

# Net Zero Emissions: un confronto tra due thread

Analisi su **Reddit** sull'obiettivo della neutralità climatica

Amedeo Onnis

Laboratorio Metodi Digitali per la Ricerca Sociale

A.A. 2023/2024



# Un'analisi sulla **neutralità climatica**

Il progetto di fine laboratorio consiste in una breve analisi linguistica sulla **neutralità climatica**. L'obiettivo *Net Zero Emissions* è stato fissato nel 2015 dopo la firma dell'Accordo sul clima di Parigi, in cui i Paesi firmatari si sono impegnati ad azzerare le emissioni di CO2 nell'atmosfera entro il 2050.



# Gli strumenti adottati



L'analisi è stata  
effettuata con il  
linguaggio **R**



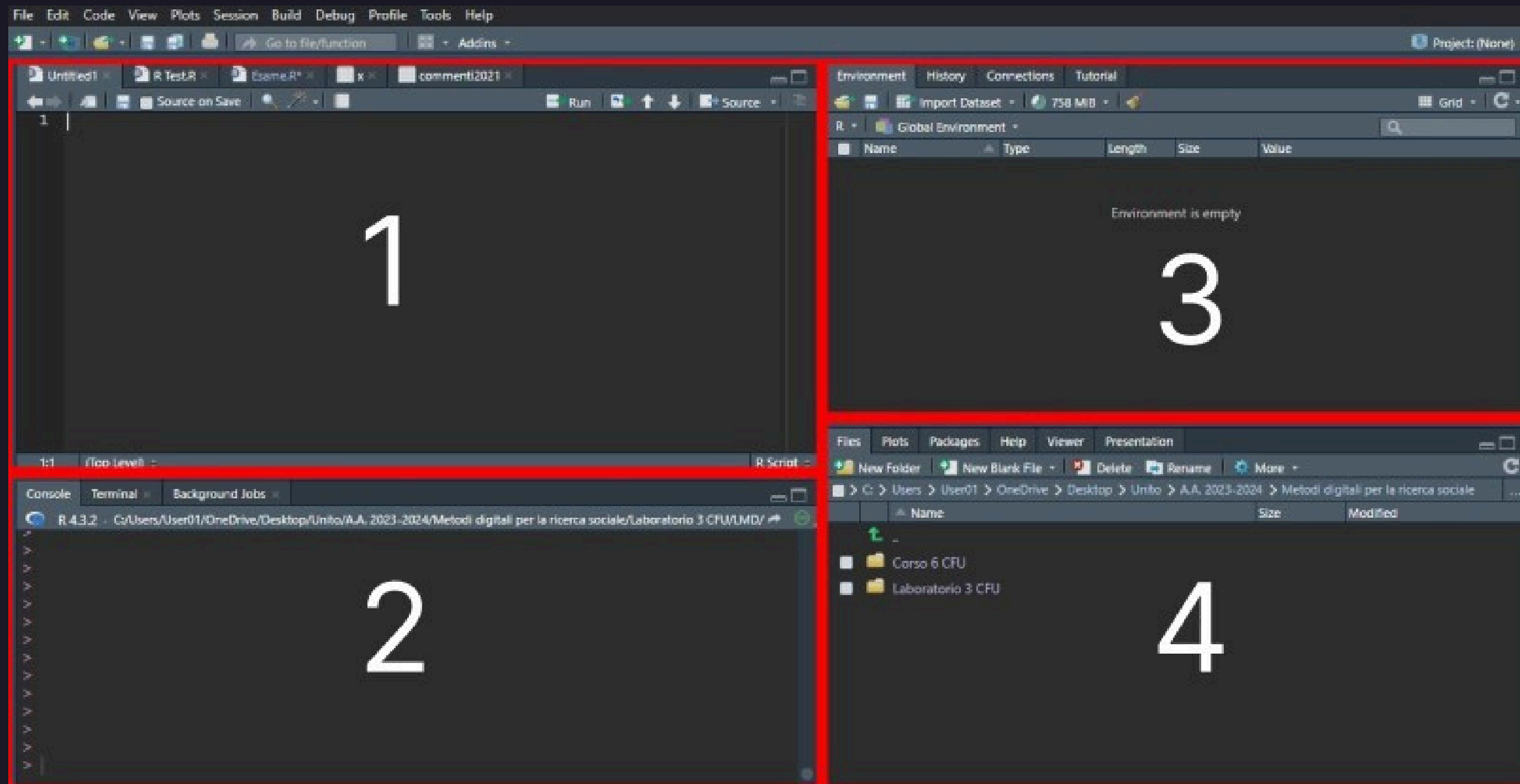
I due thread sono stati  
scaricati dalla  
piattaforma **Reddit**



