In [ ]:
```
from utils import WebSegmenter,visualize_graph,get_summary
import pandas as pd
import pprint
import seaborn as sns
from datetime import datetime
import pickle
from pprint import pprint
import string
```

Support functions

In [ ]:
```
def get_data_from_graph(graph):
    return list(graph.nodes(data=True))
```

In [ ]:
```
def get_text_from_graph(graph_data):
    txt = ''
    for el in graph_data:
        if 'payload' in el[1]:
            if el[1]['payload']:
                txt += el[1]['payload'].get('text')
                if not txt.endswith('\n'):
                      txt += '\n'

        if 'link_meta' in el[1]:
            if el[1]['link_meta'] is not None:
                if 'text_payload' in el[1]['link_meta'] and el[1]['link_meta'].g
                    for txt_tmp in el[1]['link_meta']['text_payload']:
                        if not txt_tmp.endswith('\n'):
                            txt_tmp += '\n'
                        if not txt.endswith(txt_tmp):
                            txt += txt_tmp
    return txt
```

In [ ]:
```
def replace_end_of_line_and_duplicate_by_whitespace(txt):
    #replace punctuation by whitespace with a regex
    txt = txt.translate(str.maketrans(string.punctuation, ' '*len(string.punctua
    txt = txt.replace('\n',' ')
    txt = txt.replace('\t',' ')
    txt = txt.replace('\r',' ')
    txt = txt.replace('  ',' ')
    txt = txt.replace('  ',' ')
    return txt
```

In [ ]:
```
def lower_case_and_convert_txt_to_tokens_set(txt):
    txt = txt.lower()
    txt = txt.split(' ')
    txt = set(txt)
    return txt
```

In [ ]:
```
def clean_txt(txt):
    return replace_end_of_line_and_duplicate_by_whitespace(txt)
```

In [ ]:
```
def txt_to_set(txt):
    txt = replace_end_of_line_and_duplicate_by_whitespace(txt)
```

```
    txt = lower_case_and_convert_txt_to_tokens_set(txt)
    return txt
```

In [ ]:
```
def jaccard_similarity(txt1, txt2):
    intersection = len(txt1.intersection(txt2))
    union = len(txt1.union(txt2))
    return intersection / union
```

In [ ]:
```
def set_difference(txt1, txt2):
    return txt1.difference(txt2)
```

Load data

On a sample of 1K scraped website,

In [ ]:
```
pdf = pickle.load(open('PE_relevant_sample_2023_05_23_with_analysis.pkl', 'rb'))
pdf = pdf[pdf['graph_size'] > 250]
pdf.head()
```

Out[ ]:

	url	content	
13	http://seamfix.com	\n ; Seamfix – People & Software Development C...	(d07d6b99955b3348ec7ac7ed67d75832e0
14	http://tailwindapp.com	; Tailwind Social Media & Email Marketing To...	(3984cbdd5712a6f4ed35f54bb65151b821
16	http://yoappstore.com	\n ; eCommerce App Development Company | Creat...	(db3d682260e9e7d55b4cd1dd3c8a263827
22	http://www.aldautomotive.co.uk	\n ; \r\n\tCompany Car Leasing & Vehicle Leasi...	(3d98641c2312d3d2b9fae70becade18431
26	http://www.sauter-personal.de	\n ; Die Jobbörse mit aktuellen Jobprofilen, J...	(698f1cae7b3208b2df9856db2d951ddddc

Extract and clean texts

In [ ]:
```
pdf['graph_data'] = pdf['graph'].apply(get_data_from_graph)
pdf['graph_text'] = pdf['graph_data'].apply(get_text_from_graph)
```

```
pdf['graph_text_clean'] = pdf['graph_text'].apply(clean_txt)
pdf['graph_text_clean_set'] = pdf['graph_text_clean'].apply(txt_to_set)
```

In [ ]:
```
pdf['content_text_clean'] = pdf['content'].apply(clean_txt)
pdf['content_text_clean_set'] = pdf['content_text_clean'].apply(txt_to_set)
```

In [ ]:
```
pdf[['content_text_clean','graph_text_clean']].head()
```

Out[ ]:

	content_text_clean	graph_text_clean
13	Seamfix – People Software Development Company...	Solutions Verification Suite Enrolment Suite ...
14	Tailwind Social Media Email Marketing Tool Pl...	Please Wait Connecting your Pinterest account ...
16	eCommerce App Development Company Create mobi...	info yoappstore com 91 74156 64456 Home About ...
22	Company Car Leasing Vehicle Leasing ALD Auto...	Apps ALD websites Algeria Austria Belarus Bel...
26	Die Jobbörse mit aktuellen Jobprofilen Jobs u...	Für Bewerber Jobsuche Portal Initiativbewerbu...

