WORD COUNT/TOKENIZER

PROYECTO FINAL

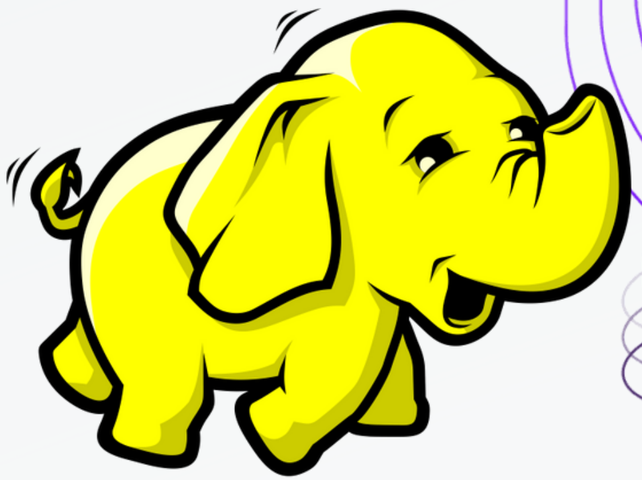
DANIEL ISAAC JUAREZ FUNES - 12141153
DIEGO ANDRÉ MOLINA VALLADARES - 12141157
SERLIO ALEJANDRO GIRÓN PAZ - 12141146

ÍNDICE

INTRODUCCIÓN

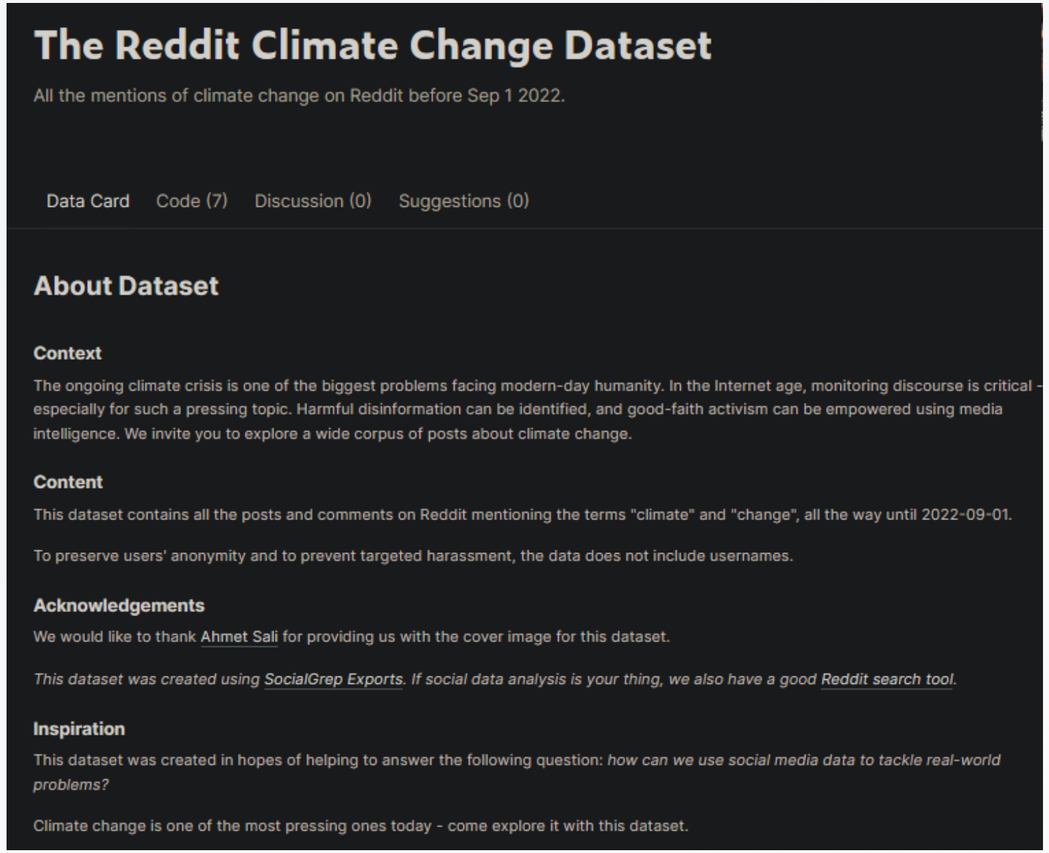
- El proyecto se enfoca en la aplicación de técnicas de Map/Reduce en el entorno Hadoop para en analisis de frecuencias de conjuntos de palabras

- Se desarrollan tres aplicaciones fundamentales. Un preprocesamiento, encargado de eliminar palabras y/o signos no deseados. Un WordCount, encargado de contabilizar la frecuencia de las palabras en un archivo de texto, y Frecuency Analysis, destinada a identificar el top 20 de los conjuntos mas frecuentes.



DATASET

Nuestro dataset contiene todas las publicaciones y comentarios en Reddit que mencionan el Cambio Climatico hasta el mes de Septiembre de 2022. Al Reddit ser una plataforma libre, nos encontramos con opiniones mixtas de diversos usuarios de esta amplia red social.



The Reddit Climate Change Dataset

All the mentions of climate change on Reddit before Sep 1 2022.

Data Card | Code (7) | Discussion (0) | Suggestions (0)

About Dataset

Context

The ongoing climate crisis is one of the biggest problems facing modern-day humanity. In the Internet age, monitoring discourse is critical - especially for such a pressing topic. Harmful disinformation can be identified, and good-faith activism can be empowered using media intelligence. We invite you to explore a wide corpus of posts about climate change.

Content

This dataset contains all the posts and comments on Reddit mentioning the terms "climate" and "change", all the way until 2022-09-01.

To preserve users' anonymity and to prevent targeted harassment, the data does not include usernames.

Acknowledgements

We would like to thank Ahmet Sali for providing us with the cover image for this dataset.

This dataset was created using SocialGrep Exports. If social data analysis is your thing, we also have a good Reddit search tool.

Inspiration

This dataset was created in hopes of helping to answer the following question: *how can we use social media data to tackle real-world problems?*

Climate change is one of the most pressing ones today - come explore it with this dataset.