L'analisi è stata  
eseguita nello  
spazio di **RStudio**

# Focus: l'ambiente RStudio

04



# Focus: l'ambiente RStudio

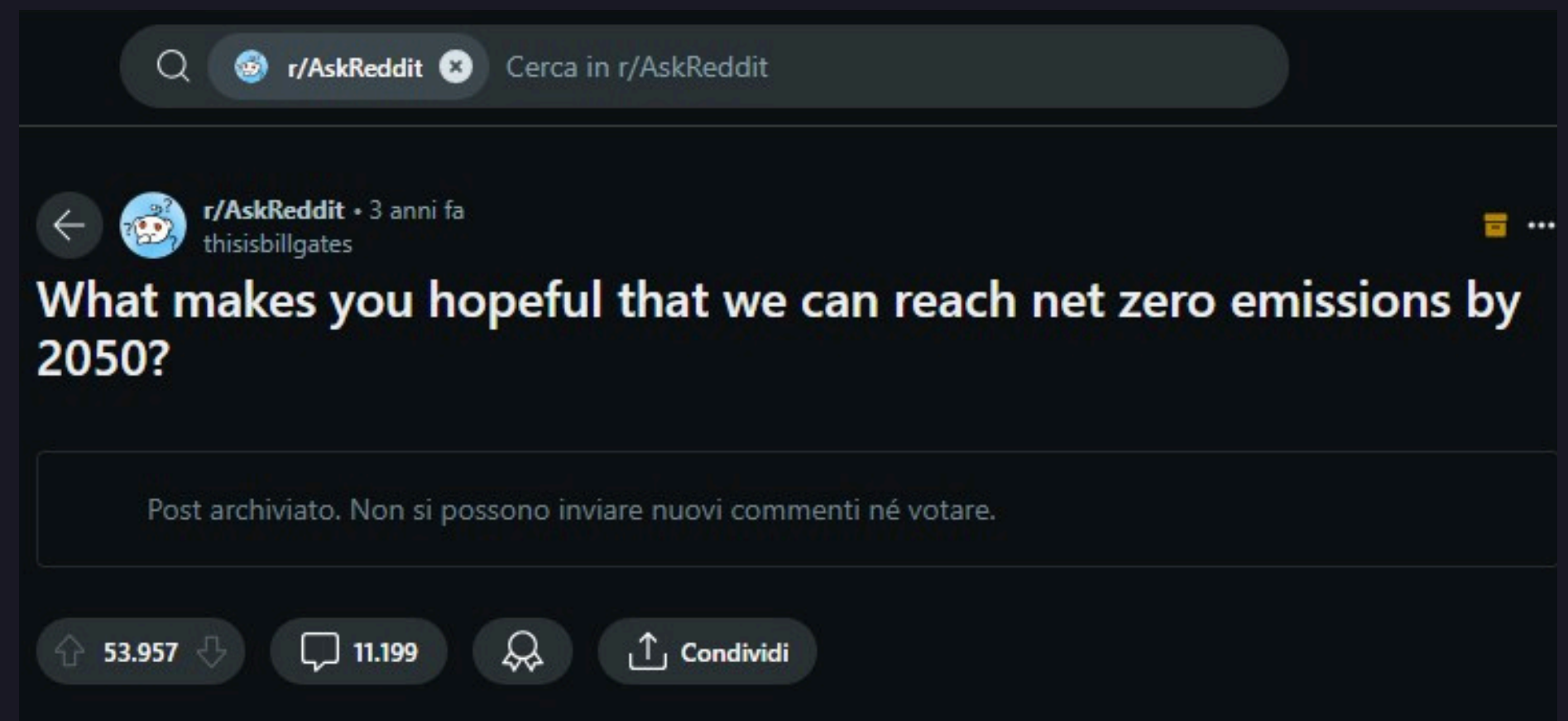
05

- 1** **SOURCE: È LO SPAZIO DEDICATO ALLA SCRITTURA DEL CODICE**
- 2** **R CONSOLE: MOSTRA I RISULTATI DEI COMANDI LANCIATI NEL SOURCE E PERMETTE DI INTERAGIRE CON R TRAMITE IL TERMINAL**
- 3** **ENVIRONMENT/HISTORY: CONTIENE TUTTI I FILE GENERATI DALLE ANALISI, OLTRE A UNA CRONOLOGIA DELLE AZIONI SVOLTE E A UN BREVE TUTORIAL**
- 4** **FILES/PLOTS/PACKAGES/HELP/VIEWER: RAPPRESENTA LA WORKING DIRECTORY SU CUI SI STA LAVORANDO E OSPITA I GRAFICI GENERATI**

# Il primo thread

*What makes you hopeful that we can reach **net zero emissions** by 2050?*

Il primo thread è stato pubblicato a marzo del 2021 e ha riscosso grande successo, raccogliendo oltre 50mila voti. Al momento dell'analisi, il thread risulta archiviato.

