwashed" into believing America is the best?	com.example.redditpositive.model.RedditPost	26	31	0.148684204	I'm sure there will be a huge age range here. But im 2…	NoStupidQuestions	2022-07-18T05:32:48.000Z	gofigure37
29	ELI5: Why are the majority of cars able to drive nearly …	com.example.redditpositive.model.RedditPost	1902	11438	0.333333343		explainlikeimfive	2022-07-03T09:25:20.000Z	LonePonderer
30	What is the deal with the American right wanting Fauc…	com.example.redditpositive.model.RedditPost	3113	14300	0.086380951	Brit here. I keep seeing American right-wingers publi…	OutOfTheLoop	2022-07-19T12:33:04.000Z	MagnusDNW
31	The last 90 days for my art account. I'm at such a los…	com.example.redditpositive.model.RedditPost	105	189	0.066666670		Instagram	2022-07-11T11:20:32.000Z	HaleyIncarnate

Dettagli dei post positivi che hanno ottenuto score maggiore in ogni subreddit

	X_id	X_class	commentsCount	score	scoreSentiment	selftext	subreddit	time	user
1	my first day in office	com.example.redditnegative.model.RedditPost	212	1961	-0.01221763	Been working remote for 2 years, and had my first da…	cscareerquestions	2022-07-14T02:05:52.000Z	largic
2	Texas woman speaks out after being forced to carry h…	com.example.redditnegative.model.RedditPost	5627	72250	-0.25000000		news	2022-07-19T12:30:56.000Z	Darth_Kahuna
3	Mychelle Johnson (Miles Bridges' wife) on Instagram: I…	com.example.redditnegative.model.RedditPost	0	20571	0.00000000	"I hate that it has come to this but I can't be silent an…	nba		
4	Complaint mega thread [lyrics + excessive ads]	com.example.redditnegative.model.RedditPost	0	648	0.00000000	Please keep complaints and posts about lyrics and ex…	spotify		
5	Am I wrong for not doing a two weeks?	com.example.redditnegative.model.RedditPost	237	532	-0.08788420	I just accepted an offer that more than doubles my cu…	jobs	2022-07-22T15:08:48.000Z	partiallypoopypants
6	A bad lip reading of Kyrgios in the Wimbledon final	com.example.redditnegative.model.RedditPost	136	5804	-0.34999999		tennis	2022-07-20T13:39:12.000Z	theguywithacamera
7	Official poster for Owen Wilson's 'Secret Headquarters'	com.example.redditnegative.model.RedditPost	1295	7208	-0.40000001		movies	2022-07-16T11:09:52.000Z	GroundbreakingSet187
8	My Girlfriend Saw Me Cry	com.example.redditnegative.model.RedditPost	0	2076	0.00000000	I (22M) am a sentimental, softie type of guy. Show me…	offmychest		
9	4 guys in our small town commited suicide in 2 mont…	com.example.redditnegative.model.RedditPost	50	276	-0.20625000	i keep saying " they are so stupid, why did they suicid…	ForeverAlone	2022-07-09T01:16:48.000Z	Obydan
10	I'm so tired of the yo-yoing.	com.example.redditnegative.model.RedditPost	0	440	0.00000000	I wish getting better was easier. My god. I hate the co…	depression		
11	Why should I choose Anaconda if I can install pandas,…	com.example.redditnegative.model.RedditPost	141	444	-0.12857144	No shade towards people behind Anaconda but all I s…	Python	2022-07-09T11:26:56.000Z	cy_narrator
12	How do u stop being mad about the past and wanting…	com.example.redditnegative.model.RedditPost	15	47	-0.35833332	Genuinely help	Anger	2022-07-07T07:45:04.000Z	geenieyeayea