DATASET ANTES DEL PROCESAMIENTO

the-reddit-climate-change-dataset-comments.csv (4.11 GB)

Detail | Compact | Column

10 of 10 columns

type Type of the datapoint	id Unique Base-36 ID of the comment	subreddit.id Unique Base-36 ID of the comment's subreddit	subreddit.name Human-readable name of the comment's subreddit	subreddit.nsfw Is the comment's subreddit NSFW?	created_utc Timestamp of the comment's creation	permalink Permalink to the comment on Reddit	body Comment's body text	sentiment Analyzed sentiment for the comment	score Comment's score
1 unique value	4600698 total values	4600698 total values	politics 8% worldnews 8% Other (3879485) 84%	true 15.5k 0% false 4.59m 100%	1.26b … 1.66b	4600698 unique values	4485881 unique values	-1 … 1	-2379 … 36.4k
comment	imlddn9	2qh3l	news	false	1661990368	https://old.reddit.com/r/news/comments/x2cszk/us_life_expectancy_down_for_second straight_year/imlddn...	Yeah but what the above commenter is saying is their base doesn't want any of that. They detest all ...	0.5719	2
comment	imldbeh	2qn7b	ohio	false	1661990340	https://old.reddit.com/r/Ohio/comments/x2awnp/state_government_may_soon_kill_a_solar_project_in/imld...	Any comparison of efficiency between solar and fossil fuels is nonsensical at best and intentionally...	-0.9877	2
comment	imldado	2qhma	newzealand	false	1661990327	https://old.reddit.com/r/newzealand/comments/x28xci/long_rant_pessimistic_asf_and_feel_like_were/iml...	I'm honestly waiting for climate change and the impacts of that to kick some fucking sense into peop...	-0.1143	1
comment	imld6cb	2qi09	sacramento	false	1661990278	https://old.reddit.com/r/Sacramento/comments/x2ruqy/hey_guyz_this_is_a_tough_one_why_do_you_think/im...	Not just Sacramento. It's actually happening all over the world. Climate change is real, believe it ...	0.0	4
comment	imld0kj	2qh1i	askreddit	false	1661990206	https://old.reddit.com/r/AskReddit/comments/x2fj3g/whats_a_controversial_topic_no_one_wants_to/imld0...	I think climate change tends to get some people riled up. When I was part of a debate club, they l...	0.6634	1
comment	imlctri	3l2gt	walkaway	false	1661990120	https://old.reddit.com/r/walkaway/comments/x2mw7x/turdeau_avoiding_accountability_as_usual/imlctri/	Naaa how could anyone be mad at a face like that... Must definitely be climate change	0.25	1
comment	imlctc0	2vvve	pastors	false	1661990114	https://old.reddit.com/r/pastors/comments/x2ilpr/lets_talk_about_it_class/imlctc0/	Can i suggest maybe honing in on LGBTQ? It's a useful grab-bag for talking about issues like law re...	0.9779	2
comment	imlcpab	2qh1i	askreddit	false	1661990065	https://old.reddit.com/r/AskReddit/comments/x2fj3g/whats_a_controversial_topic_no_one_wants_to/imlcp	They need to change laws so it's more worth selling agriculture products in the US rather than expor...	0.469	2