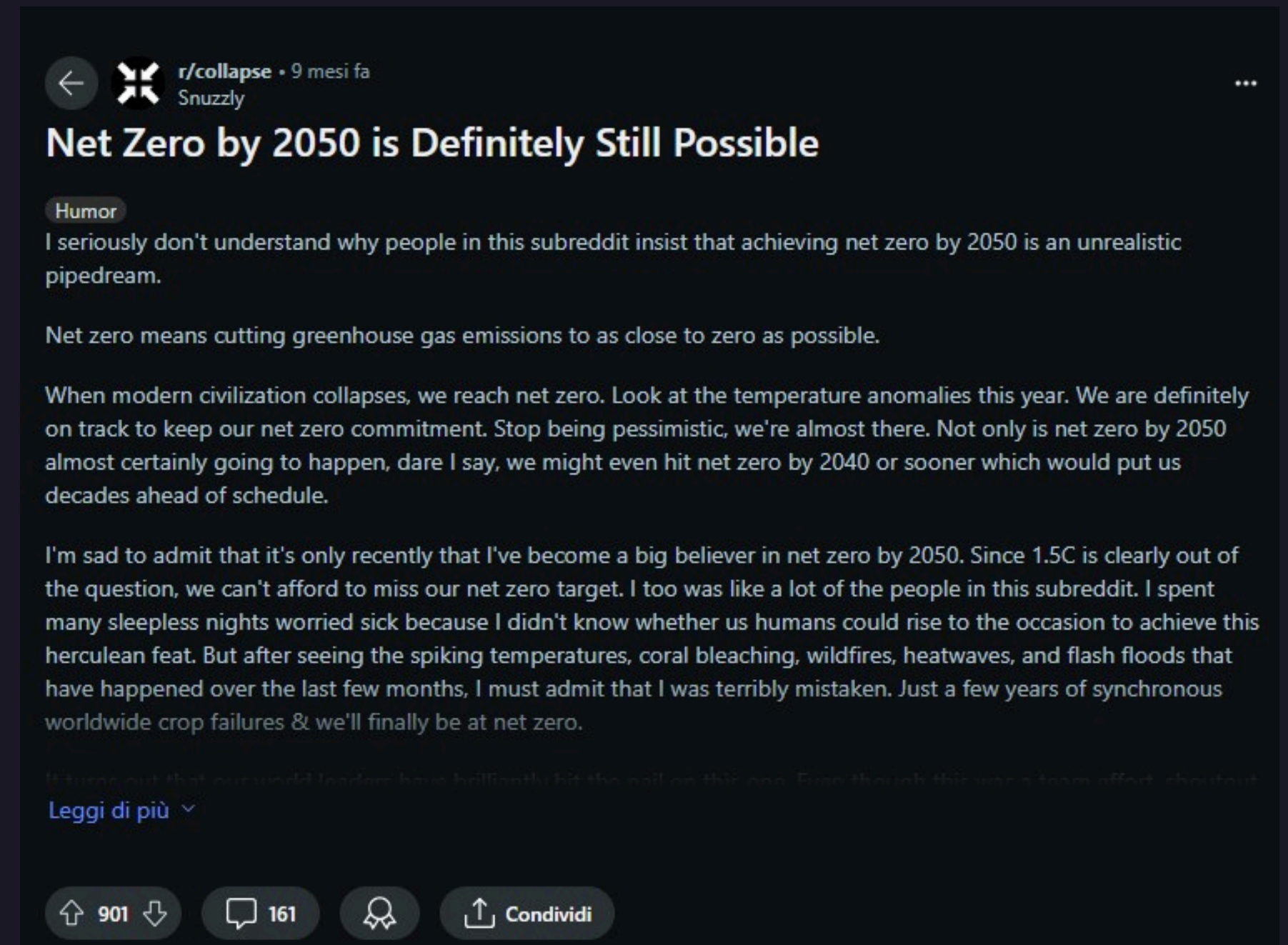




# Il secondo thread

*Net Zero by 2050 is Definitely Still Possible*

Il secondo thread è più recente: pubblicato ad agosto 2023, è ancora aperto al pubblico e ha ricevuto, al momento del data retrieving, 161 commenti.



# I due thread a confronto

## THREAD #1

- Domanda aperta
- Genera una discussione superficiale
- Contiene più commenti
- Ha raggiunto più persone
- Risale a più di tre anni fa

## THREAD #2

- Affermazione ironica e divisiva
- Genera una discussione profonda
- Contiene meno commenti
- Ha raggiunto meno persone
- Risale a meno di un anno fa



# L'analisi dei dati



Dopo aver presentato in breve le caratteristiche dei due thread, vediamo tutte le fasi dell'analisi, che includono:

- DATA RETRIEVING
- PULIZIA DEI COMMENTI
- VISUALIZZAZIONE DEI RISULTATI

# Operazioni preliminari: i pacchetti

10

Prima di svolgere qualsiasi operazione è necessario scaricare e lanciare tutti i pacchetti, oltre a verificarne il corretto funzionamento.

```
install.packages("RedditExtractor")  
install.packages("udpipe")
```

L'immagine d'esempio mostra il download di due librerie fondamentali:

- **RedditExtractor**, con cui potremo scaricare i commenti dai link reddit
- **UDPipe**, con cui misureremo le frequenze nella parte finale dell'analisi

# Data retrieving

A conclusione delle operazioni preliminari si è proceduto con l'estrazione dei commenti nella fase di **data retrieving**.

```
thread2021 <- get_thread_content("https://www.reddit.com/r/AskReddit/  
                                comments/m8o6iq/what_makes_you_hopeful  
                                _that_we_can_reach_net_zero/")  
thread2023 <- get_thread_content("https://www.reddit.com/r/collapse/  
                                comments/15uo2dx/net_zero_by_2050_is_  
                                definitely_still_possible/")  
  
commenti2021<-thread2021$comments  
commenti2023<-thread2023$comments
```

I due thread sono stati scaricati con il comando ***get\_thread\_comment***, che è intervenuto sui rispettivi link popolando altrettante variabili.

# Creazione dei data frame

I commenti individuati dal **RedditExtractor** sono stati inseriti in due data frame.

## THREAD #1

Nome: commenti2021  
Commenti trovati: 494

## THREAD #2

Nome: commenti2023  
Commenti trovati: 162

Una possibile implementazione, che non è stata eseguita in questa analisi, è la scrittura di due file **xlsx** per scaricare i data frame con il pacchetto **writexl**.



# Pulizia dei commenti

Nella fase di pulizia sono stati utilizzati i pacchetti **tm** e **dplyr** per eliminare le stopwords.

```
library(tm)
library(dplyr)

stopwords_en <- stopwords("en")

commenti2021$comment <- sapply(commenti2021$comment, function(x) {
  x <- tolower(x)
  x <- removewords(x, stopwords_en)
})
```

Pulendo i commenti prima di tokenizzarli, escluderemo tutti gli elementi che non sono utili per quest'analisi.



# Tokenizzazione dei commenti

I commenti sono stati tokenizzati: i risultati ottenuti sono stati poi aggiunti alla tabella con il comando `colnames`, inserendo i tipi nella **colonna 11** e i token nella **colonna 12**.

```
commenti2021<-cbind(commenti2021, ntype(tokenCommenti2021),  
                    ntoken(tokenCommenti2021))%>%  
  mutate(richness=ntype  
          (tokenCommenti2021)/ntoken(tokenCommenti2021))  
colnames(commenti2021)[11] ="tipo"  
colnames(commenti2021)[12] ="token"
```

# L'analisi

I dati ricavati, puliti e tokenizzati sono stati utilizzati per eseguire tre analisi:

- INDIVIDUAZIONE DEI BIGRAMMI
- CALCOLO DELLE FREQUENZE
- OCCORRENZE DELLE PARTS OF SPEECH



# Individuazione dei bigrammi

Si tratta di un'analisi basilare che isola tutte le sequenze di due parole di senso compiuto. Da questa prima analisi dovrebbero emergere bigrammi legati al contesto della **neutralità climatica**.

```
bigrammi2021<-tokenCommenti2021 %>% #INDIVIDUAZIONE BIGRAMMI
tokens_remove(stopwords("en")) %>%
tokens_select(pattern = "[a-z]", valuetype = "regex",
               case_insensitive = FALSE, padding = TRUE) %>%
textstat_collocations(min_count = 5, size=2)

dfbigrammi2021<-bigrammi2021 %>%
select(Bigramma = collocation, count)

dfbigrammi2021$Bigramma <- factor(dfbigrammi2021$Bigramma,
                                levels = dfbigrammi2021$Bigramma
                                [order(dfbigrammi2021$count,
                                       decreasing = TRUE)])
```

- Individuazione bigrammi
- Creazione data frame con due colonne: "Bigramma", che contiene i bigrammi, e "Count", che contiene le occorrenze
- Ordinamento decrescente dei bigrammi per numero di occorrenze

# Individuazione dei bigrammi

Nella rappresentazione grafica sono stati specificati:

- i due assi da tenere in considerazione (`aes x e y`)
- le caratteristiche estetiche del grafico (`stat, color, fill`)
- la direzione dei label in verticale (`axis.text.x`)

```
ggplot(dfbigrammi2021, aes(x = Bigramma, y = count)) +  
  geom_bar(stat = "identity", color = "black", fill="darkorange") +  
  theme(axis.text.x = element_text(hjust = 1, angle = 90)) +  
  order() +  
  labs(title = "Bigrammi - 2021",  
        x = "Bigramma",  
        y = "Occorrenze")
```