13	Thousands of people storm the streets of Budapest af…	com.example.redditnegative.model.RedditPost	685	16907	-0.25000000		europe	2022-07-14T11:03:28.000Z	HULYE_VIDEKI
14	Modders can now pull a dr strange in GTA online	com.example.redditnegative.model.RedditPost	430	18493	-0.05000000		gaming	2022-07-17T04:03:12.000Z	ninjawick
15	Little Yuki Tsunoda celebrating on podium	com.example.redditnegative.model.RedditPost	377	24085	-0.18750000		formula1	2022-07-22T15:21:36.000Z	AnkushTheHero
16	TIL that Billie Joe Armstrong once dropkicked a guy in…	com.example.redditnegative.model.RedditPost	2580	68237	-0.03333334		todayilearned	2022-07-22T20:13:52.000Z	derstherower
17	Tony Stark would never have lived to become a hero …	com.example.redditnegative.model.RedditPost	406	9549	-0.34999999		marvelstudios	2022-07-12T01:55:12.000Z	Giff95
18	House Republicans All Vote Against Neo-Nazi Probe o…	com.example.redditnegative.model.RedditPost	3792	51502	-0.10000000		politics	2022-07-14T11:52:32.000Z	newnemo
19	[Serious] What is your workplace horror story?	com.example.redditnegative.model.RedditPost	4	7	-0.33333334		AskReddit	2022-07-23T11:56:48.000Z	Adventurous-Pea-4925
20	826,591 (78.5%) Discord accounts from Jan-Mar this …	com.example.redditnegative.model.RedditPost	132	1199	-0.16249999		discordapp	2022-07-01T02:12:16.000Z	_-GODDERE2D-_
21	I don't think small streamers realise how much follow…	com.example.redditnegative.model.RedditPost	395	2733	-0.02500000		Twitch	2022-07-12T12:45:52.000Z	AlwaysGoBeyond
22	I mean, he could blame it on slow internet connection	com.example.redditnegative.model.RedditPost	424	36149	-0.30625001		Tinder	2022-07-17T16:17:04.000Z	RebeccaWhiteTS
23	tiktok is taking over my life	com.example.redditnegative.model.RedditPost	16	16	-0.43333334	Is there a way to block the constant feed of random ti…	techsupport	2022-07-18T03:20:32.000Z	lillyloserww
24	StubHub is selling gift cards that CANNOT BE REDEEM…	com.example.redditnegative.model.RedditPost	80	805	-0.20500000	Stop buying StubHub gift cards immediately!!! Gift Ca…	Music	2022-07-08T12:20:16.000Z	william-o
25	The Pixel 6 Pro has the worst connectivity and recepti…	com.example.redditnegative.model.RedditPost	555	2711	-1.00000000		Android	2022-07-03T16:06:24.000Z	-protonsandneutrons-
26	Nation Unable To Enjoy Baseball Without Dozens Of Pi…	com.example.redditnegative.model.RedditPost	550	3935	-0.05000000		baseball	2022-07-01T10:33:36.000Z	dwaxe
27	I fully support people who live alternative lifestyles, b…	com.example.redditnegative.model.RedditPost	44	29	-0.23181818		NoStupidQuestions	2022-07-13T16:51:12.000Z	MorrisCody
28	ELI5: Why is Chernobyl deemed to not be habitable fo…	com.example.redditnegative.model.RedditPost	741	9588	-0.16666667		explainlikeimfive	2022-07-20T17:18:56.000Z	Finnsaddlesonxd
29	What's the deal with people calling SCOTUS illegitimate?	com.example.redditnegative.model.RedditPost	1272	7748	-0.34999999	There are several people calling SCOTUS overturning …	OutOfTheLoop	2022-06-26T15:40:48.000Z	SquirrelSultan
30	What the fuck is this format? Look how they massacre…	com.example.redditnegative.model.RedditPost	91	266	-0.40000001		Instagram	2022-07-13T00:06:24.000Z	YoungGremblo

