AD: AUTO-GEMINI

A reimplementation of

AD-AutoGPT: An Autonomous GPT for Alzheimer's Disease Infodemiology

Created by –

Dev Rishi Verma

24.10.2024

Abstract

This report presents a comprehensive analysis of AD AUTO-GEMINI, a cutting-edge reimplementation of AD AUTO-GPT that leverages the AUTO-GPT framework. Developed as part of my internship at the CogAI4Sci Lab at the National University of Singapore, AD AUTO-GEMINI autonomously executes sophisticated data collection, processing, and analysis of complex health narratives related to Alzheimer's Disease. By integrating advanced machine learning algorithms and natural language processing techniques, this innovative system aims to deepen insights and enhance public awareness surrounding Alzheimer's, ultimately driving informed decision-making and fostering improved health outcomes in the community.

INTRODUCTION

In this project, I chose to use Google Colab instead of a local IDE for several reasons. Google Colab offers a hassle-free environment, eliminating the need for manual library installations or local setup. It enables users to store their work in the cloud, facilitating easy collaboration and sharing of code. This convenience and flexibility make it an ideal platform for AD-AUTOGPT, allowing me to focus on development without the burden of managing everything locally.

Regarding the API choice, I opted for the Gemini API (1.5 Flash) over the GPT-4 API because it is free and sufficiently meets my current performance needs. Thus, I refer to my project as “AD AUTO-GEMINI” rather than “AD AUTO-GPT.” The Gemini API provides reliable results without incurring additional costs. However, for tasks that demand greater precision and advanced capabilities, I would recommend utilizing OpenAPI's GPT-4 or GPT-4-turbo. These models generally yield better results for more complex tasks, but they come at a price. Therefore, investing in OpenAPI models would be advantageous for projects where quality and precision are paramount.

Note 1: For the reimplementation of the code, I have utilized BBC as my primary news retrieval platform. This approach can easily be expanded to include additional sources such as the Alzheimer’s Society, Mayo Clinic, and the National Institute on Aging (NIA) by enhancing the web scraping capabilities.

Note 2: To improve the explainability and understanding of the code's output, I incorporated the ability to save the output of each function in the instruction library to independent text files. This facilitates the retrieval and review of findings, allowing users to track and analyze performance without needing to restart the full process.

Note 3: I utilized ChatGPT and Claude AI to assist in writing certain sections of the report and code; however, the entire work is my own, entirely authentic, and self-implemented.

CODE OVERVIEW

```
SEARCH_ENGINE_ID="YOUR SEARCH ENGINE API "
GOOGLE_API_KEY="YOUR GOOGLE API"
GEMINI_API_KEY="YOUR GEMINI API"
```

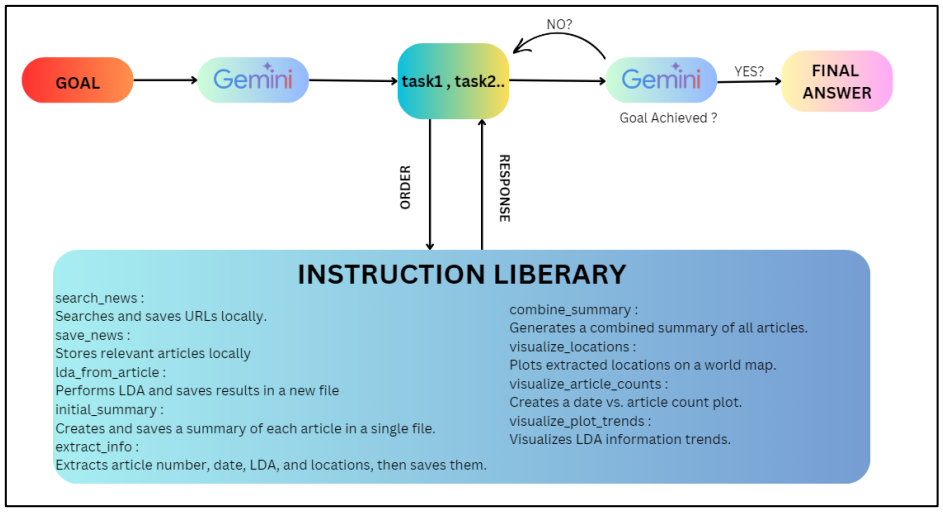
Define API and search ids at the start of implementation for future use.

```
import requests
from google.colab import files
import re
from bs4 import BeautifulSoup
import json
import google.generativeai as genai
import folium
from geopy.geocoders import Nominatim
import time
from IPython.display import IFrame
import pandas as pd
import matplotlib.pyplot as plt
from datetime import datetime
import gensim
from gensim import corpora
from gensim.models import LdaModel
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
import nltk
nltk.download('punkt')
nltk.download('stopwords')
import os
```

Importing various necessary libraries that will be utilized for building our instruction library. These include requests for making HTTP requests, BeautifulSoup for parsing HTML content, and json for handling JSON data. Additionally, google.generativeai is used for generative AI tasks, while folium and geopy help with mapping and location extraction. Libraries like pandas and matplotlib are included for data manipulation and visualization, respectively. Finally, gensim and nltk are imported for natural language processing tasks, including topic modelling and tokenization.