# Calcolo delle frequenze

Il calcolo delle frequenze è stato eseguito anche sulle singole parole grazie a una matrice dei termini nei commenti.

```
corpus2021<-VCorpus(VectorSource(tokenCommenti2021))  
  
dataset2021<-tm_map(corpus2021, content_transformer(tolower))  
dataset2021<-tm_map(dataset2021, content_transformer(removeWords),  
                     stopwords("english"))  
  
matrice_dataset2021<-TermDocumentMatrix(dataset2021)  
matrice_dataset2021<-as.matrix(matrice_dataset2021)  
  
somma_dataset2021<-sort(rowSums(matrice_dataset2021),  
                        decreasing = TRUE)  
freqDataset2021<-data.frame(parola = names(somma_dataset2021),  
                             freq=somma_dataset2021)  
  
head(freqDataset2021, 10)  
  
top20Freq2021<-freqDataset2021[1:20,]
```

- Creazione di un corpus con tutti i token dei commenti
- Creazione di una matrice dei token con le rispettive occorrenze
- Creazione del dataset con le frequenze dei token rintracciati
- Ordinamento dei primi 20 token



# Calcolo delle frequenze

Una volta individuate le frequenze, i 20 termini più frequenti sono stati selezionati, caricati in un data frame, ordinati in ordine decrescente e inseriti in un grafico.

```
top20Freq2021$parola <- factor(top20Freq2021$parola,  
                               levels = top20Freq2021$parola  
                               [order(top20Freq2021$freq,  
                                       decreasing = TRUE)])  
  
ggplot(top20Freq2021, aes(x = parola, y = freq)) +  
  geom_bar(stat = "identity", color = "black", fill = "darkorange") +  
  theme(axis.text.x = element_text(hjust = 1, angle = 90)) +  
  order() +  
  labs(title = "Parole più frequenti - 2021",  
        x = "Parola",  
        y = "Frequenza")
```

# Occorrenze delle Parts of Speech

L'ultima analisi riguardava l'individuazione delle Parts of Speech (PoS) e la creazione di un grafico con le frequenze.

Il pacchetto **udpipe** contiene un modello per annotare le PoS del data frame e generare un grafico a barre.

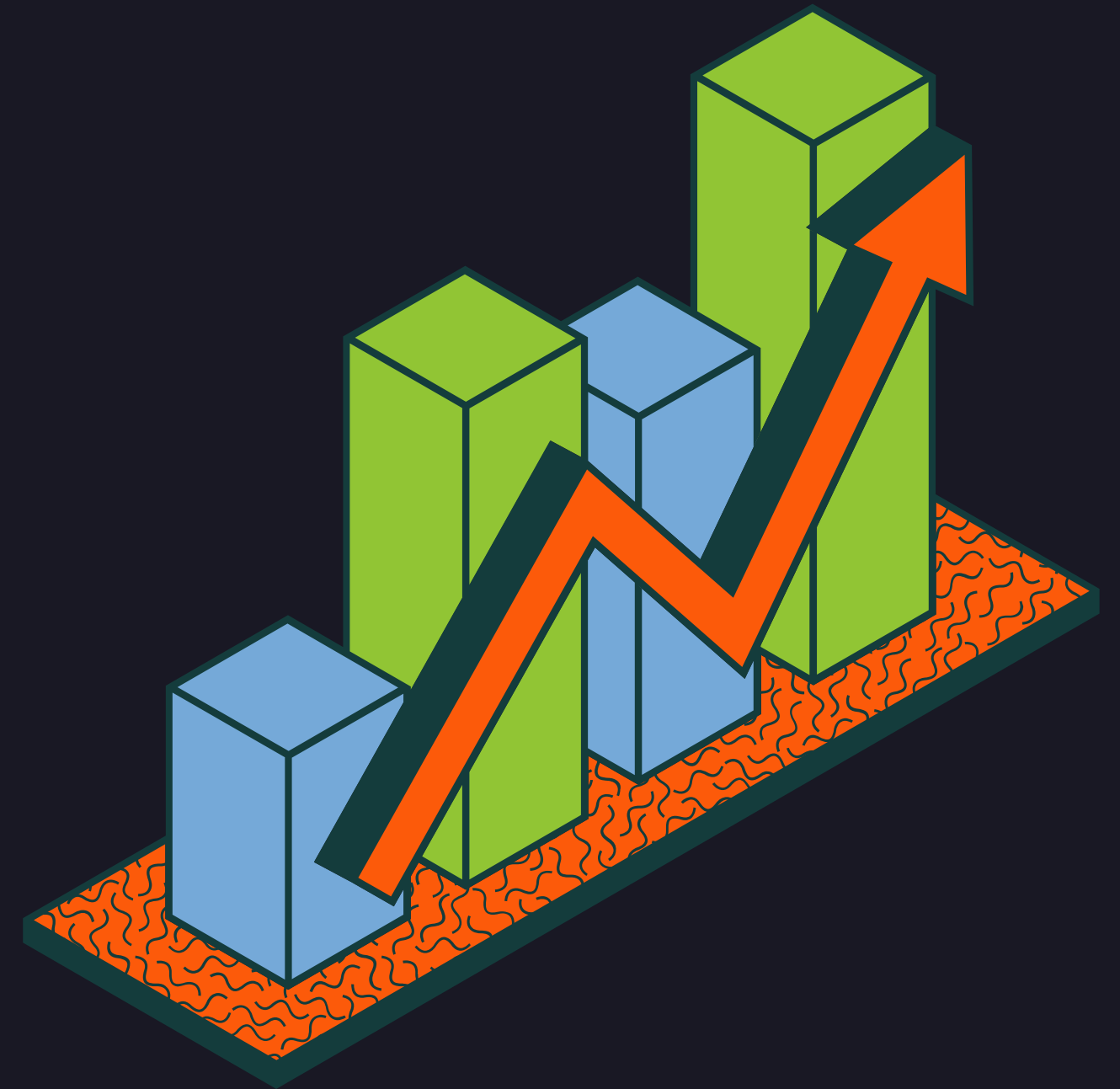
```
ud_model<-udpipe_download_model(language = "english")
ud_model<-udpipe_load_model(ud_model$file_model)
Pos2021<-udpipe_annotate(ud_model, x = commenti2021$comment,
                          doc_id = commenti2021$comment_id)
Pos2021<-as.data.frame(Pos2021)

statPos2021 <- txt_freq(Pos2021$upos)
statPos2021$key <- factor(statPos2021$key, levels = rev(statPos2021$key))
barchart(key ~ freq, data = statPos2021, col = "darkorange",
          main = "Parts of speech - 2021",
          xlab = "Frequenza")
```