Dettagli dei post negativi che hanno ottenuto score maggiore in ogni subreddit

```
group <- as.data.table(data)

minScorePost <- group[group[, .I[which.min(score)], by=subreddit]$V1]
```

	X_id	X_class	commentsCount	score	scoreSentiment	selftext	subreddit	time	user
1	best of the neighbourhood / arctic monkeys	com.example.redditpositive.model.RedditPost	0	0	1.00000000		spotify	2022-06-30T21:11:28.000Z	valhallawarrior781
2	Do jobs really care if you have an onlyfans account an…	com.example.redditpositive.model.RedditPost	7	0	0.03333334	Or will it not affect employment? It's not an illegal we…	jobs	2022-07-01T13:15:44.000Z	aquarius01gurl
3	Nadal serves be like	com.example.redditpositive.model.RedditPost	1	0	0.00000000		tennis	2022-07-01T12:03:12.000Z	GaaaX
4	'Boomerang' at 30: How Eddie Murphy's Rom-com Cl…	com.example.redditpositive.model.RedditPost	10	0	0.16666667		movies	2022-07-01T11:56:48.000Z	mike___mc
5	Mother wants daughter (12) to lift up shirt so that the…	com.example.redditpositive.model.RedditPost	0	0	0.00000000	Weird or not weird?	offmychest		
6	Being attractive doesn't get you a relationship	com.example.redditpositive.model.RedditPost	0	0	0.00000000	I never really thought about that until my ex once tol…	ForeverAlone		
7	I got my sister pregnant. What do i do?	com.example.redditpositive.model.RedditPost	6	0	0.19722222	So guys I've been having sexual relationships with my…	depression	2022-07-01T09:16:48.000Z	DoubtRelevant5
8	Umm... what	com.example.redditpositive.model.RedditPost	8	0	0.00000000		Tinder	2022-07-01T17:06:08.000Z	chaitanyathengdi
9	6 Usage Patterns for the ThreadPoolExecutor in Python	com.example.redditpositive.model.RedditPost	0	0	0.00000000		Python	2022-07-01T12:37:20.000Z	pmz
10	struggling	com.example.redditpositive.model.RedditPost	6	0	0.03524721	ive been stuck at my parents house since covid, grad…	cscareerquestions	2022-07-01T11:09:52.000Z	Mapppy
11	Blizzard Entertainment Acquires Boston-Based Studio…	com.example.redditpositive.model.RedditPost	0	0	0.00000000		news		
12	EU Court slams Lithuania's Belarus migrant pushbacks	com.example.redditpositive.model.RedditPost	19	0	0.00000000		europe	2022-07-01T13:39:12.000Z	Kairys_
13	Trump leads Biden in hypothetical 2024 match-up: poll	com.example.redditpositive.model.RedditPost	34	0	0.00000000		politics	2022-07-01T16:32:00.000Z	jmoincali
14	KD to the Blazers?	com.example.redditpositive.model.RedditPost	0	0	0.00000000	Blazers get KD Nets get Anfernee Simons, Eric Bledso…	nba		
15	Masks	com.example.redditpositive.model.RedditPost	2	0	0.01875000	I have a problem that's been going on for as long as …	Anger	2022-07-10T08:17:04.000Z	WonderFrosty3
16	What would be the best way for me to approach gami…	com.example.redditpositive.model.RedditPost	0	0	0.12189497	Hey All. Lately I've been back and forth on if the Switc…	gaming	2022-07-01T17:16:48.000Z	mean_emcee
17	Who is this woman next to Toto?	com.example.redditpositive.model.RedditPost	40	0	0.00000000		formula1	2022-07-01T17:04:00.000Z	oaz1
18	TIL The concept for the movie Good Burger was first a…	com.example.redditpositive.model.RedditPost	43	0	0.44166666		todayilearned	2022-07-12T00:14:56.000Z	swampyouthh
19	Dr Strange 2 Spoiler alert!	com.example.redditpositive.model.RedditPost	5	0	0.02325758	So I saw a doctor strange too yesterday and I really lik…	marvelstudios	2022-07-01T16:06:24.000Z	Fantastic_Forever_69
20	Six Healthy Dinner Ideas to Kick Start Your Weight Los…	com.example.redditpositive.model.RedditPost	0	1	0.27983707	https://preview.redd.it/n19vxmw2nac…	healtyfood	2022-01-17T18:46:24.000Z	Used_Assignment9425
21	They say that the whole is greater than the sum of its …	com.example.redditpositive.model.RedditPost	5	0	0.34999999		AskReddit	2022-07-01T17:21:04.000Z	Scoongili
22	I think that discord's staff is very good after not respo…	com.example.redditpositive.model.RedditPost	4	0	0.41999999	# I think that discord's staff is very good after not r…	discordapp	2022-07-01T16:06:24.000Z	TTGamerTT
23	Can I have a similar name/brand on Twitch partner?	com.example.redditpositive.model.RedditPost	11	0	0.08333334	For example, if someone already took arcane_ex and…	Twitch	2022-07-01T14:45:20.000Z	Upper_Ad6798
24	What is the highest DDR5 RAM capacity manufacturer?	com.example.redditpositive.model.RedditPost	0	0	0.47777778	Is it worthy buying ddr5 and replaced with old ddr4sd…	techsupport	2022-07-01T17:16:48.000Z	digi_pointer
25	Songs that you like to recommend that I should listen…	com.example.redditpositive.model.RedditPost	3	0	0.00000000	I want that is similar to like SoFaygo chrome. Wasp - …	Music	2022-07-01T17:04:00.000Z	Sensitive-Exit3530
26	Huawei FreeBuds Pro 2 Review: I don't review wireless…	com.example.redditpositive.model.RedditPost	15	0	0.69999999		Android	2022-06-30T17:44:32.000Z	Areyoucunt
27	Give me your chosen very okay players for the very ok…	com.example.redditpositive.model.RedditPost	31	0	0.64999998		baseball	2022-07-01T15:40:48.000Z	gopeepants
28	Does pickle juice make your hands bigger over time?	com.example.redditpositive.model.RedditPost	4	0	0.10000000	Backstory: Location: Amish market Man handling pic…	NoStupidQuestions	2022-07-01T17:10:24.000Z	SlutForTurtles
29	eli5: How do we breathe? I get that air gets sucked int…	com.example.redditpositive.model.RedditPost	19	0	0.00000000		explainlikeimfive	2022-07-01T14:49:36.000Z	Low-Ad-5229
30	What's up with the Demon Core memes lately?	com.example.redditpositive.model.RedditPost	7	0	0.00000000	https://www.reddit.com/r/surrealmemes/comments/…	OutOfTheLoop	2022-07-01T06:00:32.000Z	OmniVega
31	Help	com.example.redditpositive.model.RedditPost	0	0	0.25757575	I'm logged into my instagram but I want to deactivate…	Instagram	2022-07-20T23:04:32.000Z	traehnekorb