DATASET ANTES DEL PROCESAMIENTO

```
post,x2bwsy,2u489,truechristian,false,1661947599,https://old.reddit.com/r/TrueChristian/comments/x2bwsy/is_it_wrong_for_a_christian_to_become_an_airline/,self.truechristian,,[removed],Is it wrong for a Christian to become an airline pilot considering climate change?,0
post,x2btio,3hapm,freekarma4u,false,1661947331,https://old.reddit.com/r/FreeKarma4U/comments/x2btio/greedy_asshat_doesnt_understand_climate_change/,i.imgur.com,https://i.imgur.com/WZ0aCi6.png,,Greedy asshat doesn't understand climate change.,1
post,x2bl9l,3h9d4,autonewspaper,false,1661946668,https://old.reddit.com/r/AutoNewspaper/comments/x2bl9l/oped_tom_moyer_and_lauren_barros_utah_republicans/,sltrib.com,https://www.sltrib.com/opinion/commentary/2022/08/31/tom-moyer-lauren-barros-utah/,,[Op-Ed] - Tom Moyer a
post,x2bcuf,2tk0s,unpopularopinion,false,1661945953,https://old.reddit.com/r/unpopularopinion/comments/x2bcuf/bottling_water_is_a_huge_problem_when_it_comes_to/,self.unpopularopinion,,[removed],Bottling water is a huge problem when it comes to climate change.,2
post,x2bby3,3hssx,sltribauto,false,1661945873,https://old.reddit.com/r/SLTRIBauto/comments/x2bby3/oped_tom_moyer_and_lauren_barros_utah_republicans/,sltrib.com,https://www.sltrib.com/opinion/commentary/2022/08/31/tom-moyer-lauren-barros-utah/,,[Op-Ed] - Tom Moyer and Lau
post,x2am7a,2qh1n,environment,false,1661943615,https://old.reddit.com/r/environment/comments/x2am7a/climate_change_pakistan_emits_less_than_1_of_the/,cnn.com,https://www.cnn.com/2022/08/30/asia/pakistan-climate-crisis-floods-justice-intl/index.html,,Climate change: Pakis
post,x2ak7f,3h9d4,autonewspaper,false,1661943443,https://old.reddit.com/r/AutoNewspaper/comments/x2ak7f/oped_evs_are_merely_a_life_hack_that_wont_save_us/,thestar.com,https://www.thestar.com/opinion/contributors/2022/08/31/evs-are-merely-a-life-hack-that-wont-save-us-fro
post,x2aevq,2qpcz,backpacking,false,1661942925,https://old.reddit.com/r/backpacking/comments/x2aevq/heat_water_fire_how_climate_change_is/,nytimes.com,https://www.nytimes.com/2022/08/31/travel/climate-change-pacific-crest-trail.html?smid=nytcore-ios-share&amp;referringSo
post,x2abnw,3hbd2,torontostarauto,false,1661942612,https://old.reddit.com/r/TORONTOSTARauto/comments/x2abnw/oped_evs_are_merely_a_life_hack_that_wont_save_us/,thestar.com,https://www.thestar.com/opinion/contributors/2022/08/31/evs-are-merely-a-life-hack-that-wont-save-us
post,x2a7xo,6wzx9b,eduwriters,false,1661942260,https://old.reddit.com/r/EduWriters/comments/x2a7xo/what_are_the_effects_of_climate_change_essay/,eduwriters.pro,https://eduwriters.pro/?cid=2860/?utm_source=rdwrt&amp;utm_medium=pst&amp;utm_campaign=what_are_the_effects_of_
post,x29tpo,3l9pr,maxcactus_trailguide,false,1661940854,https://old.reddit.com/r/Maxcactus_TrailGuide/comments/x29tpo/heat_water_fire_how_climate_change_is/,nytimes.com,https://www.nytimes.com/2022/08/31/travel/climate-change-pacific-crest-trail.html,,"Heat, Water, Fire:
post,x29np7,3hv7k4,trendingquicktvnews,false,1661940264,https://old.reddit.com/r/TrendingQuickTVnews/comments/x29np7/heat_water_fire_how_climate_change_is/,nytimes.com,https://www.nytimes.com/2022/08/31/travel/climate-change-pacific-crest-trail.html,,"Heat, Water, Fire:
post,x29jkn,9wby,nytimes,false,1661939873,https://old.reddit.com/r/nytimes/comments/x29jkn/heat_water_fire_how_climate_change_is/,nytimes.com,https://www.nytimes.com/2022/08/31/travel/climate-change-pacific-crest-trail.html?partner=IFTTT,,"Heat, Water, Fire: How Climate
post,x29dc1,2t5w9,newry,false,1661939205,https://old.reddit.com/r/newry/comments/x29dc1/people_are_now_conscious_of_the_reality_of/,twitter.com,https://twitter.com/NewryDemo/status/1564899862218522624,,'People are now conscious of the reality of climate change' https://t
post,x298fk,2qh1i,askreddit,false,1661938687,https://old.reddit.com/r/AskReddit/comments/x298fk/what_is_one_thing_that_makes_you_think_yes_we_can/,self.askreddit,https://www.reddit.com/r/AskReddit/comments/x298fk/what_is_one_thing_that_makes_you_think_yes_we_can/,,"What
post,x297is,2qh1i,askreddit,false,1661938593,https://old.reddit.com/r/AskReddit/comments/x297is/is_there_still_one_thing_that_makes_you_think_yes/,self.askreddit,,[removed],"Is there still one thing that makes you think “yes, we can fight climate change”?",1
post,x28w9s,5r8j6u,phnewsfeed,false,1661937409,https://old.reddit.com/r/phnewsfeed/comments/x28w9s/indonesia_calls_for_more_g20_action_on_climate/,mb.com.ph,https://mb.com.ph/2022/08/31/indonesia-calls-for-more-g20-action-on-climate-change/?utm_source=rss&amp;utm_medium=
post,x28rbl,3h9d4,autonewspaper,false,1661936879,https://old.reddit.com/r/AutoNewspaper/comments/x28rbl/travel_heat_water_fire_how_climate_change_is/,nytimes.com,https://www.nytimes.com/2022/08/31/travel/climate-change-pacific-crest-trail.html,,"[Travel] - Heat, Water, F
post,x28q5g,3h8h4,nytauto,false,1661936755,https://old.reddit.com/r/NYTauto/comments/x28q5g/travel_heat_water_fire_how_climate_change_is/,nytimes.com,https://www.nytimes.com/2022/08/31/travel/climate-change-pacific-crest-trail.html,,"[Travel] - Heat, Water, Fire: How Cli
post,x2815m,6wzx9b,eduwriters,false,1661936252,https://old.reddit.com/r/EduWriters/comments/x2815m/how_to_address_climate_change_essay/,eduwriters.pro,https://eduwriters.pro/?cid=2860/?utm_source=rdwrt&amp;utm_medium=pst&amp;utm_campaign=how_to_address_climate_change_ess
post,x28fy2,2tlqg,samplesize,false,1661935671,https://old.reddit.com/r/SampleSize/comments/x28fy2/climate_change_and_global_warming_survey_for/,self.samplesize,,"Hello, I am conducting a survey so I can study how different ages use different language devices to describe

Here it is: [https://forms.gle/6fwALrZGpkYwES1i8](https://forms.gle/6fwALrZGpkYwES1i8)",Climate change and global warming survey for English Language coursework. (Everyone),3
post,x28efu,2tlqg,samplesize,false,1661935508,https://old.reddit.com/r/SampleSize/comments/x28efu/climate_change_and_global_warming_survey_for/,self.samplesize,,[deleted],Climate change and global warming survey for English Language coursework. (Academic),1
post,x28eby,3h9d4,autonewspaper,false,1661935497,https://old.reddit.com/r/AutoNewspaper/comments/x28eby/world_indonesia_calls_for_more_g20_action_on/,mb.com.ph,https://mb.com.ph/2022/08/31/indonesia-calls-for-more-g20-action-on-climate-change/?utm_source=rss&amp;utm_medi
post,x28byg,2tlqg,samplesize,false,1661935232,https://old.reddit.com/r/SampleSize/comments/x28byg/academic_climate_change_and_global_warming_survey/,self.samplesize,,[removed],[**Academic**] Climate change and global warming survey for English Language coursework.,1
post,x28a35,2qh1n,environment,false,1661935024,https://old.reddit.com/r/environment/comments/x28a35/bible_demands_action_on_climate_change/,washingtonpost.com,https://www.washingtonpost.com/religion/2022/08/30/evangelicals-climate-change-bible/,,"Bible demands action on