Length difference

In [ ]:
```
pdf['content_text_clean_size'] = pdf['content_text_clean'].apply(len)
pdf['graph_text_clean_size'] = pdf['graph_text_clean'].apply(len)
```

In [ ]:
```
pdf[['content_text_clean_size', 'graph_text_clean_size']].corr(method='pearson')
```

Out[ ]:

	content_text_clean_size	graph_text_clean_size
content_text_clean_size	1.000000	0.974188
graph_text_clean_size	0.974188	1.000000

In [ ]:
```
pdf[['content_text_clean_size', 'graph_text_clean_size']].corr(method='spearman'
```

Out[ ]:

	content_text_clean_size	graph_text_clean_size
content_text_clean_size	1.000000	0.936507
graph_text_clean_size	0.936507	1.000000

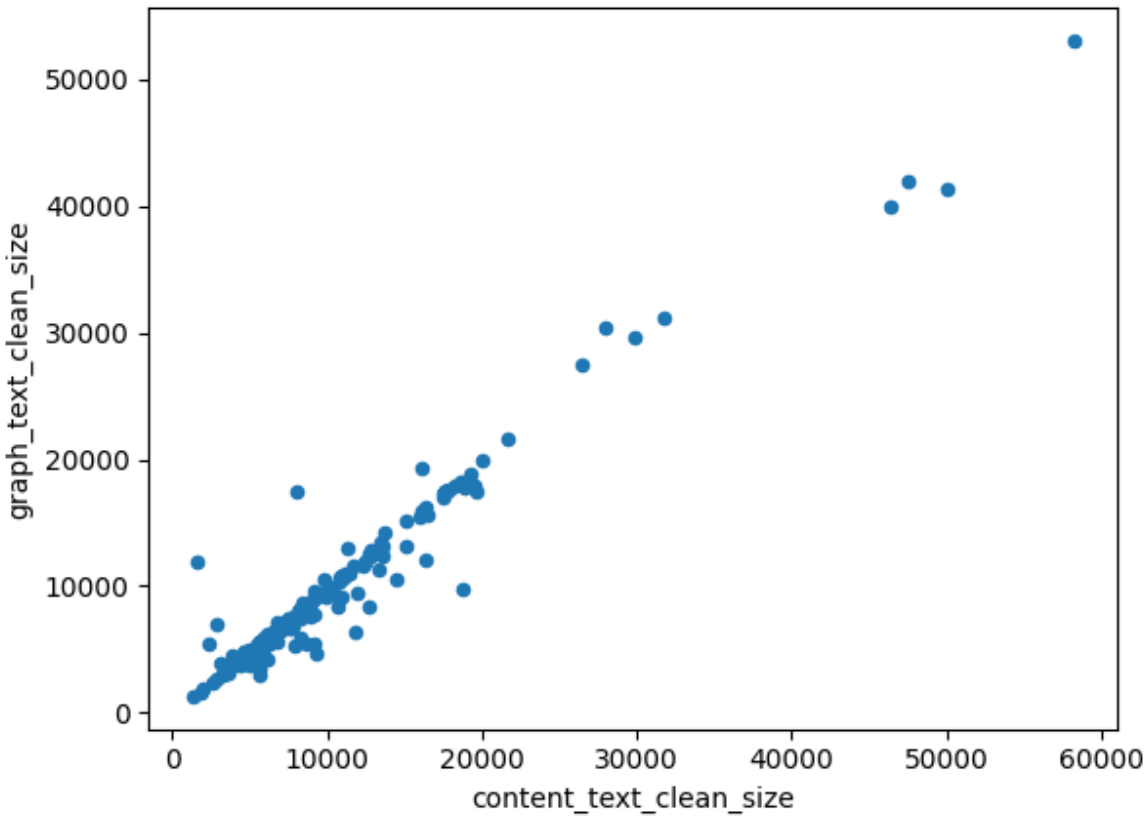
In [ ]:
```
pdf[['content_text_clean_size', 'graph_text_clean_size']].corr(method='kendall')
```

Out[ ]:

	content_text_clean_size	graph_text_clean_size
content_text_clean_size	1.000000	0.849794
graph_text_clean_size	0.849794	1.000000

In [ ]:
```
pdf.plot.scatter(x='content_text_clean_size', y='graph_text_clean_size')
```

Out[ ]: <Axes: xlabel='content_text_clean_size', ylabel='graph_text_clean_size'>



In [ ]:
```
# Display the the rows where the absolute relative ratio of the graph text size
pdf['ratio'] = 1 - pdf['graph_text_clean_size'] / pdf['content_text_clean_size']
pdf['ratio'] = pdf['ratio'].abs()
for r in pdf[pdf['ratio'] > .5].iterrows():
    print(r[0])
    print(r[1]['url'])
    pprint(r[1]['content_text_clean'])
    print('\n----------------\n')
    pprint(r[1]['graph_text_clean'])
    print(r[1]['ratio'])
    print('\n\n\n\n')
```