Dettagli dei post positivi che hanno ottenuto score minore in ogni subreddit

	X_id	X_class	commentsCount	score	scoreSentiment	selftext	subreddit	time	user
1	Indeed assessment	com.example.redditnegative.model.RedditPost	13	0	-0.019545455	Hi guys, sorry if this is a question that's already been …	cscareerquestions	2022-06-30T20:01:04.000Z	freddDeeP
2	Kyiv and Moscow agree deal to resume Ukraine grain …	com.example.redditnegative.model.RedditPost	0	2	-0.166666672		news	2022-07-23T12:20:16.000Z	AgileNetwork7
3	Portland getting defensive with Utah's help	com.example.redditnegative.model.RedditPost	73	0	-0.039646465	While highly unlikely, an idea I've had (therefore a ter…	nba	2022-07-01T14:19:44.000Z	DJ_Drayen
4	Radical Islamist on Spotify?	com.example.redditnegative.model.RedditPost	0	0	0.000000000	I was listening to my weekly mix which mainly consist…	spotify		
5	Work from Europe for USA companies without USA W…	com.example.redditnegative.model.RedditPost	1	0	-0.043055557	What is the easiest way to find companies that do not…	jobs	2022-07-01T12:35:12.000Z	a_veseli
6	Why are there no Wimbledon 2022 highlights on You…	com.example.redditnegative.model.RedditPost	3	0	-0.192727268	I get that they want to monetize content. Until last ye…	tennis	2022-07-01T13:28:32.000Z	anonymousyoshi42
7	I dont underatand this conversation in the movie "Dar…	com.example.redditnegative.model.RedditPost	0	0	0.000000000	JOKER : when you and Rachel was being abducted I w…	movies		
8	technoblade fucking died lol	com.example.redditnegative.model.RedditPost	0	0	0.000000000	i just couldnt stop thinking about it since this mornin…	offmychest		
9	I'm average on looks and with a bit of work i could lo…	com.example.redditnegative.model.RedditPost	6	0	-0.017857144	would you do the trade? I'm really not interested in gi…	ForeverAlone	2022-07-04T17:06:08.000Z	X_Anonymous_2020
10	What's wrong with me	com.example.redditnegative.model.RedditPost	0	0	-0.116049379	All of my friends are going out with girls but all my lif…	depression	2022-07-01T11:56:48.000Z	OGbudsandtha
11	Would you classify writing a BFS pathfinder algorithm …	com.example.redditnegative.model.RedditPost	5	0	-0.178571433	I made this pathfinder and visualizer in python using …	Python	2022-06-30T14:30:24.000Z	HesAMagicalPoney55
12	dealing with destructive angry people	com.example.redditnegative.model.RedditPost	0	0	-0.500000000	(english): confused as to why people get angry at me …	Anger	2022-06-27T23:40:48.000Z	RunItAndSee2021
13	Systematic Abuses at EU External Border: Greek Police…	com.example.redditnegative.model.RedditPost	4	0	-0.166666672		europe	2022-07-01T14:36:48.000Z	aknb
14	Dead Space 2 Chapter 10 part 1 Ishimura Again. Fixin…	com.example.redditnegative.model.RedditPost	1	0	-0.200000003	Dead Space 2 Chapter 10 part 1 Ishimura Again. Fixi…	gaming	2022-07-01T17:21:04.000Z	wolfsolus
15	The New Cars Are Violently Bouncing, and F1 Is Looki…	com.example.redditnegative.model.RedditPost	4	0	-0.331818193		formula1	2022-07-01T12:35:12.000Z	AsstBalrog
16	TIL President Nixon allegedly smuggled 3 pounds of c…	com.example.redditnegative.model.RedditPost	0	1	-0.100000001		todayilearned	2022-07-06T19:31:12.000Z	smn2clay
17	You DO NOT want sentry in the mcu	com.example.redditnegative.model.RedditPost	21	0	-0.023121966	Seen a couple fan casts so let me just say how much …	marvelstudios	2022-07-01T17:18:56.000Z	lelwood6_16
18	The Supreme Court's EPA decision could have been m…	com.example.redditnegative.model.RedditPost	4	0	-0.400000006		politics	2022-07-01T15:36:32.000Z	grist
19	Conservatives of reddit, how do you reconcile with th…	com.example.redditnegative.model.RedditPost	15	0	-0.166666672		AskReddit	2022-07-01T17:12:32.000Z	chicagolandnative93
20	Why wont this noti go away??? i closed discord over a…	com.example.redditnegative.model.RedditPost	3	0	-0.050000001		discordapp	2022-06-30T23:38:40.000Z	melonnn11
21	Modded game streams?	com.example.redditnegative.model.RedditPost	11	0	-0.003219697	I'm new to twitch, as per rules I won't advertise my ch…	Twitch	2022-07-01T16:12:48.000Z	RajputDynasty
22	Please stay off tinder	com.example.redditnegative.model.RedditPost	15	0	-0.291666657	Can you poly-partnered people and Fatherless mothe…	Tinder	2022-07-01T01:08:16.000Z	dj55i
23	PC slowing down when multitasking	com.example.redditnegative.model.RedditPost	1	0	-0.085470088	I recently upgraded my PC but Ive noticed that it beco…	techsupport	2022-07-02T12:09:36.000Z	Boiko_boba
24	Dated - Endless Night [Dark Lofi Hip Hop Mix]	com.example.redditnegative.model.RedditPost	0	0	-0.137500003		Music	2022-07-01T16:34:08.000Z	illadvisedrecords
25	Alleged Moto Razr 2022 shows up on Weibo at Lenov…	com.example.redditnegative.model.RedditPost	9	8	-0.100000001		Android	2022-07-20T04:52:16.000Z	HolidayJesus
26	Is there a way to get rid of ads on MLB Film Room?	com.example.redditnegative.model.RedditPost	3	0	-0.187500000	Just started messing around with it again since it first…	baseball	2022-07-01T15:38:40.000Z	I_Smashed_Girl_1x
27	Are people who have no arms allowed to drive cars?	com.example.redditnegative.model.RedditPost	7	0	-0.100000001	I just watched a video of a woman who has no arms d…	NoStupidQuestions	2022-07-01T16:38:24.000Z	AdPrudent1593
28	ELI5: App/Website Development	com.example.redditnegative.model.RedditPost	3	0	-0.208333328	For complex apps doing a lot of work not released by…	explainlikeimfive	2022-07-01T16:06:24.000Z	DGADK
29	What's up with the hate for Short Video DJs?	com.example.redditnegative.model.RedditPost	1	0	-0.169155851	What's up with the hate for Short Video DJs? I think al…	OutOfTheLoop	2022-07-01T00:34:08.000Z	CarsPlanesTrains
30	Account deactivation bug	com.example.redditnegative.model.RedditPost	3	0	-0.166666672	Can't deactivate my account in any possible way, I trie…	Instagram	2022-07-01T16:51:12.000Z	Lonely-Tumbleweed-56

