*Folders:*

1. **data:** given raw data. There are 2 versions of csv in this folder. The original list of 35k companies and the second version with city, country, and foundation year columns added.
2. **get\_features:** This folder contains the script to get the external features
3. active.ipynb: This notebook takes the list of websites of 35k companies and gets the new domain when a browser opens the websites.
4. crunchbase.ipynb: The data from crunchbase is obtained from manually inputting a list of companies to the Crunchbase search engine. Then I used a web scraper extension [Web Scraper Google Chrome extension](https://chrome.google.com/webstore/detail/web-scraper-free-web-scra/jnhgnonknehpejjnehehllkliplmbmhn) to get the details of the companies (Acquisition Status, Operating Status, and SEMrush Website Monthly Visits). Because Crunchbase only allows a search of <1000 companies at once, there are multiple csv files to complete getting the full set of 5963 companies. This notebook concats and merges these files into crunchbase.csv that has details of all 5963 companies.
5. keyword.ipynb: This notebook uses the library pytrends to obtain the number of weekly Google searches for keywords associated with the companies for the past year (52 weeks). This library also has a rate limit up to 1600 keyword requests per account, so I split the list of 35k companies into smaller sets and run.
6. Tax.ipynb: This notebook accesses the API of the corporate registration website for the state of California and Idaho. It only searches for companies that were founded in the 2 states (CA and ID) and get the details about the latest date of a company’s tax fillings or their tax status.
7. **external\_features:** This folder contains the data scraped for each external feature
8. corporate registration: This folder contains a google sheet that lists all of the states in the US and each state’s corporate registration website. This website contains the name search and the tax status of any registered company (some are free, some are charged).
9. crunchbase: This folder contains the data scraped by a web scraped extension from the Crunchbase website with a pro account.

i. raw: This contains the csv files of details of the set of smaller lists of companies.

ii. formatted: This contains the formatted csv files of details of the companies with the columns organized.

iii. news: This contains the news appearances on each date of the companies list that I input into Crunchbase.

iv. cb\_free\_api.csv: This file has the search results of the list of CompanyName to get the id of each company in the Crunchbase database. These ids can be used to access the more detailed information of each company if we have access to the Enterprise API.

v. crunchbase.csv: This file contains the set of companies that match Crunchbase’s database from the set of 5963 companies. Crunchbase doesn’t have information about the whole set.

vi. news\_crunchbase.csv: This file contains the news appearances of the set of companies that match Crunchbase’s database from the set of 5963 companies.

1. interest\_df: This folder contains the Google Trends search results of the 35k companies using the pytrends library. The weeks might not be matching exactly because I ran the script over the period of 2-3 weeks.

i. interest\_df\_num1\_num2.csv: num1 and num2 is the id of the company at the beginning and the end of the set.

ii. test2.csv: This file just contains the list of the keywords related to 35k companies. I used this as the input for the keyword.ipynb to get the interest\_df.

1. linkedIn: This folder contains the 2 main files from the LinkedIn data that I received. I only unzipped these 2 files and they didn’t match a lot of the founders in the final set of zombie companies.
2. tax: This folder contains the data that I scraped from the tax websites in California and Idaho. The input is the CompanyName of the companies that were founded in California and Idaho in our dataset.

i. tax\_scraping.csv: The tax\_scraping.csv file is the output dataframe with which you can merge with the original dataset using the CompanyName column.

ii. Us\_cities\_states\_counties.csv: I used this file to match the City with the State because the dataset only has a column for City to match it with the tax website.

1. **result:** This folder contains different versions of the predicted set of zombie companies. The zombie\_companies\_v1.csv has the predicted set without using the proxy burn rate. The zombie\_companies\_v2.csv has the final predicted set of zombie companies using all the analysis detailed in the notebook analysis burn rate.ipynb

*Files:*

1. **analysis burn rate.ipynb:** This notebook contains all the analysis including the existing and external features to identify zombie companies. It takes the original raw data and filters the set using the existing features first, then merges the data frame with the external features to do further analysis.
2. **zombie companies.gslides:** This presentation summarizes the approach and the filtering process to the list of zombie companies.

Project Progress Check:

1. Identify Zombie company (inactive companies but still listed as active) ie. rechecking the active/ inactive feature

|  | Features |  | Progress | Data |
| --- | --- | --- | --- | --- |
| 1 | Redirect | If the listed company’s website redirects to a new irrelevant domain | Completed |  |
| 2 | Social | If the company’s Twitter/ Linkedin account has an recent activity (within the last 3 months) | Not Started | Need to get API? |
| 3 | Web Traffic | If the website has no/ minimum traffic within the last 12 months | Completed | Instrumenting with Google Trends search |
| 4 | Founder’s LinkedIn | If all the founders end their position at the companies | Completed |  |
| 5 | Tax Filings/ Status | Government website to check if the corporate registration is still active or if they have any tax violation/ suspension/ forfeiture  US: (different website for each state) <https://www.nass.org/business-services/corporate-registration>  UK: <https://find-and-update.company-information.service.gov.uk/company/07182051/filing-history> | Completed | scraping |