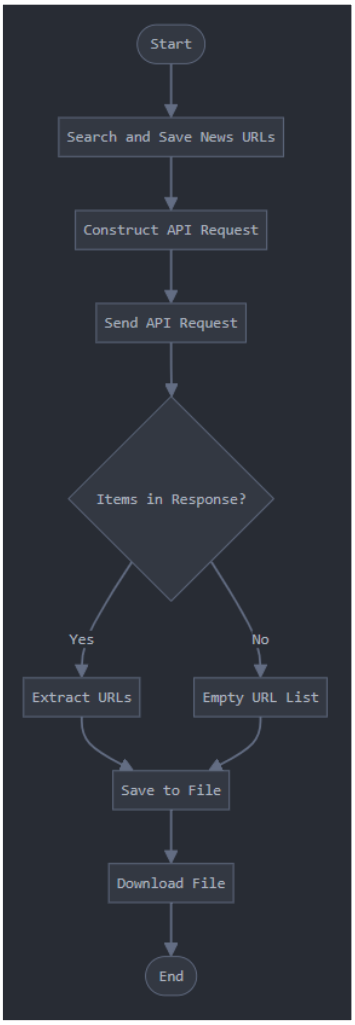
Building Instruction Library

In this section I'll brief out the instruction library



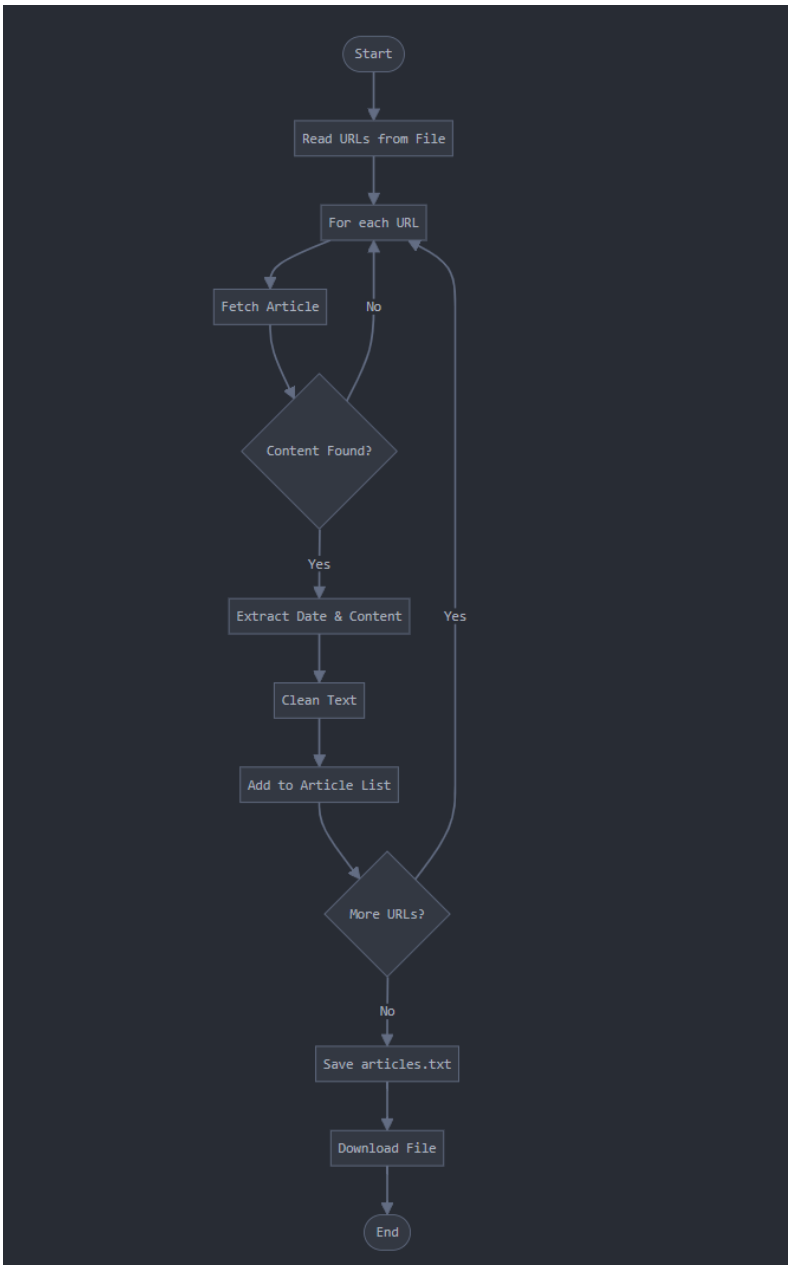
AUTO-GEMINI framework

Search_news



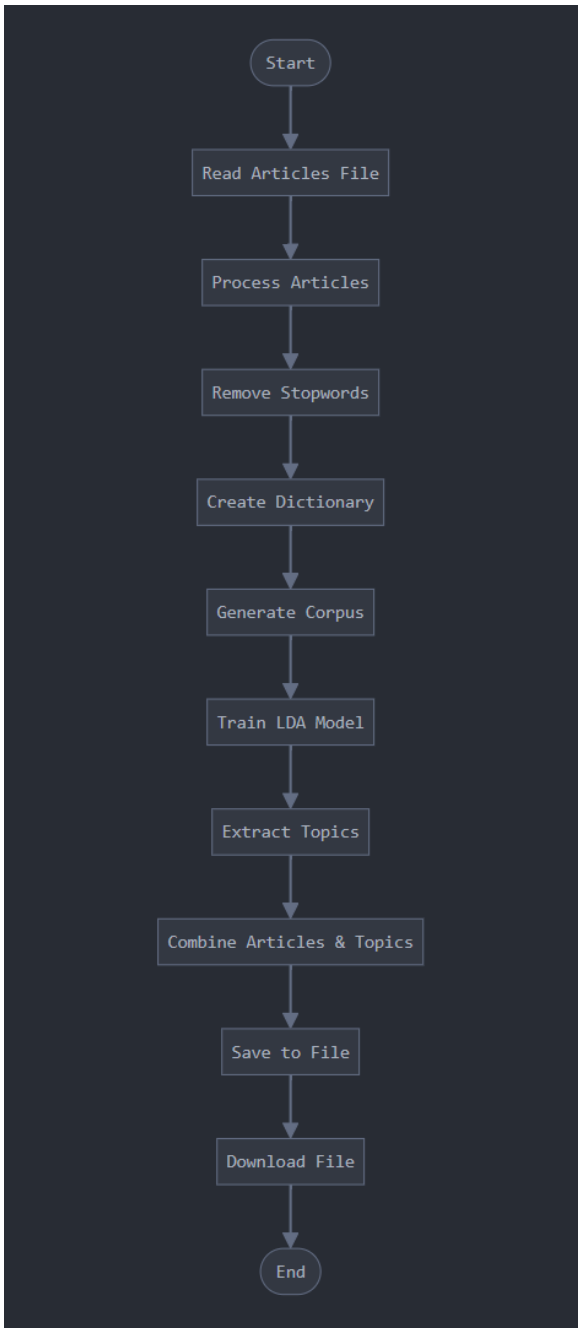
The “search_news” function queries the Google Custom Search API to find news articles from BBC News based on a provided search query. It constructs a request URL using the Google API key and search engine ID, then retrieves the JSON response containing the search results. The function extracts the article URLs from the response and saves them to a text file named “extracted_urls.txt”. Finally, it initiates a download of this file.