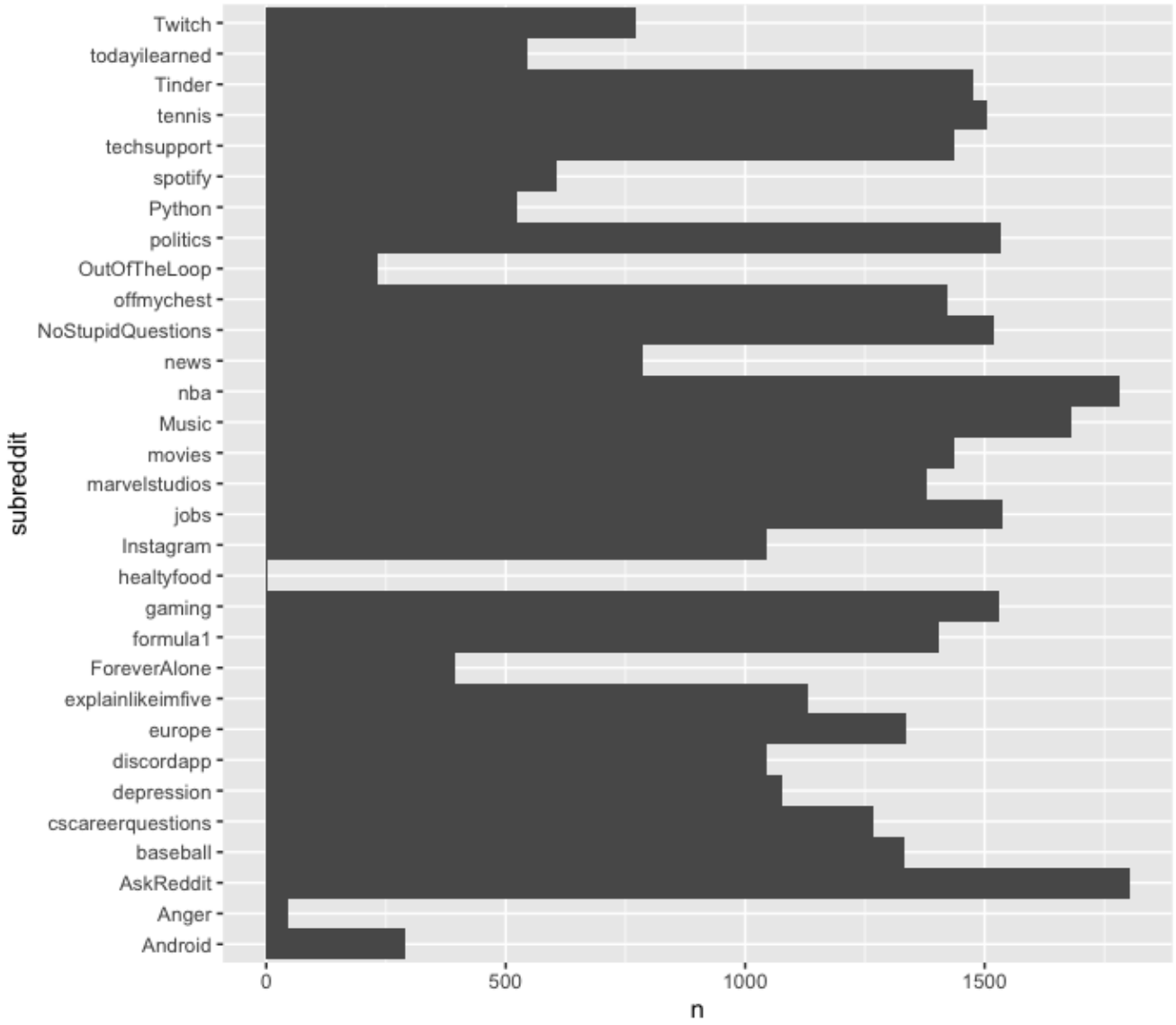
Dettagli dei post negativi che hanno ottenuto score minore in ogni subreddit