post,x289h8,6yqjag,soundslikeajoke,false,1661934954,https://old.reddit.com/r/soundslikeajoke/comments/x289h8/climate_change_is_evidence_that_earth_is_a_woman/,self.soundslikeajoke,https://www.reddit.com/r/soundslikeajoke/comments/x289h8/climate_change_is_evidence_that_ea
post,x288dr,2tlqg,samplesize,false,1661934828,https://old.reddit.com/r/SampleSize/comments/x288dr/climate_change_and_global_warming_survey_for/,self.samplesize,,[removed],Climate change and global warming survey for English Language coursework.,1
post,x280gz,3jaf8,mbauto,false,1661933956,https://old.reddit.com/r/MBauto/comments/x280gz/world_indonesia_calls_for_more_g20_action_on/,mb.com.ph,https://mb.com.ph/2022/08/31/indonesia-calls-for-more-g20-action-on-climate-change/?utm_source=rss&amp;utm_medium=rss&amp;utm
post,x27x2z,6wzx9b,eduwriters,false,1661933573,https://old.reddit.com/r/EduWriters/comments/x27x2z/how_can_we_prevent_climate_change_essay/,eduwriters.pro,https://eduwriters.pro/?cid=2860/?utm_source=rdwrt&amp;utm_medium=pst&amp;utm_campaign=how_can_we_prevent_climate_ch
post,x27uiv,3h9d4,autonewspaper,false,1661933297,https://old.reddit.com/r/AutoNewspaper/comments/x27uiv/business_climate_change_alibaba_boosts/,scmp.com,https://www.scmp.com/business/article/3190822/climate-change-alibaba-boosts-clean-energy-purchases-become-largest-buye
post,x27tjk,6q3cs1,sciencenopolitics,false,1661933183,https://old.reddit.com/r/ScienceNoPolitics/comments/x27tjk/exposure_to_past_temperature_variability_may_help/,eurekalert.org,https://www.eurekalert.org/news-releases/963009,,Exposure to past temperature variability ma
post,x27szs,mouw,science,false,1661933122,https://old.reddit.com/r/science/comments/x27szs/exposure_to_past_temperature_variability_may_help/,eurekalert.org,https://www.eurekalert.org/news-releases/963009,,Exposure to past temperature variability may help forests cope wi
post,x27rtd,2qh3l,news,false,1661932999,https://old.reddit.com/r/news/comments/x27rtd/england_faces_longer_drier_summers_due_to_climate/,cde.news,https://cde.news/?p=450652,,"England faces longer, drier summers due to climate change -Met Office",1
post,x27h9k,3hbcw,scmpauto,false,1661931887,https://old.reddit.com/r/SCMPauto/comments/x27h9k/business_climate_change_alibaba_boosts/,scmp.com,https://www.scmp.com/business/article/3190822/climate-change-alibaba-boosts-clean-energy-purchases-become-largest-buyer?utm_sour
post,x27grb,6wzx9b,eduwriters,false,1661931828,https://old.reddit.com/r/EduWriters/comments/x27grb/what_are_the_causes_of_climate_change_essay/,eduwriters.pro,https://eduwriters.pro/?cid=2860/?utm_source=rdwrt&amp;utm_medium=pst&amp;utm_campaign=what_are_the_causes_of_cl
post,x26sge,2e9vth,u_joggler-66,false,1661929264,https://old.reddit.com/r/u_joggler-66/comments/x26sge/climate_change_scam_pope_francis_the_jesuits/,youtube.com,https://youtube.com/watch?v=5v_YIc2s9Co&amp;feature=share,,"Climate Change Scam, Pope Francis &amp; the Jesuit
post,x263b1,2qjpg,memes,false,1661926684,https://old.reddit.com/r/memes/comments/x263b1/first_climate_change_now_this/,i.redd.it,https://i.redd.it/ddmfbotrrzk91.jpg,,"First climate change, now this?",144
post,x25ppz,5r8j6u,phnewsfeed,false,1661925299,https://old.reddit.com/r/phnewsfeed/comments/x25ppz/england_faces_longer_drier_summers_due_to_climate/,newsinfo.inquirer.net,https://newsinfo.inquirer.net/1656005/england-faces-longer-drier-summers-due-to-climate-change-nati
post,x25iiu,36vxjo,churchofcovid,false,1661924556,https://old.reddit.com/r/ChurchOfCOVID/comments/x25iiu/another_victim_of_the_climate_change/,i.redd.it,https://i.redd.it/fm4ih9vflzk91.jpg,,Another victim of the Climate Change? 😔,311
post,x251nx,2wtau,dirtykikpals,true,1661922937,https://old.reddit.com/r/dirtykikpals/comments/x251nx/24_m4f_using_sex_to_distract_from_rampant_climate/,self.dirtykikpals,,"If you’re looking to have some fun with abs and at least a 3rd grade reading level, I’m your guy! I

[Here’s a little sneak preview of me!](https://i.imgur.com/1z1uX6x.jpg)",24 [M4F] Using sex to distract from rampant climate change,1
post,x24qxr,36ywi,climate_science,false,1661921952,https://old.reddit.com/r/climate_science/comments/x24qxr/probabilistic_projections_of_increased_heat/,nature.com,https://www.nature.com/articles/s43247-022-00524-4,,Probabilistic projections of increased heat stress driv
post,x24p2i,2rzkz,osp,false,1661921769,https://old.reddit.com/r/osp/comments/x24p2i/cries_in_anachronism_and_climate_change/,i.redd.it,https://i.redd.it/6zaqs0n5dzk91.jpg,,*cries in anachronism and climate change*,765
post,x24izd,2v3da,futurewhatif,false,1661921199,https://old.reddit.com/r/FutureWhatIf/comments/x24izd/fwi_by_2050_most_areas_within_300_miles_of_chinas/,self.futurewhatif,https://www.reddit.com/r/FutureWhatIf/comments/x24izd/fwi_by_2050_most_areas_within_300_miles_of_chi
post,x24hp8,ik2qt,u_lateroyal,false,1661921080,https://old.reddit.com/r/u_LateRoyal/comments/x24hp8/hightech_wooden_cities_could_fight_fire_climate/,thehill.com,https://thehill.com/policy/equilibrium-sustainability/3621018-high-tech-wooden-cities-could-fight-fire-climate
post,x23kmm,2ryk4,kzoo,false,1661918134,https://old.reddit.com/r/kzoo/comments/x23kmm/looks_good_and_makes_sense_i_think_there_could_be/,imgur.com,https://imgur.com/a/qvmqjUi,,"Looks good and makes sense! I think there could be trees in median, since people will run over
post,x22u2c,37kq2,skincareaddicts,false,1661915902,https://old.reddit.com/r/SkincareAddicts/comments/x22u2c/is_this_due_to_dry_skin_currently_live_in_cool/,i.redd.it,https://i.redd.it/jga63e0pvyk91.jpg,,"Is this due to dry skin? Currently live in cool sometimes warm weat
post,x222wd,61n8ns,stopthenwo,false,1661913698,https://old.reddit.com/r/StopTheNWO/comments/x222wd/their_solution_to_climate_change_is_to_price_out/,self.conspiracy,/r/conspiracy/comments/x1gmh0/their_solution_to_climate_change_is_to_price_out/,,Their solution to climate
post,x21oiv,mouw,science,false,1661912569,https://old.reddit.com/r/science/comments/x21oiv/can_exceptionalism_withstand_crises_an_evaluation/,academic.oup.com,,[deleted],Can Exceptionalism Withstand Crises? An Evaluation of the Arctic Council's Response to Climate Change
post,x214bj,xpnox,askapriest,false,1661910999,https://old.reddit.com/r/AskAPriest/comments/x214bj/is_it_wrong_for_a_catholic_to_become_an_airline/,self.askapriest,,[removed],Is it wrong for a Catholic to become an airline pilot considering climate change?,2
post,x210is,3cct3,capitalismvsocialism,false,1661910690,https://old.reddit.com/r/CapitalismVSocialism/comments/x210is/once_we_run_low_on_water_because_of_climate/,self.capitalismvsocialism,,"UN-Water, the United Nations’ inter-agency coordination mechanism for  all water
```