Save_news



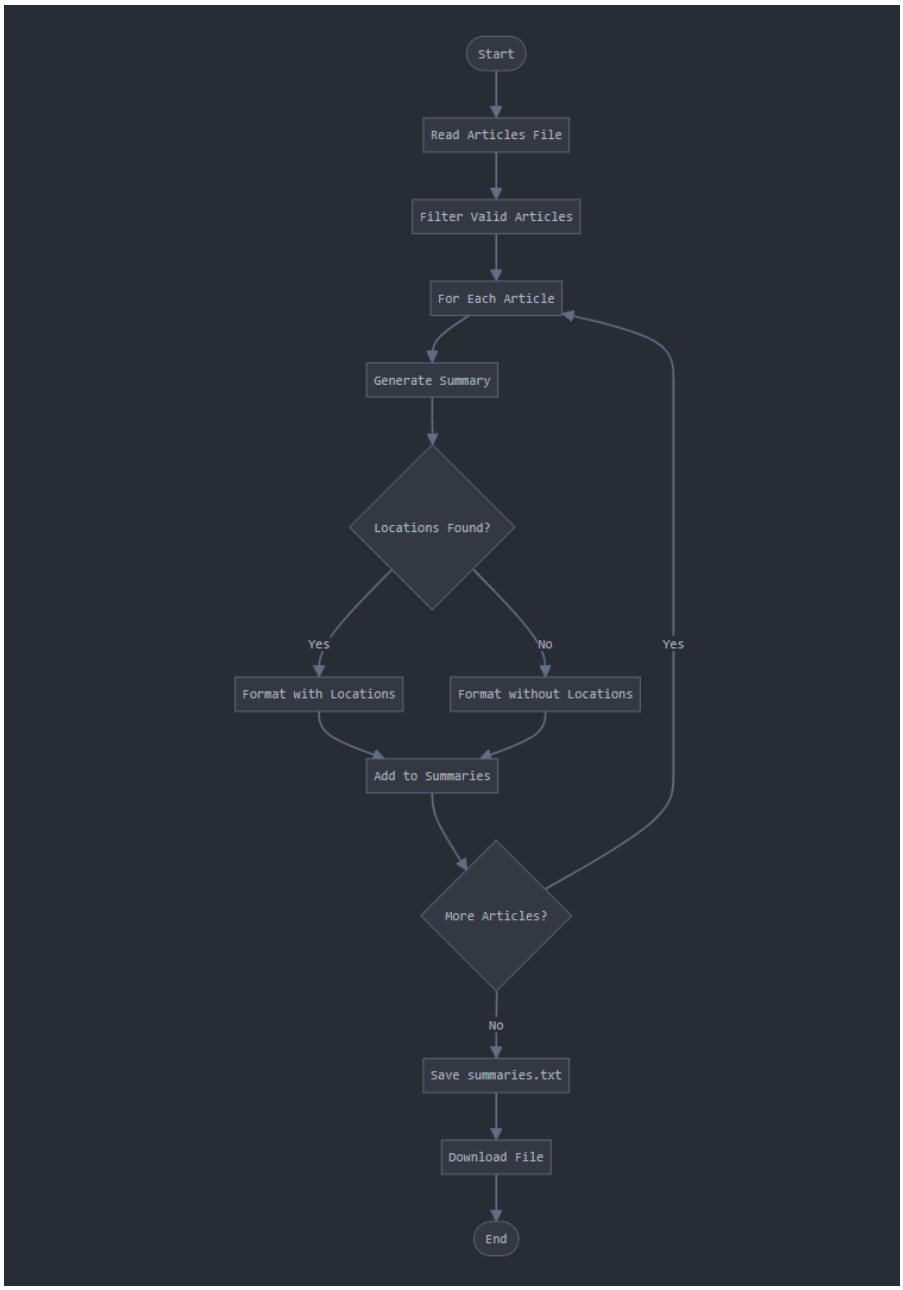
The “save_news” function extracts articles from a list of URLs provided in “extracted_urls.txt” and saves the content locally in a new file. It begins by reading the URLs and initializing a content storage list. For each URL, the function sends a request to retrieve the webpage, then parses the HTML using BeautifulSoup. It extracts the publication date from the JSON-LD script tag and the main article content. A cleaning function is applied to format the text by removing unwanted characters and excessive whitespace. Finally, the function compiles the article details, including the URL, date, and cleaned content, and saves them to a file named “articles.txt”, which is then made available for download.

Lda_from_articles



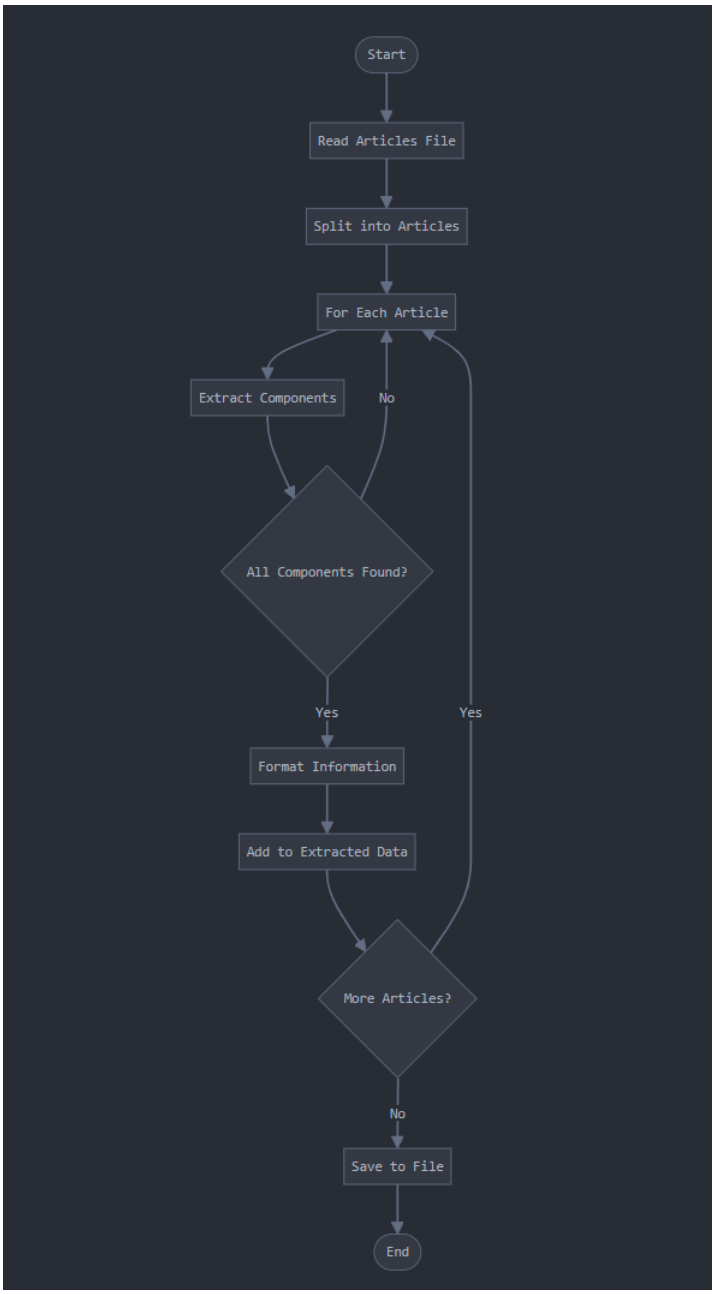
The “lda_from_articles” function processes our saved articles to extract Latent Dirichlet Allocation (LDA) topics. It begins by defining a set of custom stopwords and a helper function to preprocess the article content, which includes tokenization and filtering out stopwords. The main function reads articles from a specified file, skipping any entries with missing content or dates. It then tokenizes the cleaned articles to create a bag-of-words corpus and builds an LDA model to identify topics. Finally, the function saves the articles along with their corresponding LDA topics to a text file named “articles_with_lda.txt” and initiates a download of this file for the user.

