ProfileManager

Table of contents

Introduction

A common theme for companies with a lot of data is that they cannot make use of it. Even though our data is represented many ways (HTML, text, lists of words, vectors, etc.), it is important that it is always traceable back to the source as well as accessible and thus queryable. It needs to be traceable back to the source so we can justify the conclusions we make with tangible evidence rather than simply pointing to the output of a system. Aggregating this information in a queryable format makes it usable for query formulation and building company profiles. The variety of our data, integration needs, and the desire to have random access to source documents (like PDF/HTML) mean that we had to make our own data storage solution. To fulfill these needs, we have made a **Profile Manager** for company profile information and a **Web Resource Manager** for web-based data.

ProfileManager was designed for the aggregation of information related to corporate entities to support building business profiles. It uses the United States Securities and Exchange Commission (SEC) Central Index Key (CIK) to act as universally unique identifiers (UUIDs) and allows the user to compile a variety of information on corporate entities in an easy to use and query format because each profile is a dictionary. Assisting the accessibility of information, **Profile Manager** includes a series of mappings from CIK codes to names and back, names to aliases, and mappings from industry codes (namely The North American Industry Classification System (NAICS) and Standard Industrial Classification (SIC) codes) and descriptions of them. The hope to provide for a flexible data solution for complex business oriented applications.

Documentation

`__init__(rel_path=None)`

Sets the rel_path variable and reads in various mappings including:

- CIK (Central Index Key) to name
- Name to CIK (Central Index Key)

- Name to (a list of) aliases
- NAICS (North American Industry Classification System) to description of classification
- SIC (Standard Industrial Classification) to description of Classification
- NAICS to SIC
- SIC to NAICS

Parameters

- rel_path (string) : the relative path from the script to the parent folder of where your data will be housed. ProfileManager always assumes that data will be held in "data/profilemanager/data" and profiles will be held in "data/profilemanager/profiles" so this allows you to orient your ProfileManager instance to your data source.

Returns

- None

`__contains__(key)`

Returns true if the key is found in the ProfileManager instance. Checks the list of CIK numbers, names, and aliases for matches.

Parameters

- key (string) : can be a CIK (Central Index Key) code, name, or alias

Returns

- None

`__getitem__(key)` and get(key)

Gets profile identified by the key

Parameters

- key (string) : can be a CIK (Central Index Key) code, name, or alias

Returns

- dict : A dictionary which is the profile if found, else None

`__iter__(instances=1, iam = 0)`

A generator that yields the profiles of the contained corporate entities with the ability to be accessed by multiple instances at once.

Parameters

- instances (int) : the number of instances using the iterator (default = 1)
- iam (int) : the current instance's assignment [0-*instances*) (default = 0)

Returns

- dict : A dictionary which is the profile if found, else None

`__len__()`

Returns the number of CIK (Central Index Key) codes in the instance.

Returns

- int : number of CIK (Central Index Key) codes in the instance

`__repr__()` and `__str__()`

Returns a sorted and indented dictionary of CIK (Central Index Key) codes to names of the profiles contained.

Returns

- str : a sorted and indented representation of the CIK (Central Index Key) to names map

build_aliases()

Builds an internal list of aliases from the contained items' names and aliases fields.

Returns

- None

cik_to_alias(cik)

Returns a list of aliases of the business entity identified by the CIK (Central Index Key) code.

Parameters

- cik (string) : is the CIK (Central Index Key) code of a business

Returns

- list of strings : a list of aliases associated with the CIK code

cik_to_description(cik)

Returns a list of descriptions of the industries of the NAICS (North American Industry Classification System) and SIC (Standard Industrial Classification) codes of the business entity identified by the CIK (Central Index Key) code.

Parameters

- cik (string) : is the CIK (Central Index Key) code of a business

Returns

- list of strings : a list of descriptions of business activities associated with the CIK code

cik_to_naics(cik)

Returns the NAICS (North American Industry Classification System) codes of the business entity identified by the CIK (Central Index Key) code.

Parameters

- cik (string) : is the CIK (Central Index Key) code of a business

Returns

- list of strings : the NAICS codes associated with the CIK code

cik_to_name(cik)

Returns the name of the business entity identified by the CIK (Central Index Key) code.