3. subredditCount

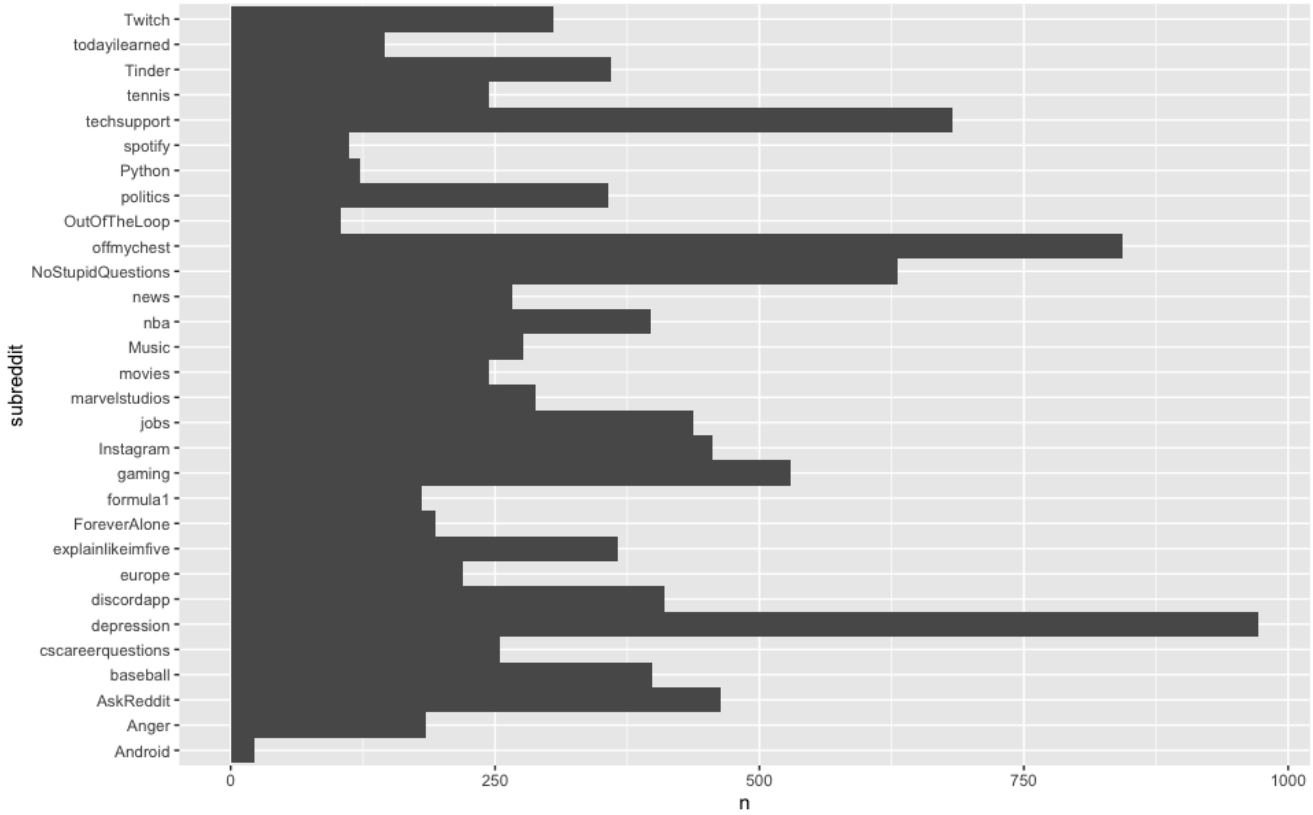
La terza analisi si è incentrata sull'osservare quale fosse il subreddit più presente all'interno di un databse specifico, per monitorare quindi se un certo subreddit contiene più post positivi o negativi.

```
subredditCount = count(data, subreddit = data$subreddit)

ggplot(subredditCount, aes(x=n, y=subreddit)) + geom_bar(stat='identity',
width=1)
```



Positive subredditCount



Negative subredditCount

4. mostPositiveUsers/mostNegativeUsers

La quarta analisi mira a scoprire quali utenti sono più presenti all'interno di diversi subreddit, ciò può essere utile ad identificare quali subvreddit possono avere una community condivisa e quindi più interazioni tra gli stessi utenti o comunque tra utenti con interessi/opinioni comuni.

```
#nel database dei post positivi
mostPositiveUser = count(data, data$user)

#nel database dei post negativi
mostNegativeUser = count(data, data$user)
```

mostPositiveUser

	data$user	n
1		378
15018	moreice45	82
8213	FragmentedChicken	63
9224	GroundbreakingSet187	59
14151	MarvelsGrantMan136	58
1815	Aratho	41
5209	curated_android	39
11811	jovanmilic97	38
19128	RobertGracie	38
19853	SealDrop	38
3406	brandon_the_bald	36
12675	KostisPat257	35
9446	handlit33	34
18378	racingfan96	34
21049	Soupjoe5	34
21496	Stock412	34
16974	Own_Ad6388	33
3518	Bruhmangoddman	31
7534	F1-Bot	31
2424	BalticsFox	29
10737	indig0sixalpha	29
6833	Ecomystic	28
15799	Nexusu	28
1220	AlienSomewhere	27
14772	Miserable-Lizard	27
2486	BaseballBot	25
18283	Quartz1992	25
7123	emkaerr	24
13464	LoneWolfInCyberia	24
15795	Next-Winner279	24
17080	PanEuropeanism	24
1364	amatom27	23
21892	Sweep145	23
12045	KaamDeveloper	22
14764	misana123	22
16197	nosotros_road_sodium	22
24217	vancouver_reader	22
4836	ContentPuff	21