Initial_summarize



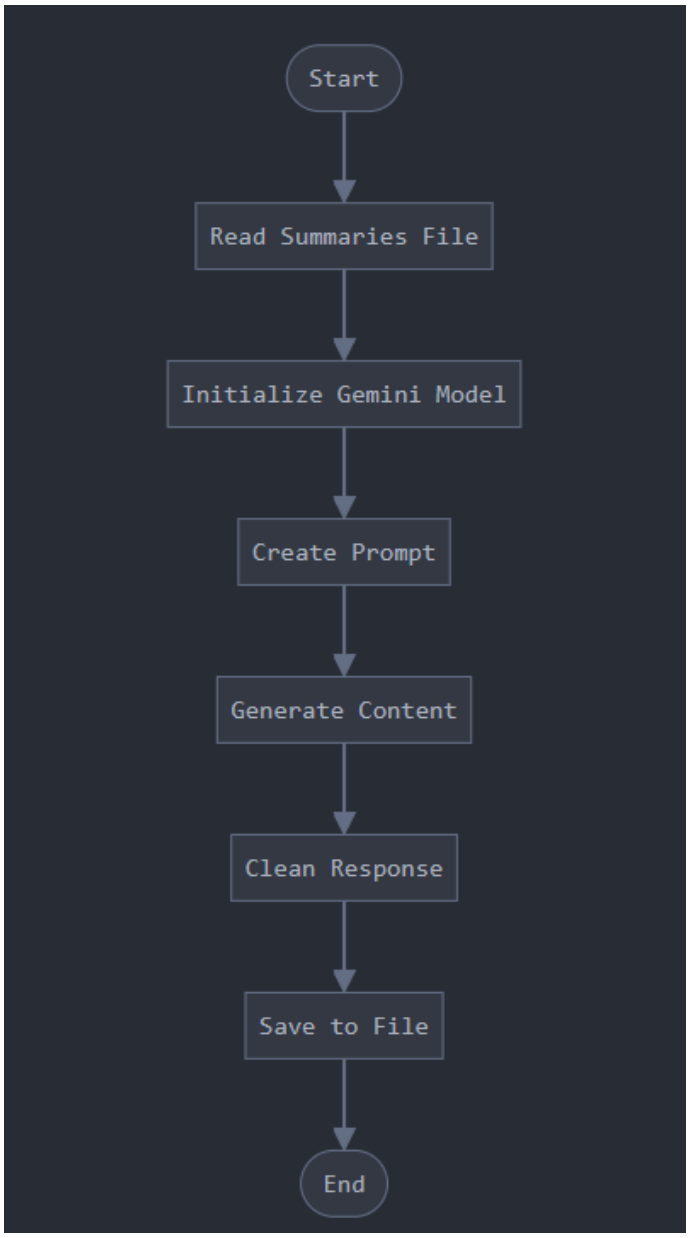
The “initial_summarize” function summarizes articles and extracts relevant locations. It reads articles from a file and skips any that don’t have enough content. For each valid article, it creates a prompt for the Gemini API to generate a summary and identify locations mentioned in the text. The function saves the summaries, along with URLs, dates, LDA topics, and extracted locations, to a text file called “summaries.txt”, which the user can download.

Extract_info



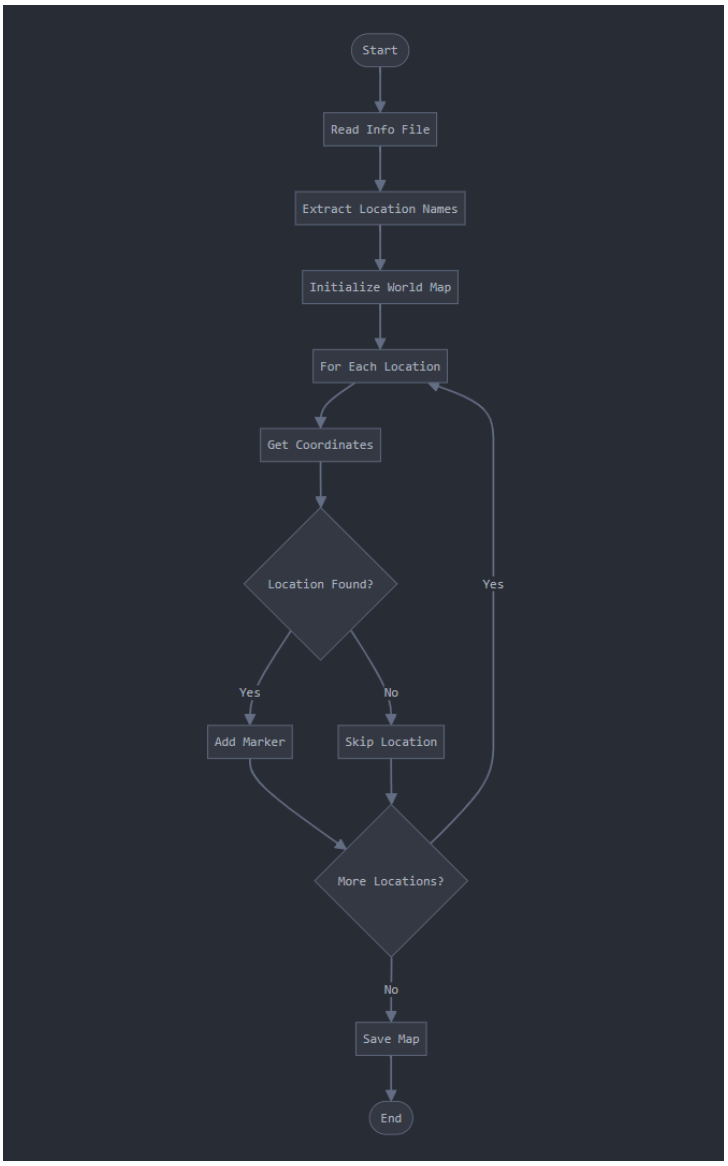
The “extract_info” function extracts key details from articles, including the article title, publication date, LDA topic, and mentioned locations. It reads the articles from a specified file, splitting them by double newlines. Using regular expressions, the function identifies and retrieves the relevant information. If all required details are found, it formats them and appends them to a list. Finally, the extracted information is saved to a file named “extracted_info.txt”, which the user can download.