DICCIONARIO

- Stop Words: Palabras comunes o frases que se utilizan con frecuencia y que generalmente no aportan un significado específico a una oración (ejemplo: "the", "is", "a", "an").

- Conjunciones: Palabras que se utilizan para unir cláusulas, frases o palabras dentro de una oración (ejemplo: "and", "but", "or", "nor").

- Preposiciones: Palabras que indican una relación espacial, temporal o lógica entre otras palabras en la oración (ejemplo: "in", "on", "at", "under").

- Pronombres: Palabras que se utilizan en lugar de un sustantivo para evitar la repetición (ejemplo: "he", "she", "they", "it").

- Caracteres Especiales y Números: Símbolos y dígitos que se utilizan en el lenguaje escrito y en la numeración (ejemplo: "%", "3", "$", "&").

- Emojis: Iconos expresivos digitales utilizados para representar emociones, estados de ánimo o conceptos en el texto escrito (ejemplo: 😊, 👍, 🎉).

- Malas Palabras: Términos ofensivos, vulgares o inapropiados que pueden causar incomodidad o ser considerados inadecuados en ciertos contextos.

APLICACIONES DESARROLLADAS

Preprocesamiento

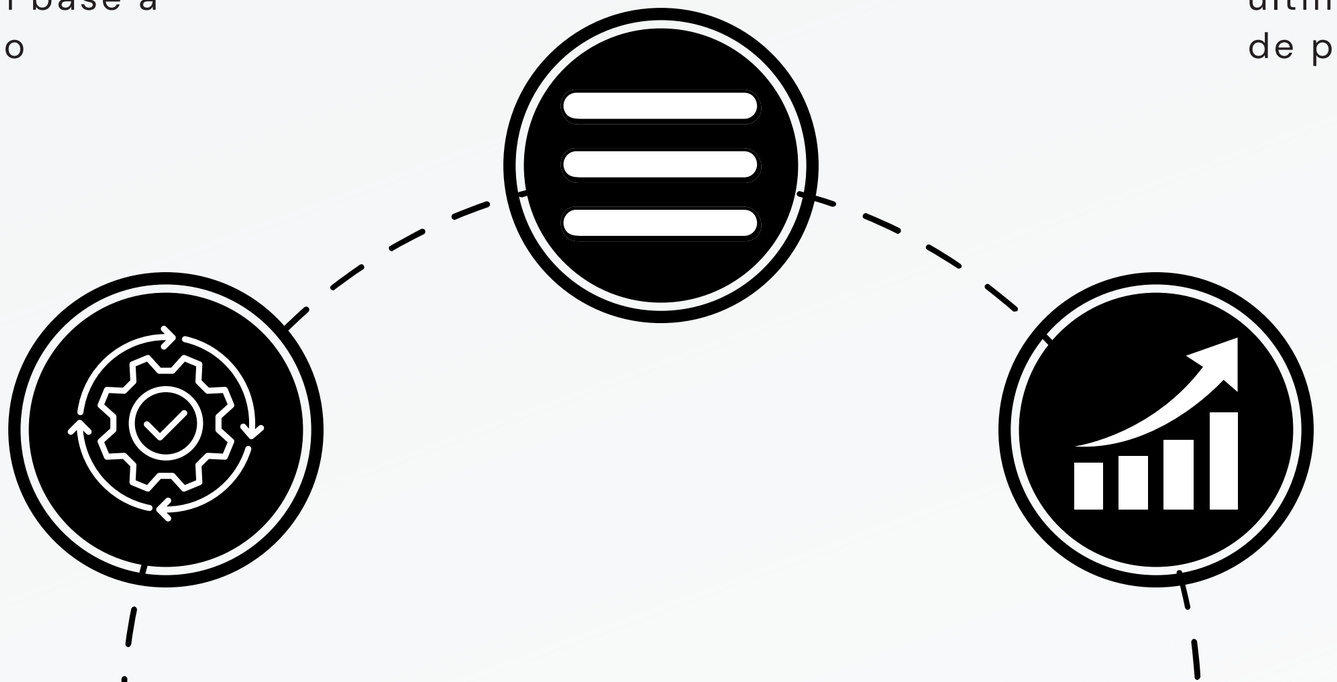
Programa realizado en java en el cual se remueve toda aquella informacion que no sea requerida para el proyecto. Se eliminaron columnas innecesarias y se filtraron palabras en base a un diccionario

Word Count

Programa en Java, utilizando librerías de Hadoop; encargado de contar todas las ocurrencias de una palabra. Dos iteraciones: Word Count para conjuntos de 1 palabra y Word Count para conjuntos de 2 palabras de acuerdo al Minimum Support

Frecuency Analysis

Programa escrito en Python. Entre las funciones importantes: Ordenar las palabras de mayor frecuencia a menor frecuencia, y por ultimo, filtrar el top 20 de palabras con mayor frecuencia.



PASO A PASO DE LA EJECUCIÓN

1. Ejecución de PreProcesamiento.java en conjunto con Dictionary,txt para realizar el procesamiento y obtener como resultado nuestro dataset final

```java
import java.io.BufferedReader;
import java.io.FileReader;
import java.io.FileWriter;
import java.io.IOException;
import java.util.ArrayList;

You, 2 hours ago | 1 author (You)
public class preProcesamiento {

    Run | Debug
    public static void main(String[] args) {
        long startTime = System.nanoTime();

        String datasetFile = "workingFiles\\datasetCompleto.txt";
        String dictionaryFile = "workingFiles\\Dictionary.txt";
        String outputFile = "workingFiles\\datasetProcesado.txt";
        cleanDataset(datasetFile, dictionaryFile, outputFile);

        long endTime = System.nanoTime();
        long elapsedTimeNano  = endTime - startTime;
        double elapsedTimeMinutes = (double) elapsedTimeNano / 1_000_000_000 / 60;
        System.out.println("Execution time: " + elapsedTimeMinutes + " minutes");
    }

    private static String removePunctuation(String line) {
        return line = line.replaceAll(regex:"&'.*?\\w+;", replacement:" ")
                .replaceAll(regex:"[^a-zA-Z ]", replacement:" ")
                .replaceAll(regex:"\\s+", replacement:" ");
    }
```

```java
    private static ArrayList<String> loadDictionary(String dictionaryFile) {
        ArrayList<String> dictionary = new ArrayList<>();
        try (BufferedReader dictReader = new BufferedReader(new FileReader(dictionaryFile))) {
            String word;
            while ((word = dictReader.readLine()) != null) {
                dictionary.add(word.trim().toLowerCase());
            }
        } catch (IOException e) {
            System.out.println("An error occurred: " + e.getMessage());
        }
        return dictionary;
    }

    public static void cleanDataset(String datasetFile, String dictionaryFile, String outputFile) {
        try {
            ArrayList<String> dictionary = loadDictionary(dictionaryFile);

            try (BufferedReader br = new BufferedReader(new FileReader(datasetFile));
                 FileWriter fw = new FileWriter(outputFile, append:true)) {
                String line;
                while ((line = br.readLine()) != null) {
                    line = removePunctuation(line);
                    String[] words = line.split(regex:"\\s+");
                    for (String word : words) {
                        String cleanedWord = word.toLowerCase();
                        if (!dictionary.contains(cleanedWord)) {
                            fw.write(cleanedWord + " ");
                        }
                    }
                    fw.write(str:"\n");
                }
            }
            System.out.println("Dataset cleaned successfully and saved to " + outputFile);
        } catch (IOException e) {
            System.out.println("An error occurred: " + e.getMessage());
        }
    }
```