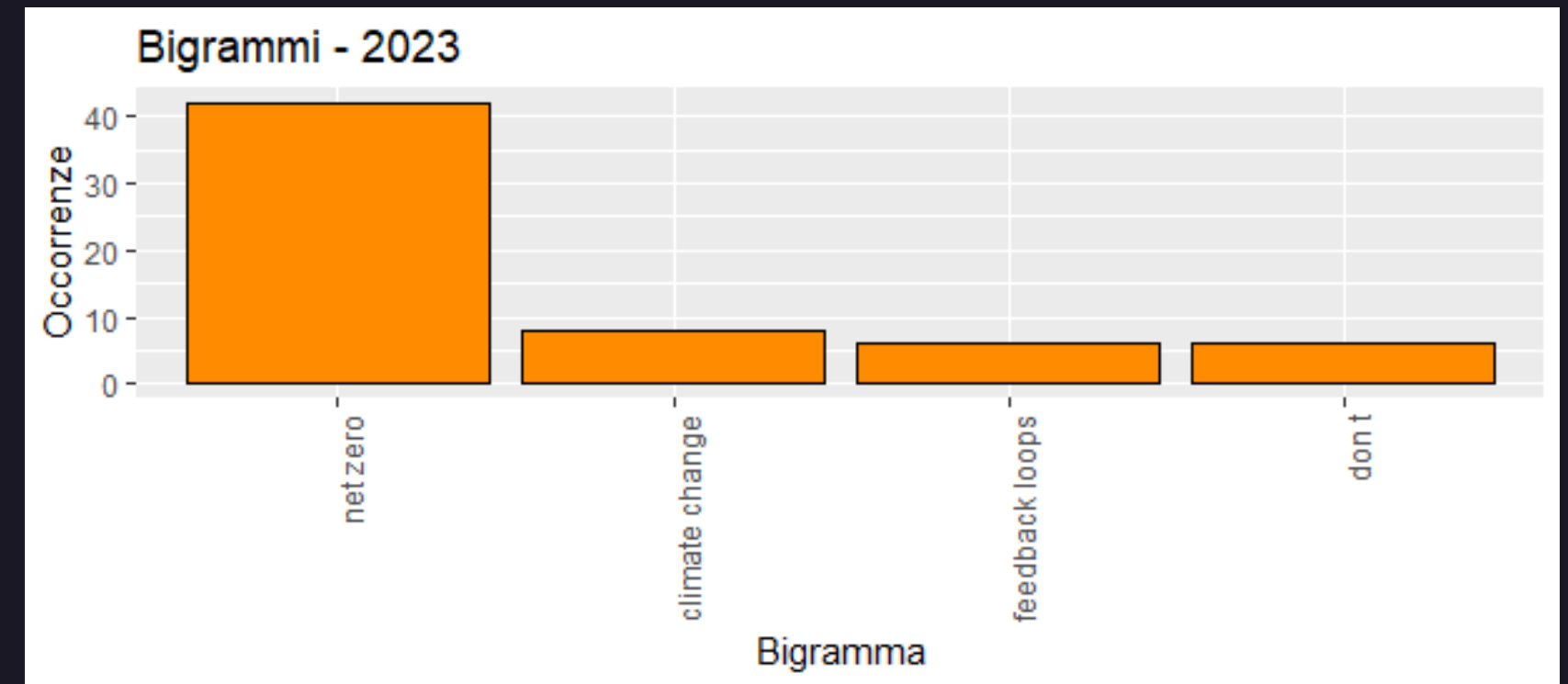
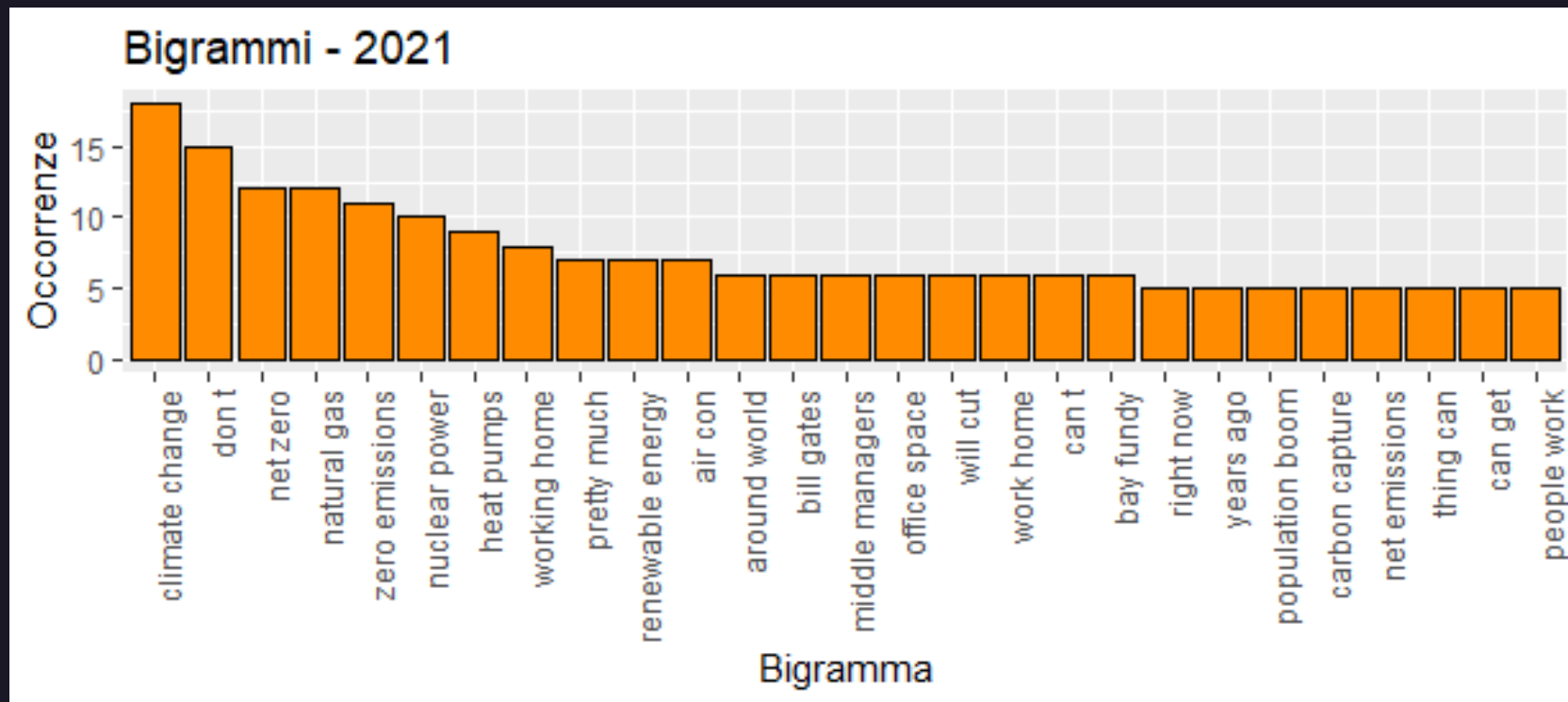
# La visualizzazione dei dati