Combined_summary



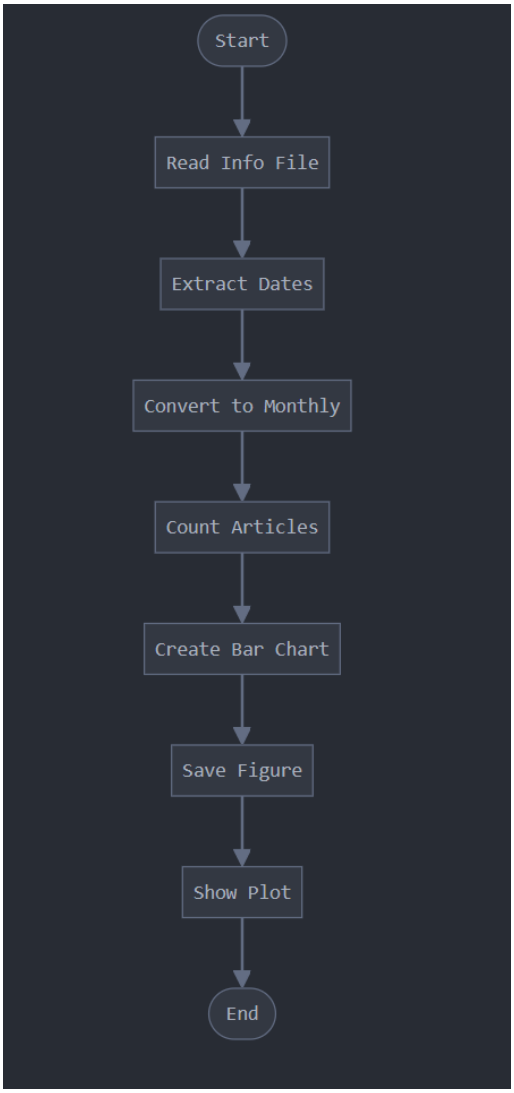
The “combined_summary” function generates a cohesive summary, and a compelling title based on a set of article summaries. It reads the summaries from a specified file and initializes a generative model to create the content. The function constructs a prompt that requests a comprehensive response regarding Alzheimer's disease, asking for key insights, relevant statistics, and spatiotemporal data from the provided summaries. Once the model generates the response, the function cleans up the text and saves the final summary to a file named “combined_summary.txt”, which the user can download.

Visualize_locations



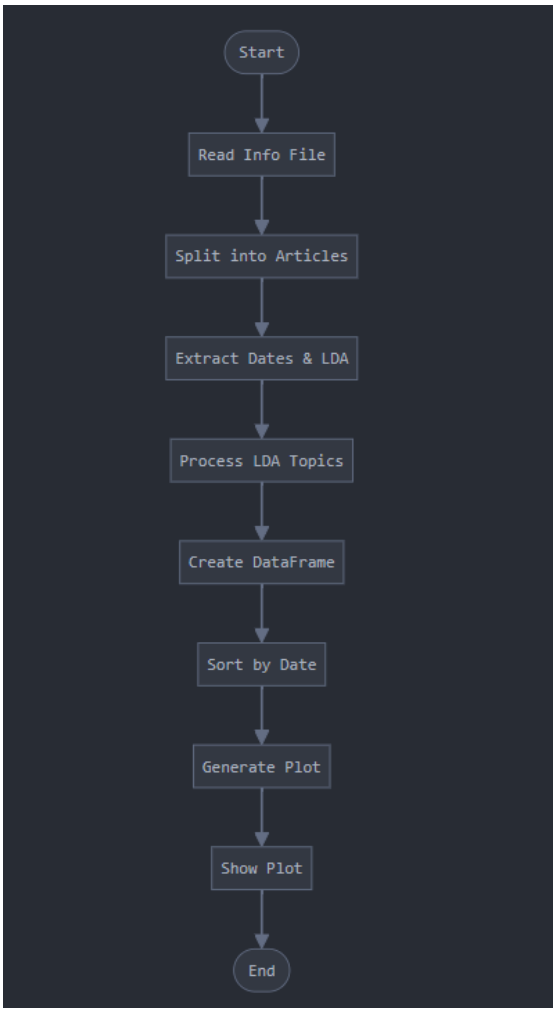
The “visualize_locations” function creates a world map highlighting key locations extracted from article summaries. It reads location data from a specified file, uses a regular expression to identify unique names, and fetches geographic coordinates with the Nominatim geocoder. Markers are added for each found location, and the map is saved as an HTML file named “world_map.html”, which is then available for download. This function provides a visual representation of important locations discussed in the articles.

Visualize_article_count



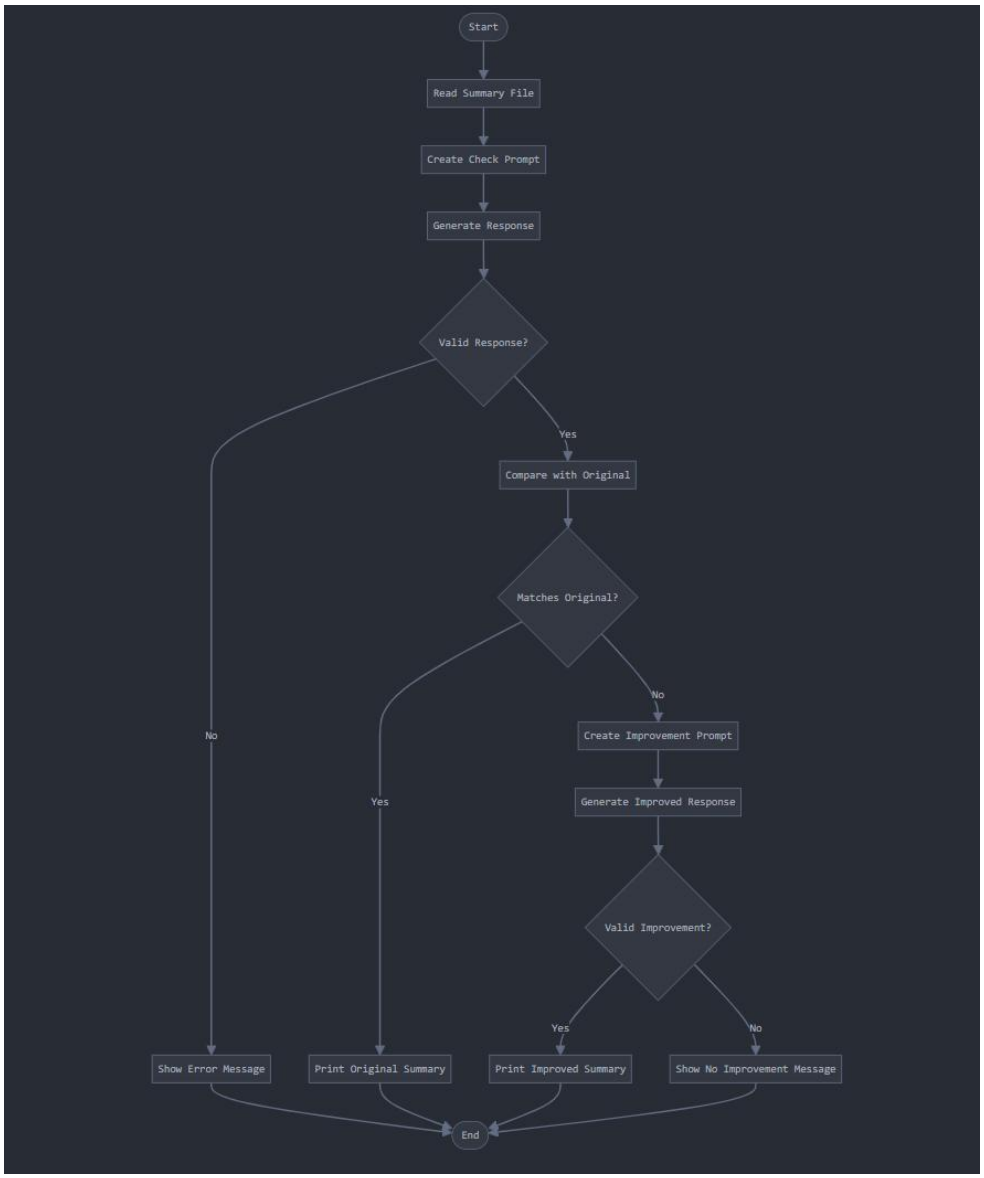
The “visualize_article_counts” function generates a bar chart to visualize the number of articles published each month. It reads dates from the “extracted_info.txt” file, extracts and converts them to monthly periods, and counts occurrences for each month. The function then creates a bar chart with a clear title and axis labels, saves the chart as “articles_per_month.png”, and provides it for download. Finally, the plot is displayed to provide a visual overview of article publication trends over time.