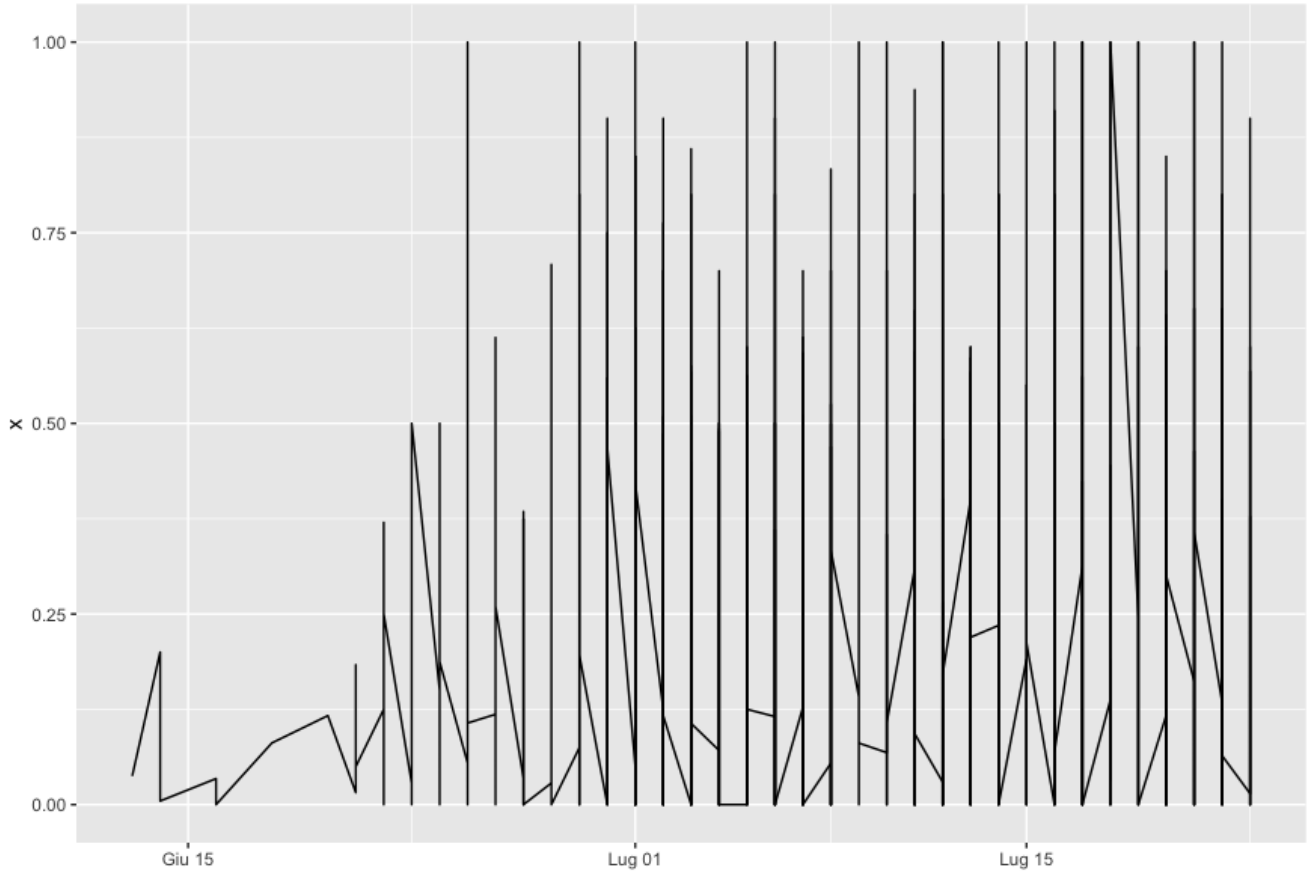
mostNegativeUser

	data$user	n
1		139
1838	CSCQMods	24
5259	moreice45	19
341	ahhlo134	12
3308	handlit33	12
6555	racingfan96	12
1184	brandon_the_bald	11
6042	Own_Ad6388	11
6966	SadMathematician7799	11
7698	Stock412	10
81	[deleted]	9
5166	Miserable-Lizard	9
6435	Professor_Tanaka	8
537	anechointhedark	7
868	BaseballBot	7
1439	Ccbm2208	7
1916	DaFunkJunkie	7
3747	indig0sixalpha	7
4409	koavf	7
5814	ODB95	7
1174	BoysenberryStatus767	6
1177	Brady331	6
2433	EdwardBliss	6
2747	fetuswut	6
4059	JJPJ	6
4137	jovanmilic97	6
5161	misana123	6
5532	NevermoreSEA	6
5709	nosotros_road_sodium	6
5771	Numani99	6
655	Aratho	5
1043	Black_wolf_disease	5
1250	Bruhmangoddman	5
1282	Bulletz4Brkfzt	5
1461	ChamberDavs	5
1568	Cinderace1	5
1574	city_basso	5
3260	Gullible_Peach	5
3389	HeinieKaboobler	5