2. Obtención de nuestro Archivo Final, preparado para su manejo en Hadoop y la aplicación de ambos programas de WordCount

```
speaking uk user feels recent development working environment impacts covid urgency macro agendas climate change trickling team management
climate change trigger hard ship massive scale common comforts enjoyed decades
 funny read reverse social economic labour weak crime repercussions national austeric focused widening gaps foster crime place pushing governance structures worsen majority poor working class maori fucked neo
 implies single event magnitude set events leading single magnitude earth sciences forgive basic question plate attachments locations late jurassic movie set climate compare recent history understanding earthq
busier climate change local ac units arent keeping normal replacing youll cheese sooner
projected future scarcity climate change ravages worlds ecosystems
irony majority people claim concerned climate change
armchair activists world issues worth fighting massive human rights battles fought globe victory lgbtqa rights fought usa saudi arabia muslim world swathes asia europe russia ignores issues racism ableism sexi
 decade impending catastrophe realized changes john kerry telling public climate change sea levels rising buying beach house biggest fuel skepticism npr sections thetwo years sugar industry paid scientists bla
problem advance faster profit expense environment systematic growth extracting earth transform products energy decentralised solution emerges reason highly reduce impact environment stop systems destruction co
 teeth slaughter dinner chew raw bone lol chuckle generally discussions morality killing animals empathetic empathetic kids grandkids future humans borrowing planet dirty leave worse condition childfree due cl
intergovernmental panel climate change ipcc provide executive summaries reports idea language simple understandable completely objective people questioning objectivity vested interests ipcc ch report ar wg dow
generating refugees tactic instance actors iraq fund arms wars end chaos victims end problem faster permanently actors viable political groups iraq helping refugees money relief organizations political freedom
 people taking action agains government civil war worse civil war abject tyranny hands authoritarian government trump president people defended political violence blm riots righteous struggle evil oppressor de
flying fossil fuel commercially climate change
 talking strangers internet conversation makes sad completely adoption thinking momentary omg world worsening climate change wars minute world kids future people change opinions explore process
climate change scaring worse
brilliant economists understood massive short long term economic benefits finally passing legislation aimed addressing climate change
helping climate change disgusting
slay climate change
 scientists hundreds thousands disagree verbiage emergency speaking climate change climate change global warming headline found read article nutshell scientists climate change uniform measures assessing conside
congratulations hiked portion young girl scout years chance hear nyt article nytimes travel climate change pacific crest trail searchresultposition nytimes travel climate change pacific crest trail searchresul
climate change fake news
considered purpose meeting change scenery live south primarily family men guys wanna hookup luckily great job relocate pretty considered living country thinking research singles population importantly crime ra
 true republican votes infrastructure bill talked climate terms packed fair amount climate punch passed bipartisan basis growing silence climate open opportunities absolutely single republican state lifted fin
climate change deniers fucked hard
bbc future article water shortages brewing warshttps bloomberg news articles global warming destroying crops stopped kw js press gaef doc htmhttps usnews news countries articles climate change increase global
climate crisis solved single person planet contributes reality corporations causing climate change long people continue giving companies money continue destroy planet easiest simply stop supporting corporation
heres thinking somethings complaining climate change ftr climate change festival attended somethings waste post pilfering care rest end land fill absolutely gross
climate change happening billionaires suddenly died
deflection rife current political scenario important understand history learn consequences actions easy politicians deflect current crises face climate change inequality corruption diminishing freedom press le
man washington state requires smog testing support climate change measures tax
tax revenue huge consideration case development flood fringe frowned climate change room future growth
 suppose nasty habit dismissing problems existed born kids kinda bad people expecting expecting eventual societal collapse climate change pressures people care idea
high temperatures stress californias electrical system warned state grid operator energy demands rise largely due air conditioning weekend temperatures supposed hottest electricity conservation avoid outages f
 israel official total mortality numbers released cbs gov il pages search tablemaps aspx cbssubject century pandemic vaccines total mortality cbs gov il publications lochuttlushim xlsx dramatically higher coun
 posted graph meme attempted discredit climate change claiming shitty data deserved notoriety actual climate change claiming amount coverage people hid post climate change
 main points personal transportation responsible part global carbon emissions environment impact general actual comparison environmental impact ices evs difficult quantify address absolutely difficult lithium
part solar panel company cap vote climate change measures increases companies worth cap resign immediately price
speedy bullets racist fascist caused climate change
image shows horrific amount pollution pollution caused climate change day heat wave drought stupid
character arc support fighting climate change
```

```
PS C:\Users\danie\Escritorio\datasetWorking> cd "c:\Users\danie\Escritorio\datasetWor
king\" ; if ($?) { javac preProcesamiento.java } ; if ($?) { java preProcesamiento }
Dataset cleaned successfully and saved to workingFiles\datasetProcesado.txt
Execution time: 22.7164150966666665 minutes
PS C:\Users\danie\Escritorio\datasetWorking>
```

01

HADOOP

Instalar Hadoop

preparar nuestro WordCount.java

02

UPLOADING

Subir nuestro input (dataset despues del preprocesamiento) a hadoop, igualmente crear una carpeta para nuestro output

03

CLASSES

En Linux creamos una carpeta para que aloje nuestras classes.

javac -classpath ${HADOOP_CLASSPATH} -d <class folder> <.java>

corremos este comando para crear nuestras clases provenientes de nuestro archivo .java

04

JAR

Creamos un jar con este comando

jar -cvf <nombre jar> -C <class folder>

05

RUNNING

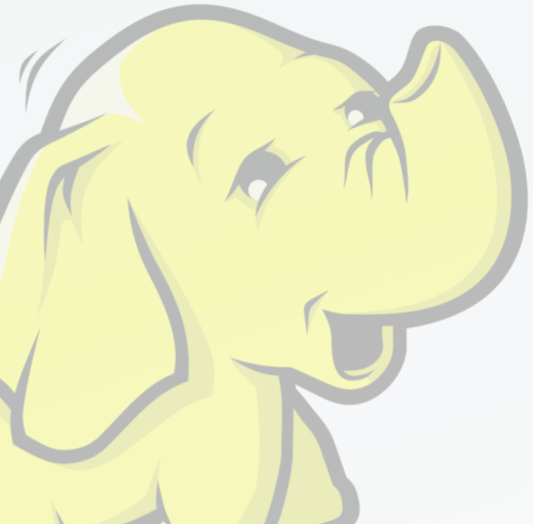
Corremos nuestro wordcount con hadoop de esta manera

hadoop jar <.jar> <class name> <input hadoop> <output hadoop>

06

OUTPUT

Hadoop guarda el output un archivo con el nombre part-r-00000 en nuestra carpeta de output



FRECUENCIA DE UNO

```
hdoop@DESKTOP-6FGOBRI:~/proyecto$ ls
WordCount.java   classes   datasetProcesadoFinal.txt    outputNuevo   outputSingapur  wordcount.jar
WordCount2.java  classes2  datasetProcesadoTesting.txt  outputNuevo2  singapur.txt    wordcount2.jar
hdoop@DESKTOP-6FGOBRI:~/proyecto$ hdfs dfs -mkdir /wordcount/inputFINAL
hdoop@DESKTOP-6FGOBRI:~/proyecto$ hdfs dfs -put datasetProcesadoFinal.txt /wordcount/inputFINAL
hdoop@DESKTOP-6FGOBRI:~/proyecto$ hadoop jar wordcount.jar WordCount /wordcount/inputFINAL /wordcount/outputFINAL
2024-03-18 22:28:27,109 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /127.0.0.1:8032
2024-03-18 22:28:27,538 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2024-03-18 22:28:27,562 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hdoop/.staging/job_1710821039990_0002
2024-03-18 22:28:27,821 INFO input.FileInputFormat: Total input files to process : 1
2024-03-18 22:28:28,291 INFO mapreduce.JobSubmitter: number of splits:13
2024-03-18 22:28:28,851 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1710821039990_0002
2024-03-18 22:28:28,851 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-03-18 22:28:29,048 INFO conf.Configuration: resource-types.xml not found
2024-03-18 22:28:29,049 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-03-18 22:28:29,161 INFO impl.YarnClientImpl: Submitted application application_1710821039990_0002
2024-03-18 22:28:29,207 INFO mapreduce.Job: The url to track the job: http://DESKTOP-6FGOBRI.:8088/proxy/application_1710821039990_0002/
2024-03-18 22:28:29,208 INFO mapreduce.Job: Running job: job_1710821039990_0002
```