Visualize_plot_trends



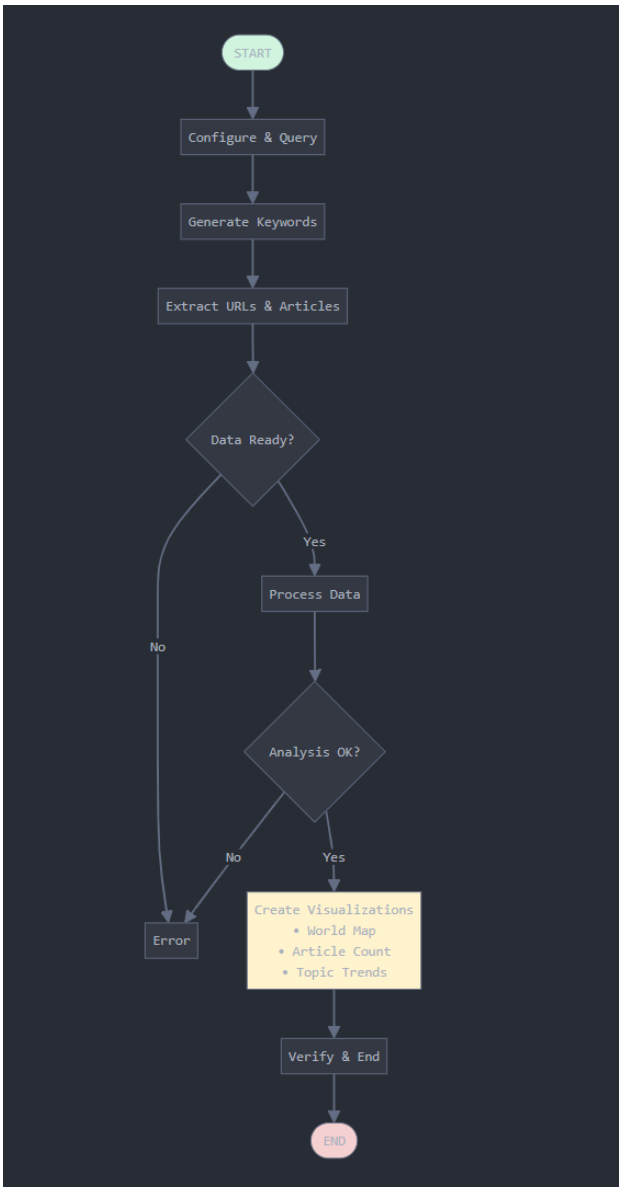
The “visualize_plot_trend” function creates a stacked area plot to visualize the trend of LDA topics related to Alzheimer's disease over time. It reads summary text from a specified file, extracts the publication date and LDA topic information for each article, and collects the importance of each topic. The function organizes this data into a DataFrame, sorts it by date, and then generates a stacked area plot to illustrate how the importance of various topics has changed over time. The plot is labeled clearly, providing insights into trends in Alzheimer's disease research topics across the timeline.

Check_summary



The “check_summary” function evaluates if a generated summary addresses a user's question. It reads the summary from “combined_summary.txt” and constructs a prompt combining the user's question with the summary content. The model generates a response based on this prompt. If the response includes the summary, it confirms it as correct. If not, the function prompts the model to improve the summary by adding relevant information. Finally, it outputs the enhanced summary if available or notifies the user if no valid improvement was generated, while also handling any processing errors.

Main_function



The main function of my code logic is to use the Gemini API to break down the user query into two major keywords. For example, “Can you help me to know something new about Alzheimer’s disease cure?” gets decomposed to “Alzheimer's cure.” This is then sent to the “search_news” function to retrieve URLs, which are saved locally. After confirming that the URLs have been successfully saved, “save_news” is called. This chain repeats until the final visualizations and summary are saved, after which Gemini checks if the summary matches the user query. If it doesn't, Gemini improves it and prints it; if it does, it prints it as is. I am aware that this approach is quite different from what is outlined in the actual paper, but the results I've obtained are quite similar, and I am currently working on optimizing it according to the paper's methodology.

OUTPUT ANALYSIS

In this section I'll be discussing the output of each function of the instruction library from the very start.

User Query: Can you help me to know something new about Alzheimer's disease cure?

Extracted search query: Alzheimer's Cure (decomposing the user query into 2 words makes it simpler to search and gave better results)

Thought: I need to use the tool "Search and Save news" to find answers to Can you help me to know something new about Alzheimer's disease cure?

Extracting URLs and saving locally

Action: Search and Save news URLs

Searching for: Alzheimer's Cure

Extracted 10 URLs and saved to 'extracted_urls.txt'.

Observation: URLs extracted and saved successfully.

Output:

```
https://www.bbc.com/news/health-66221116
https://www.bbc.com/news/health-66816268
https://www.bbc.com/news/health-24478942
https://www.bbc.com/news/uk-wales-65166225
https://www.bbc.com/news/health-54531075
https://www.bbc.com/news/world-latin-america-13428265
https://www.bbc.com/news/health-65471914
https://www.bbc.com/news/uk-wales-north-east-wales-11532508
https://www.bbc.com/news/articles/c20m4e27enko
https://www.bbc.com/news/videos/cvgl0z34ergo
```

Note: The output clearly shows 10 URLS getting saved from BBC News and all of them are related to decomposed user query of "Alzheimer's cure".

Thought: Now that URLs are extracted, the next step is to save the articles

Action: Save news articles

Action Input: File containing extracted URLs

Observation: Articles saved successfully.

Output:

```
Article 1
URL: https://www.bbc.com/news/health-66221116
Date: 2023-07-17T14:16:16.000Z
Content:
Drug donanemab seen as turning point in dementia fi
```

The output in the image above is for article 1 there are 10 more articles below this in same format

Note: As one can see the 10th article is a video therefore it does not get saved and is skipped

---

Thought: Now that articles have been saved, it's time to get LDA information from the articles

Action: Extract LDA information for the articles

Action Input: File containing articles

Observation: Successfully extracted LDA information from the articles and saved to a new file

```
Article 1
URL: https://www.bbc.com/news/health-66221116
Date: 2023-07-17T14:16:16.000Z
LDA Topic: 0.001*"alzheimers" + 0.001*"disease" + 0.001*"brain" + 0.001*"dementia" + 0.001*"drugs" + 0.001
*"research" + 0.001*"family" + 0.001*"drug" + 0.001*"scientists" + 0.001*"says"
 Content: Drug donanemab seen as turning point in dementia fightGetty ImagesA new drug, donanemab, is being hailed
as a turning point in the fight against Alzheimers, after a global trial confirms it slows cognitive decline.The
```

Note: As you can see LDA Topics have successfully been extracted and are saved below URL & Date this is done for each article

---

Thought: Now that LDA information has been extracted, it's time to summarize the articles

Action: Summarize the articles

Action Input: File containing articles

Skipping article 10 due to insufficient content.