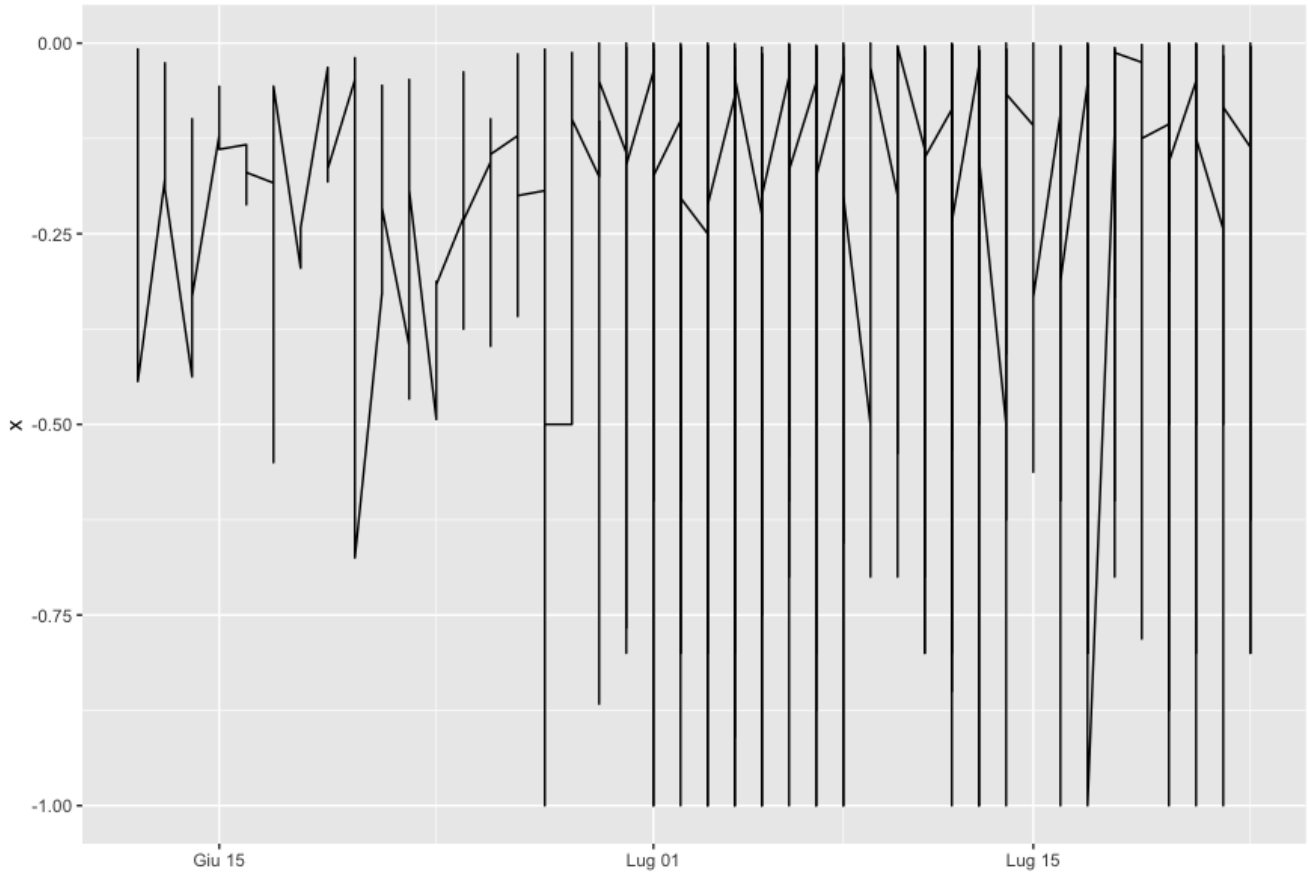
5. scoreSentimentTrend

La quinta analisi mira a scoprire i trend temporali dello score sentiment, ciò può essere utile per monitorare l'andamento dell'umore all'interno delle community prese d'esame.

```
# andamento nel tempo dello score tra i post positivi
scoreByDate = aggregate(data$scoreSentiment, by=list(Category=data$time),
FUN=mean)
scoreByDate[['Category']] <- as.POSIXct(scoreByDate[['Category']],format =
"%Y-%m-%d")
q <- subset(scoreByDate, Category> "2022-06-06" & Category < "2022-12-12")
p <- ggplot(q, aes(x=Category, y=x)) + geom_line() + xlab("")
show(p)
```



Andamento del sentiment score dei post positivi



Andamento del sentiment score dei post negativi

Conclusioni

In sintesi, durante lo svolgimento di questo progetto è stata realizzata ed analizzata un'architettura Big Data, a partire dalla data ingestion fino ad arrivare all'analisi dei dati estratti.

Sono stati estratti e memorizzati dati a partire dalla piattaforma Reddit mediante appositi script/microservizi Python e Java (con l'ausilio del framework Spring) su database non relazionali MongoDB, per poi essere analizzati attraverso R per ricavare le informazioni descritte precedentemente.

E' stato senza dubbio interessante approfondire alcune tra le tecnologie studiate durante il corso di Big Data (tra le quali Kafka, MongoDB, R) ottenendo dei risultati certamente soddisfacenti.