Al termine dell'analisi sono stati generati tre grafici di frequenza per ogni dataset, per un totale di sei grafici:

- FREQUENZA DEI BIGRAMMI
- FREQUENZA DELLE PAROLE
- FREQUENZA DELLE PARTS OF SPEECH



# Individuazione dei bigrammi



**Bigrammi trovati: 27**

**Primi bigrammi su neutralità climatica:**

**Climate change - 18 occorrenze**

**Net zero - 12 occorrenze**

**Natural gas - 12 occorrenze**

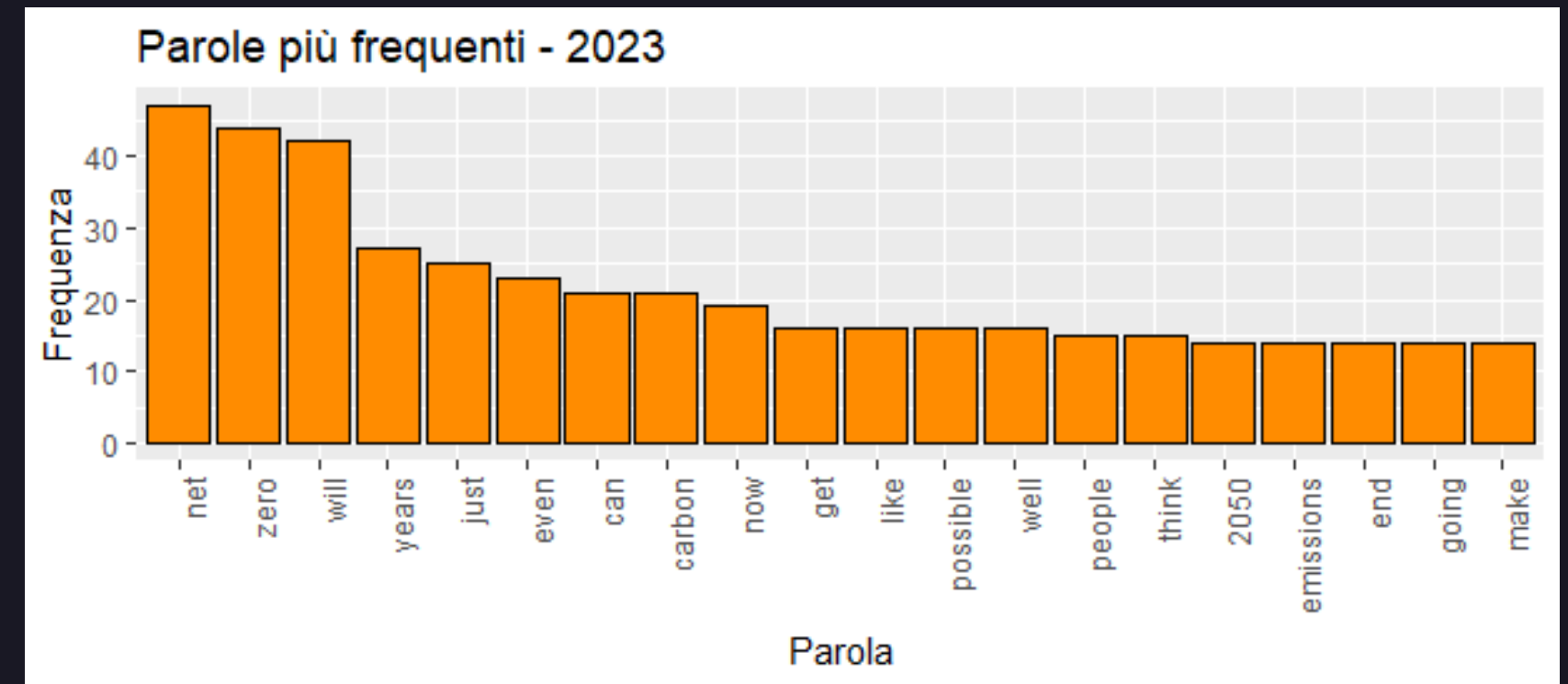
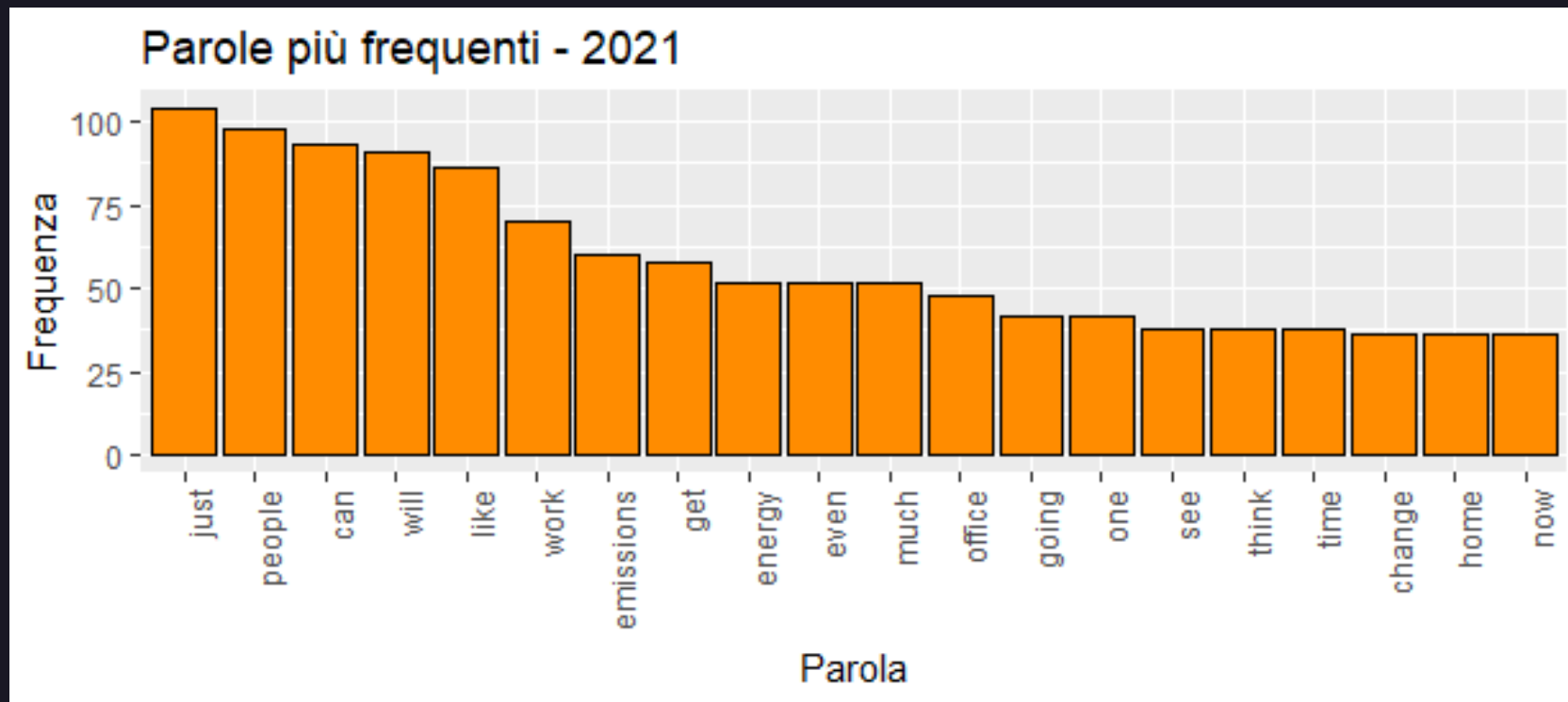
**Bigrammi trovati: 4**