Observation: Successfully summarized the articles and saved to a new file

```
Article 1
URL: https://www.bbc.com/news/health-66221116
Date: 2023-07-17T14:16:16.000Z
LDA Topic: 0.001*"alzheimers" + 0.001*"disease" + 0.001*"brain" + 0.001*"dementia" + 0.001*"drugs" + 0.001
*"research" + 0.001*"family" + 0.001*"drug" + 0.001*"scientists" + 0.001*"says"
Locations:
**

* UK
* Kent
* London
* England
* Wales
* US
Summary:
A new drug, donanemab, has shown promise in slowing cognitive decline in patients with early-stage Alzheimer's
disease. In a global trial involving 1,736 people aged 60 to 85, donanemab demonstrated a significant reduction in
amyloid buildup in the brain, leading to improved cognitive function and daily living abilities. The drug's effect
was more pronounced in patients with less amyloid buildup at the start of the trial. While donanemab has shown
potential, it is not without risks, with brain swelling being a common side effect, and two volunteers, and
possibly a third, died due to dangerous swelling. The trial's results are a significant step forward in the fight
against Alzheimer's, and the UK's drug watchdog is currently assessing donanemab for possible use in the NHS.
Experts emphasize the need for continued research and investment to develop a "statin for the brain" that could
prevent Alzheimer's disease altogether. With around 720,000 people in the UK potentially benefiting from new
Alzheimer's treatments, the Alzheimer's Society highlights the need for better diagnostic infrastructure and
preparedness within the NHS to deliver these treatments effectively.

**

Article 2
```

Note: All the articles have been summarized and locations have been saved along with them

As article 10 was a video it was skipped.

Thought: Now that the articles have been summarized, it's time to extract date, lda and locations separately

Action: Extract date, LDA, and locations from the articles

Action Input: File containing articles

Observation: Successfully extracted date, lda and locations from the articles and saved to a new file

Note: This function can totally be skipped, and we can proceed with the above output for creating a combined summary and visualizations

---

Thought: Now that the information has been extracted, it's time to combine the summaries

Action: Create a final summary

Action Input: File containing article summaries

Observation: Successfully combined the summaries and saved to a new file

```
##  The Global Pursuit of Alzheimer's Solutions: A Spatiotemporal Analysis of Research Efforts

This synthesis examines the global research landscape of Alzheimer's disease, drawing insights from recent and historical research efforts.
The analysis reveals a surge in Alzheimer's research in the UK, particularly in recent years, focusing on both drug development and
understanding the disease's complex mechanisms.

**Key Findings:**

* **UK as a Hub:** Multiple articles highlight the UK's prominent role in Alzheimer's research, with active research centers in London,
Edinburgh, Bristol, and Cardiff.
* **Drug Development:** Significant focus on drug development, with articles discussing recent advances in potential treatments like
donanemab, alongside ongoing efforts to identify new targets and therapies.
* **Early Intervention:**  A growing interest in early intervention, with research exploring potential approaches to slow disease
progression and potentially even prevent Alzheimer's altogether.
* **Longitudinal Research:**  A shift towards understanding the disease's long-term progression, with researchers exploring the role of
environmental factors and genetic predisposition over time.

**Spatiotemporal Trends:**

* **Recent Surge:**  A noticeable increase in research activity in the UK, especially since 2011, with a particular concentration of efforts
in the last 3 years.
* **Global Collaboration:** Research efforts are often collaborative, with institutions in Belgium and the US partnering with UK
counterparts.

**Overall, this analysis suggests a concentrated, multi-faceted effort within the UK to combat Alzheimer's disease. With a focus on drug
development, early intervention, and longitudinal studies, research in the UK and globally is poised to make significant progress in
understanding and treating this devastating disease.**
```

Note: A comprehensive summary which combined the information from all the different articles was created and saved later this summary was passed to check_summary() function to make sure it is what user is asking about

---

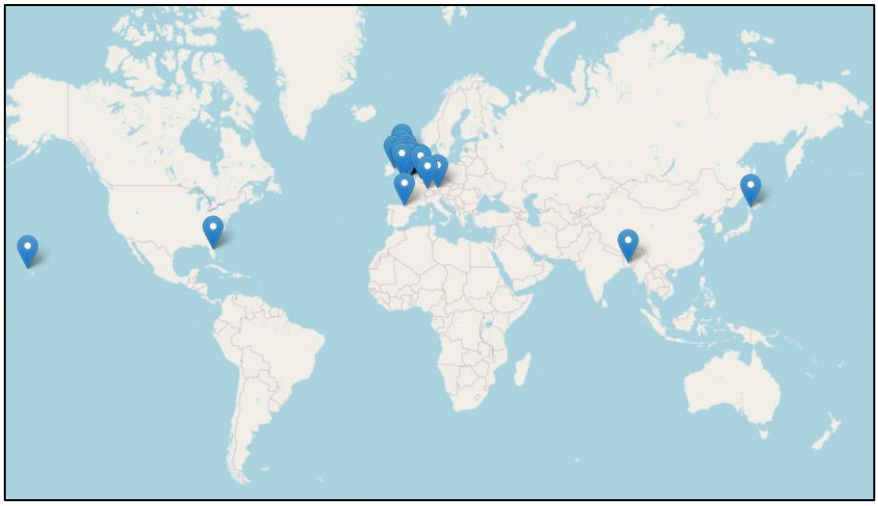
Thought: Now we'll proceed to build a map of the locations

Action: Draw a world map of important locations

Action Input: File containing locations (extracted_info.txt)

Map saved as world_map.html

Observation: Successfully created world map with important locations marked



Note: All the extracted locations have been saved and marked on world map (due to error with API some were skipped)

Again, most of the markers are around UK showcasing that it was covered more by BBC than rest of the world