FREQUENCIA DE DOS

```
hdoop@DESKTOP-6FGOBRI:~/proyecto$ hadoop jar wordcount2.jar WordCount2 /wordcount/inputFINAL/ /wordcount/output2
2024-03-21 00:58:47,066 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /127.0.
0.1:8032
2024-03-21 00:58:47,363 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Im
plement the Tool interface and execute your application with ToolRunner to remedy this.
2024-03-21 00:58:47,396 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/
staging/hdoop/.staging/job_1711004240233_0001
2024-03-21 00:58:48,192 INFO input.FileInputFormat: Total input files to process : 1
2024-03-21 00:58:49,066 INFO mapreduce.JobSubmitter: number of splits:13
2024-03-21 00:58:49,571 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1711004240233_0001
2024-03-21 00:58:49,571 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-03-21 00:58:49,710 INFO conf.Configuration: resource-types.xml not found
2024-03-21 00:58:49,710 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-03-21 00:58:49,910 INFO impl.YarnClientImpl: Submitted application application_1711004240233_0001
2024-03-21 00:58:49,946 INFO mapreduce.Job: The url to track the job: http://DESKTOP-6FGOBRI.:8088/proxy/applica
tion_1711004240233_0001/
2024-03-21 00:58:49,947 INFO mapreduce.Job: Running job: job_1711004240233_0001
2024-03-21 00:58:57,048 INFO mapreduce.Job: Job job_1711004240233_0001 running in uber mode : false
2024-03-21 00:58:57,049 INFO mapreduce.Job:  map 0% reduce 0%
2024-03-21 00:59:16,487 INFO mapreduce.Job:  map 3% reduce 0%
2024-03-21 00:59:17,524 INFO mapreduce.Job:  map 5% reduce 0%
```

RESULTADOS - TOPS 20

1

```
climate 6573438
change 6123095
people 2536033
world 968694
years 863891
global 752283
science 619737
trump 600299
energy 533309
warming 495085
problem 483502
government 477833
carbon 467456
power 423618
human 421047
money 420800
life 407081
earth 402779
issue 396016
political 384653
```

2

```
climate change 5359453
global warming 332551
fossil fuels 130939
fossil fuel 99723
change deniers 88565
long term 87244
change climate 81418
united states 78292
carbon tax 72916
global climate 67956
fight climate 64252
change denier 63096
change denial 62231
renewable energy 62193
due climate 61034
donald trump 59589
nuclear power 59507
effects climate 59015
change people 57613
man climate 54779
```

HALLAZGOS INTERESANTES

trump 600299
donald trump 59589

biden 262533

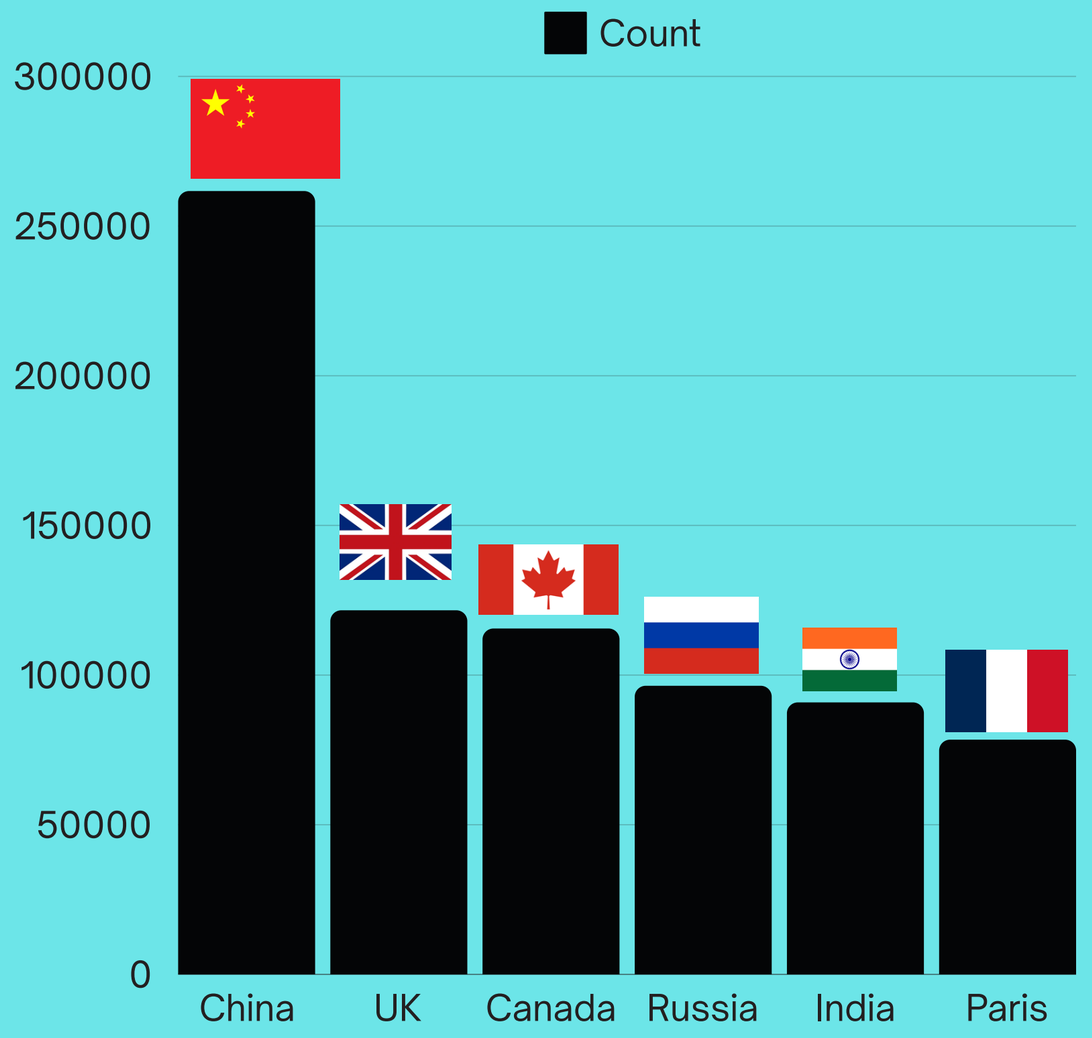
bernie 198680
bernie sanders 23659

```
donald trump 59589
supreme court 28228
republican party 27996
climate bernie 27408
nasa gov 25513
paris agreement 25084
bernie sanders 23659
democratic party 22765
united nations 21766
hillary clinton 17084
prime minister 16267
trump climate 16043
trump administration 14966
```