Parameters

- cik (string) : is the CIK (Central Index Key) code of a business

Returns

- string : the name associated with the CIK code

cik_to_sic(cik)

Returns the SIC (Standard Industrial Classification) codes of the business entity identified by the CIK (Central Index Key) code.

Parameters

- cik (string) : is the CIK (Central Index Key) code of a business

Returns

- list of strings : the SIC codes associated with the CIK code

clean_financial_statements()

Removes the text field of all ten_k's, eight_k's, and EX21's which are either the empty string or None.

Returns

- None

generate_profiles()

The code uses the various mappings to aggregate the information into one profile for each business entity.

Currently the code uses two additional JSON files to generate profiles:

- a map from CIK (Central Index Key) to SIC (Standard Industrial Classification)
- a map from SIC (Standard Industrial Classification) to NAICS (North American Industrial

Classification System)

Returns

- None

get_aliases()

A getter function for the the list of aliases

Returns

- list of strings : a list of names and aliases of entities in the instance

get_docs_by_sentence(instances, iam)

An generator function of the sentences of the documents contained with the ability to be accessed by multiple instances at once in a safe way.

Parameters

- instances (int) : the number of instances using the iterator (default = 1)
- iam (int) : the current instance's assignment [0-*instances*) (default = 0)f

Returns

- list of tuples (string, string) : A list tuples representing the sentences of the documents contained and the IDs of the sentences (Yields)

get_resources_by_company(item)

A getter for the documents and associated URLs of the resources by company

Parameters

- item (dict) : a profile

Returns

- list of tuples (string, string) : A list tuples representing the text of the documents contained and the URLs of the documents

get_texts()

A getter for the documents contained

Returns

- list of tuples (string, string) : A list tuples representing the text of the documents contained and the IDs of the documents (Yields)

naics_to_description(naics)

Returns a list of descriptions of the industrial code.

Parameters

- naics (string) : is a NAICS (North American Industry Classification System) code.

Returns

- list of strings : a list of descriptions of the industrial code

naics_to_sic(naics)

Returns the SIC (Standard Industrial Classification) most closely associated with naics.

Parameters

- naics (string) : a NAICS (North American Industry Classification System) code.

Returns

- string : the SIC (Standard Industrial Classification) most closely associated with naics

name_to_aliases(name)

Returns a list of aliases of the business entity.

Parameters

- name (string) : the name of a business entity.

Returns

- list of strings : a list of aliases associated with the business entity

name_to_cik(name)

Returns the CIK (Central Index Key) of the named business entity.

Parameters

- name (string) : the name of a business entity.

Returns

- string : the CIK (Central Index Key) of the named business entity.

name_to_description(name)

Returns a list of descriptions of the industries of the named business entity.

Parameters

- name (string) : the name of a business entity.

Returns

- list of strings : a list of descriptions of the industries of the named business entity

parse_sec_docs(filename)

The code parses the filings which haven't already been parsed, iterating on the CIK codes contained in filename. This means if the CIK codes are disjoint, this method can safely be run in parallel.

Parameters

- filename (string) : the name of a JSON file in "data/profilemanager/data/edgardata/JSON". It must be a dictionary from CIK (Central Index Key) to a dictionary containing the keys "10K", "8K", and "EX21". These keys must map to a list of dictionaries each containing the keys "time_of_filing" and "url".

Returns

- None

parse_wikipedia(parse_list)

Gets the information on the Wikipedia pages for the companies in parse_list

Iterates on the companies contained and searching Wikipedia for the name field, then saves the returned page's parsed table and text in dictionaries ("wiki_table" and "wiki_page" respectively). The table is a dictionary from heading to values and the page is a dictionary from section headings to content.

Parameters

- parse_list (list of dicts) : list of profiles which you would like to get the Wikipedia information for

Returns

- None

save_aliases()

Saves the aliases list to "profilemanager/data/aliases.json" using rel_path

Returns

- None

update_profile(profile)

Writes the current instance of the profile to the JSON

Parameters

- profile (dict) : profiles which you would like update the saved version of

Returns

- None

write_to_raw_text()

Writes all of the contained documents to "profilemanager/raw_text"

Returns

- None