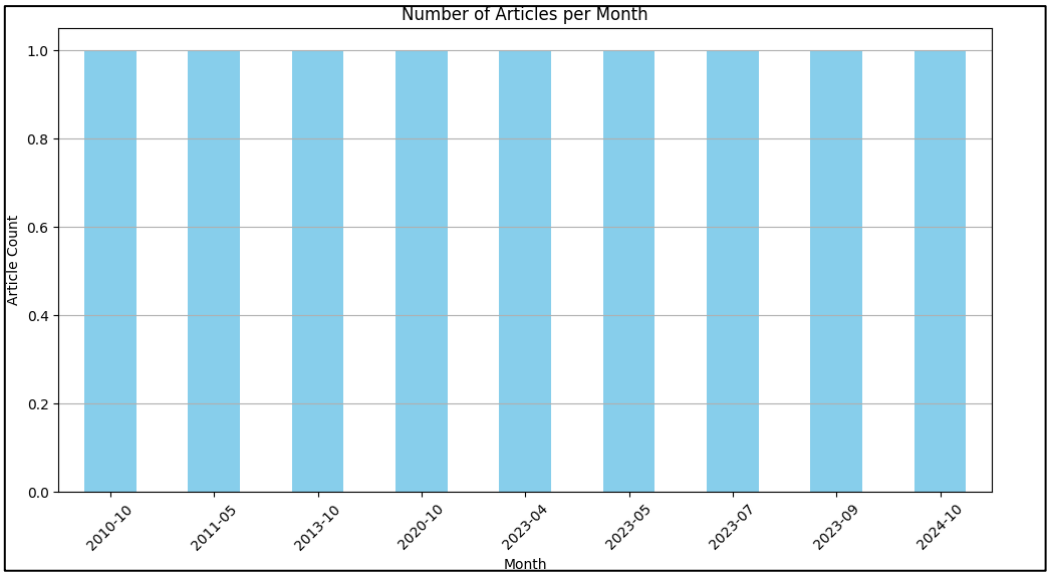
---

Thought: Now we'll proceed to visualize article counts per month

Action: Draw a visualization representing articles per month

Action Input: File containing dates (extracted_info.txt)

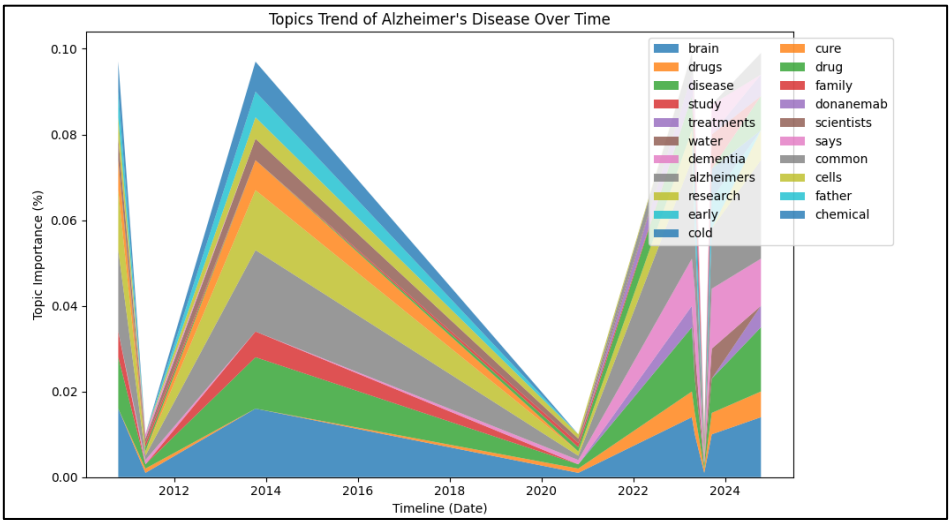
Observation: Successfully created a plot of articles per month



Personal Observations: As I am not considering date ranges for now, older articles, including those from a decade ago, are also being included. The official implementation, on the other hand, uses 227 articles over a one-year period. This makes the plot differ from the official one, but with adjustments in date extraction and limiting the analysis to news within the one-year period, we can achieve similar results to the official implementation.

Thought: Now we'll proceed to visualize article trends

Observation: Successfully created a plot of article trends



Personal Observations: As I've only worked on 10 articles, the LDA topic distribution isn't as sharply defined over the period, but we can still spot some early trends and shifts. There are noticeable spikes in topics related to Alzheimer's research, particularly in 2012 and 2024, which could hint at breakthroughs or increased focus in these areas during those times.

After Plotting all the plots and saving all the information in documents check_summary() function is called which checks our combined_summary.txt using Gemini

Checking the final summary...

Thought: Now I've got the final answer.

FINISHED CHAIN

Correct Response:

The Global Pursuit of Alzheimer's Solutions: A Spatiotemporal Analysis of Research Efforts

This synthesis examines the global research landscape of Alzheimer's disease, drawing insights from recent and historical research efforts. The analysis reveals a surge in Alzheimer's research in the UK, particularly in recent years, focusing on both drug development and understanding the disease's complex mechanisms.

Key Findings:

UK as a Hub: Multiple articles highlight the UK's prominent role in Alzheimer's research, with active research centers in London, Edinburgh, Bristol, and Cardiff.

Drug Development: Significant focus on drug development, with articles discussing recent advances in potential treatments like donanemab, alongside ongoing efforts to identify new targets and therapies.

Early Intervention: A growing interest in early intervention, with research exploring potential approaches to slow disease progression and potentially even prevent Alzheimer's altogether.

Longitudinal Research: A shift towards understanding the disease's long-term progression, with researchers exploring the role of environmental factors and genetic predisposition over time.

Spatiotemporal Trends:

Recent Surge: A noticeable increase in research activity in the UK, especially since 2011, with a particular concentration of efforts in the last 3 years.

Global Collaboration: Research efforts are often collaborative, with institutions in Belgium and the US partnering with UK counterparts.

Overall, this analysis suggests a concentrated, multi-faceted effort within the UK to combat Alzheimer's disease. With a focus on drug development, early intervention, and longitudinal studies, research in the UK and globally is poised to make significant progress in understanding and treating this devastating disease.

Personal Observations: This summary effectively addresses the user's question, "Can you help me to know something new about Alzheimer's disease cure?" by showcasing recent advancements in Alzheimer's research, particularly in drug development and early intervention. It emphasizes ongoing efforts and collaborations, conveying a sense of urgency and progress in the field. To enhance the response, it could include more specific details about innovative treatments or breakthroughs directly related to curing Alzheimer's. This could be achieved by scraping additional research papers and exploring more global sources.

CONCLUSION

In conclusion, AD AUTO-GEMINI effectively addressed the user's question about new developments in Alzheimer's disease cures. By breaking down the query into key terms, I was able to search for relevant articles using the Gemini API, extracting 10 URLs from BBC News and saving the articles for in-depth analysis.

As done in original paper I utilized LDA to identify important topics within the articles and created summaries that highlighted significant areas such as drug development and early intervention. I also gathered essential data like publication dates and locations, culminating in a comprehensive summary of findings. Visualizations, including article counts per month and trend analysis, were created to illustrate the data clearly.

However, I acknowledge a few shortcomings in my work. Using older articles may have affected the overall analysis, and the LDA topic distribution could be clearer with a larger dataset. To improve, I'm focusing on refining the AI's thinking process algorithm and bringing in a wider range of sources, including more recent articles. I also plan to enhance the visualizations to better showcase the data and trends. These adjustments will help strengthen the project and provide a more insightful analysis of Alzheimer's research.

Through these enhancements, I aim to strengthen the project's findings and provide a more accurate and insightful analysis of Alzheimer's research.