**Primi bigrammi su neutralità climatica:**

**Net zero - 42 occorrenze**

**Climate change - 8 occorrenze**

# Calcolo delle frequenze



Parole trovate: 3267

Prime parole su neutralità climatica:

Emissions - 60 occorrenze

Energy - 52 occorrenze

Change - 36 occorrenze

Parole trovate: 1585

Prime parole su neutralità climatica:

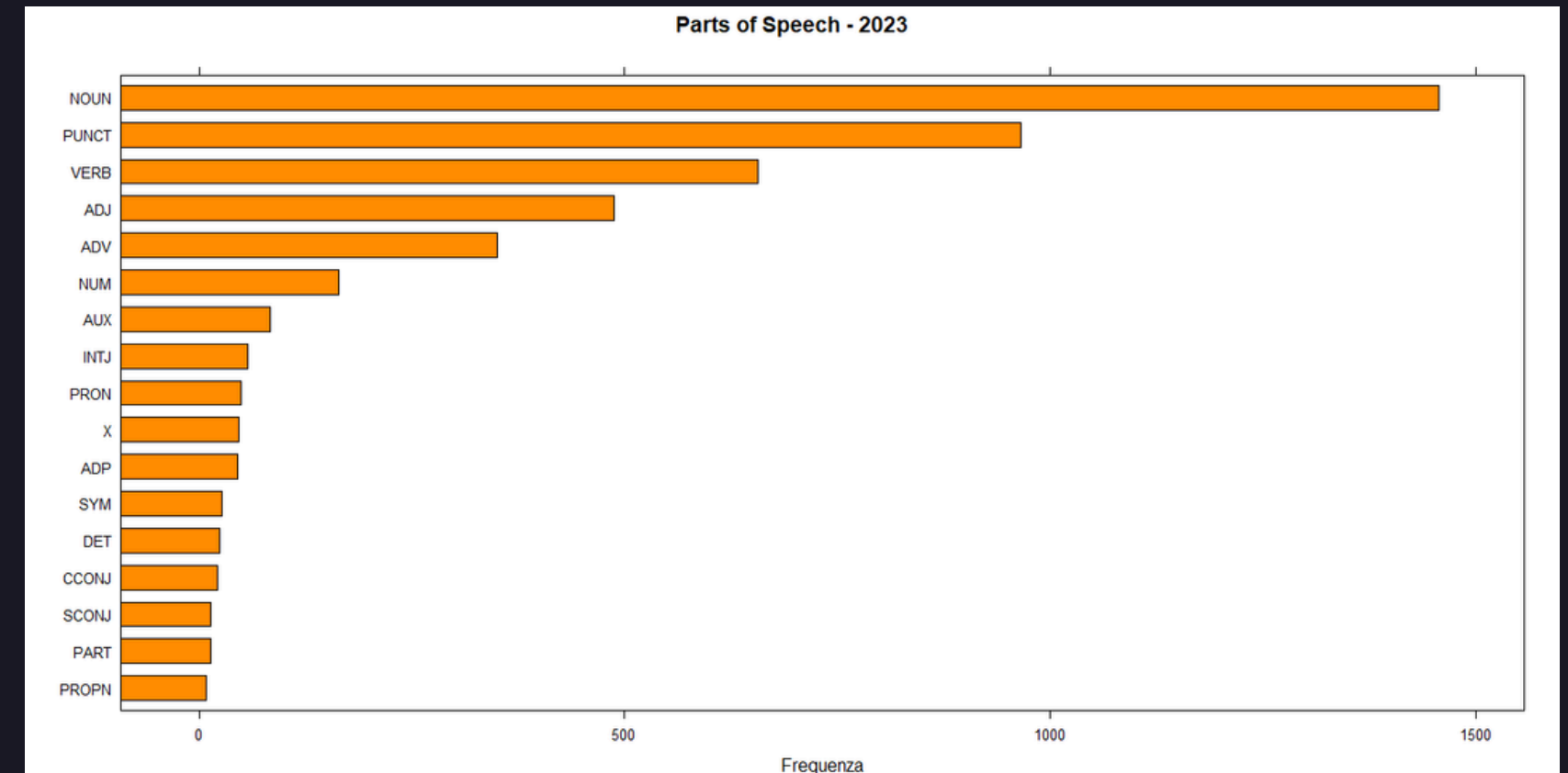
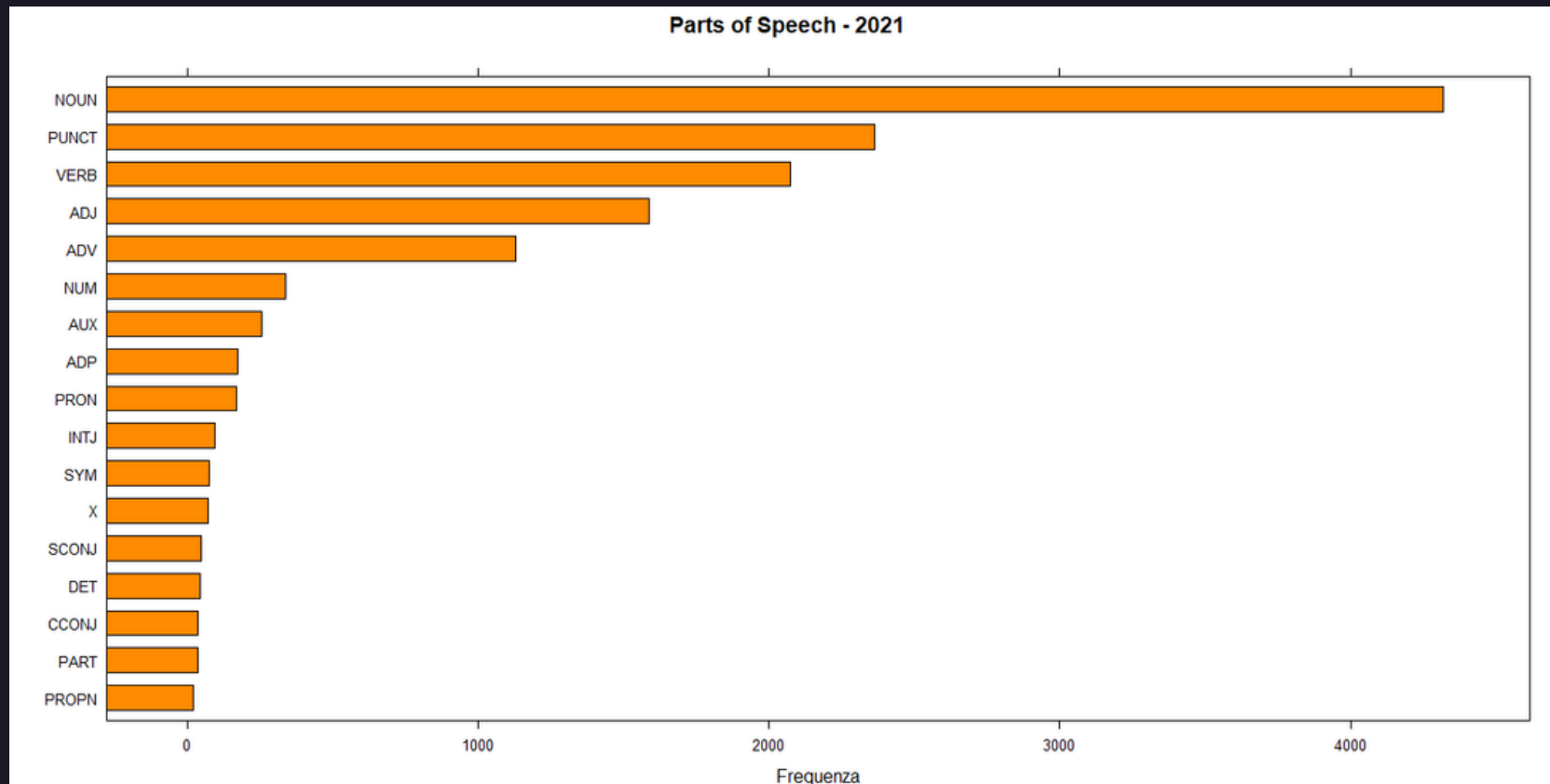
Net - 47 occorrenze

Zero - 44 occorrenze

Carbon - 21 occorrenze



# Occorrenze delle Parts of Speech



Parts of Speech più utilizzate\*:

Nomi - 4318

Verbi - 2074

Aggettivi - 1587

Parts of Speech più utilizzate\*:

Nomi - 1456

Verbi - 657

Aggettivi - 488

\* = sono stati esclusi i segni di punteggiatura (PUNCT)

# Le conclusioni dell'analisi

Uso crescente del termine **net zero**

THREAD #1

Occorrenze **net zero**: 12  
2° bigramma più usato

THREAD #2

Occorrenze **net zero**: 42  
1° bigramma più usato

# Le conclusioni dell'analisi

## Equilibrio nell'uso delle Parts of Speech

### THREAD #1

NOUN (Nomi): 33.7%  
VERB (Verbi): 16.1%  
ADJ (Aggettivi): 12.3%

### THREAD #2

NOUN (Nomi): 32.5%  
VERB (Verbi): 14.6%  
ADJ (Aggettivi): 10.9%

Sono stati esclusi i segni di punteggiatura (PUNCT)

# Possibili sviluppi dell'analisi

L'analisi rappresenta un punto di partenza per altri approfondimenti che indaghino l'atteggiamento globale verso la **neutralità climatica**. Le implementazioni includono:

- 1 L'utilizzo di più fonti da confrontare con **Reddit** (es. **YouTube**)
- 2 Un confronto su periodi **più lunghi** e con dataset **più ricchi**
- 3 Una sentiment analysis con le librerie **udpipe**