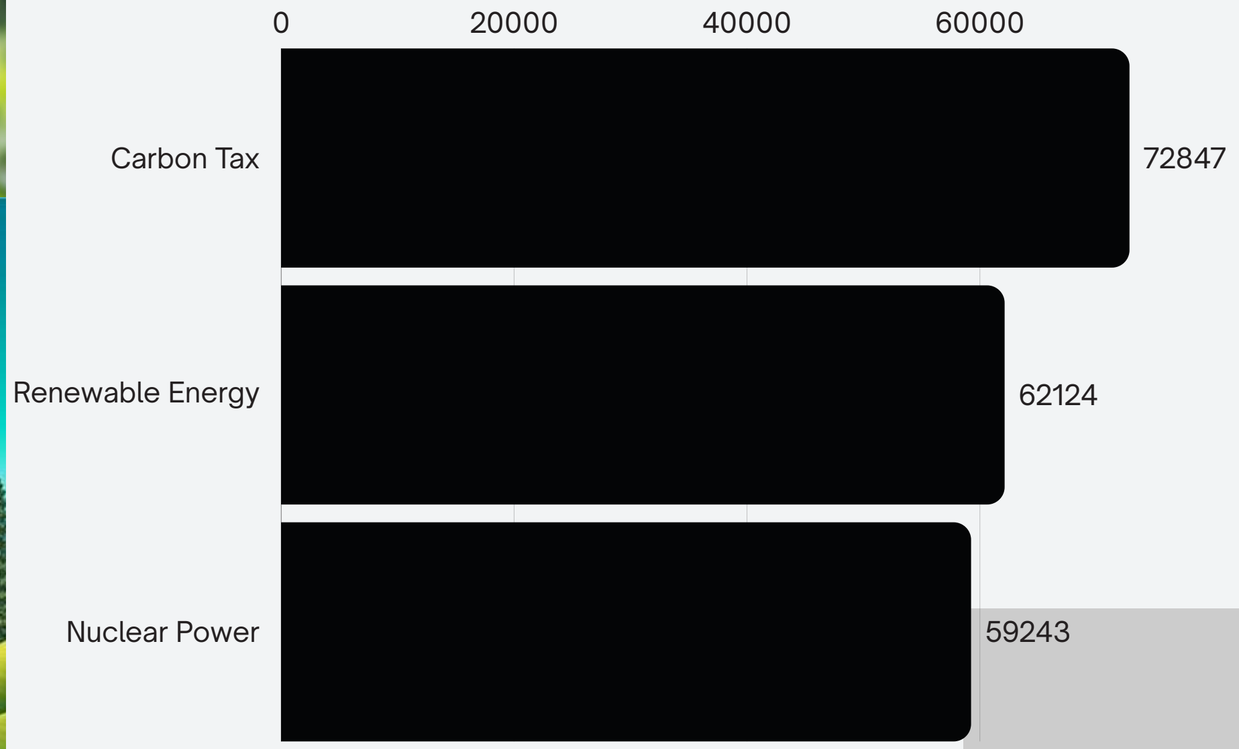
```
trump 600299
biden 262533
politics 225635
bernie 198680
president 177143
republicans 171780
policies 170226
major 169764
conservative 152602
politicians 132563
republican 132330
obama 121667
```

HALLAZGOS INTERESANTES

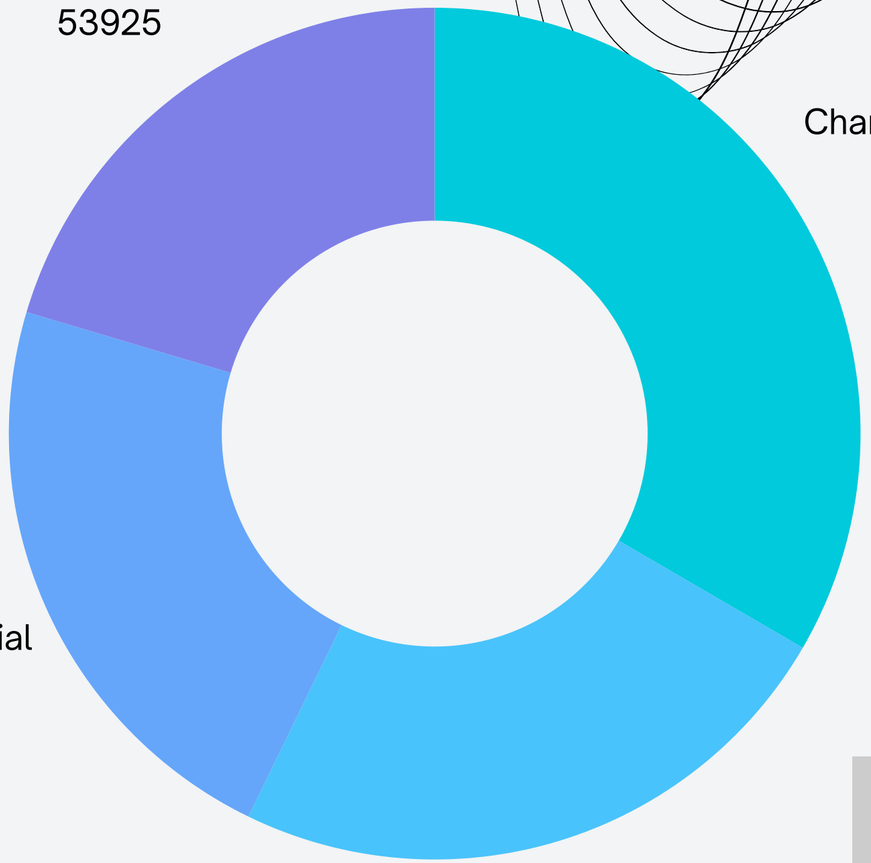
america 190689
american 184275
united states 78292



HALLAZGOS INTERESANTES



HALLAZGO INTERESANTE



EN RESUMEN

- **Conciencia Global**
  - La alta frecuencia de "climate change" y "global warming" tanto en conjunto como por separdado indica que estos son los conceptos centrales en las discusiones sobre el cambio climático, también una preocupación generalizada por el aumento de la temperatura global y sus efectos.

- **Efectos a Largo Plazo**
  - "long term" y "effects climate" sugiere un interés en los efectos a largo plazo del cambio climático en el medio ambiente y la sociedad. Esto podria indicar una preocupación por los impactos futuros del cambio climático y la necesidad de tomar medidas preventivas y adaptativas.

- **Energía, Recursos y Tecnología**
  - "fossil fuels", "renewable energy", "power", y "carbon" pueden girar en torno a la transición energética y la reducción de la de los combustibles fósiles. Esto reflejajaria una preocupación por las emisiones de carbono y el impacto ambiental de la quema de combustibles fósiles.

EN RESUMEN

- **Negacionismo**
  - Términos como "change deniers" y "change denial" sugieren que hay un debate activo por parte de aquellos que niegan la existencia o la importancia del cambio climático.

- **Política y Gobierno**
  - La presencia de palabras como "government", "donald trump", y "political" sugiere un interés en el papel de los gobiernos y los líderes políticos en la gestión y la respuesta al cambio climático. Esto sugiere que las políticas gubernamentales y las acciones políticas son puntos importantes al abordar el tema del cambio climático.

- **Impacto Económico**
  - Palabras como "money" sugiere que el impacto económico del cambio climático es una preocupación importante. Esto podría implicar discusiones sobre los costos de la mitigación y adaptación al cambio climático, así como el potencial económico de las soluciones sostenibles.

MUCHAS GRACIAS POR SU ATENCIÓN