Token set size difference

In [ ]:
```
pdf['content_text_clean_set_size'] = pdf['content_text_clean_set'].apply(len)
pdf['graph_text_clean_set_size'] = pdf['graph_text_clean_set'].apply(len)
```

In [ ]:
```
pdf[['content_text_clean_set_size', 'graph_text_clean_set_size']].corr(method='p
```

Out[ ]:

	content_text_clean_set_size	graph_text_clean_set_size
content_text_clean_set_size	1.000000	0.986521
graph_text_clean_set_size	0.986521	1.000000

In [ ]:
```
pdf[['content_text_clean_set_size', 'graph_text_clean_set_size']].corr(method='s
```

Out[ ]:

	content_text_clean_set_size	graph_text_clean_set_size
content_text_clean_set_size	1.000000	0.974968
graph_text_clean_set_size	0.974968	1.000000

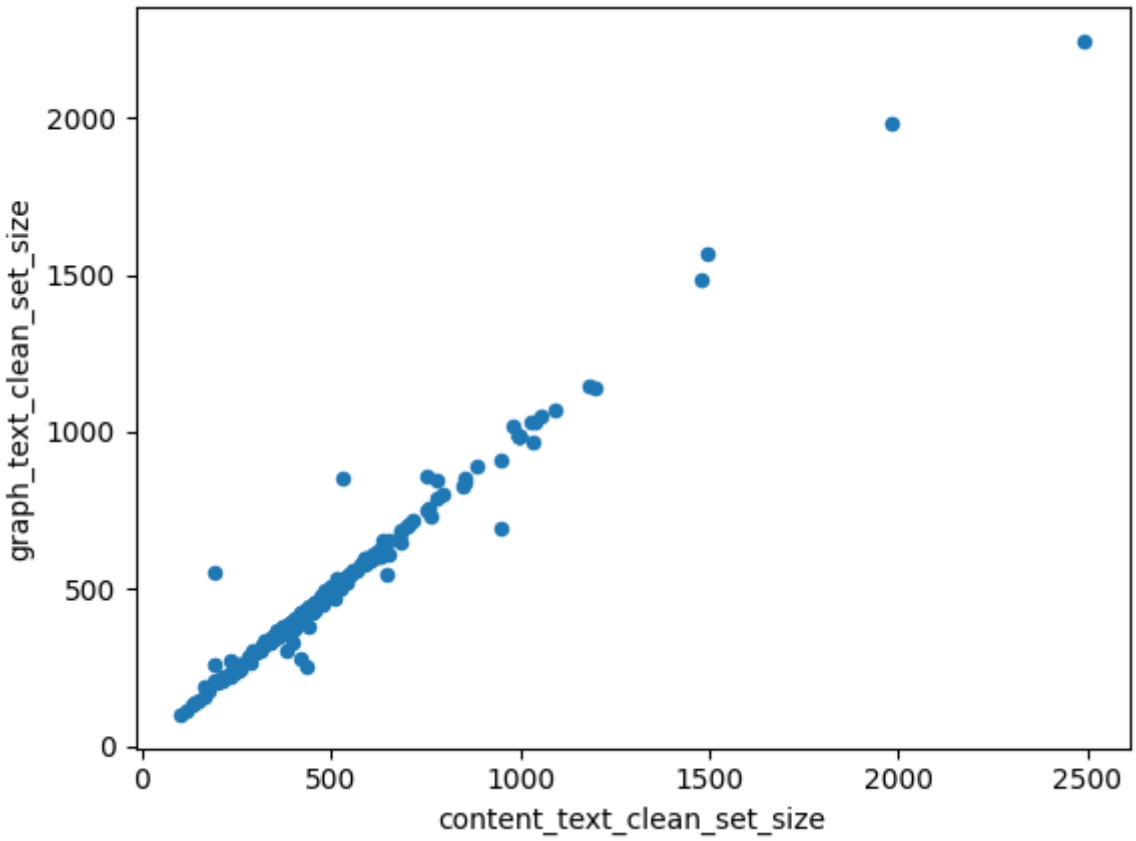
In [ ]:
```
pdf[['content_text_clean_set_size', 'graph_text_clean_set_size']].corr(method='k
```

Out[ ]:

	content_text_clean_set_size	graph_text_clean_set_size
content_text_clean_set_size	1.000000	0.928571
graph_text_clean_set_size	0.928571	1.000000

In [ ]:
```
pdf.plot.scatter(x='content_text_clean_set_size', y='graph_text_clean_set_size')
```

Out[ ]:
```
<Axes: xlabel='content_text_clean_set_size', ylabel='graph_text_clean_set_siz
e'>
```



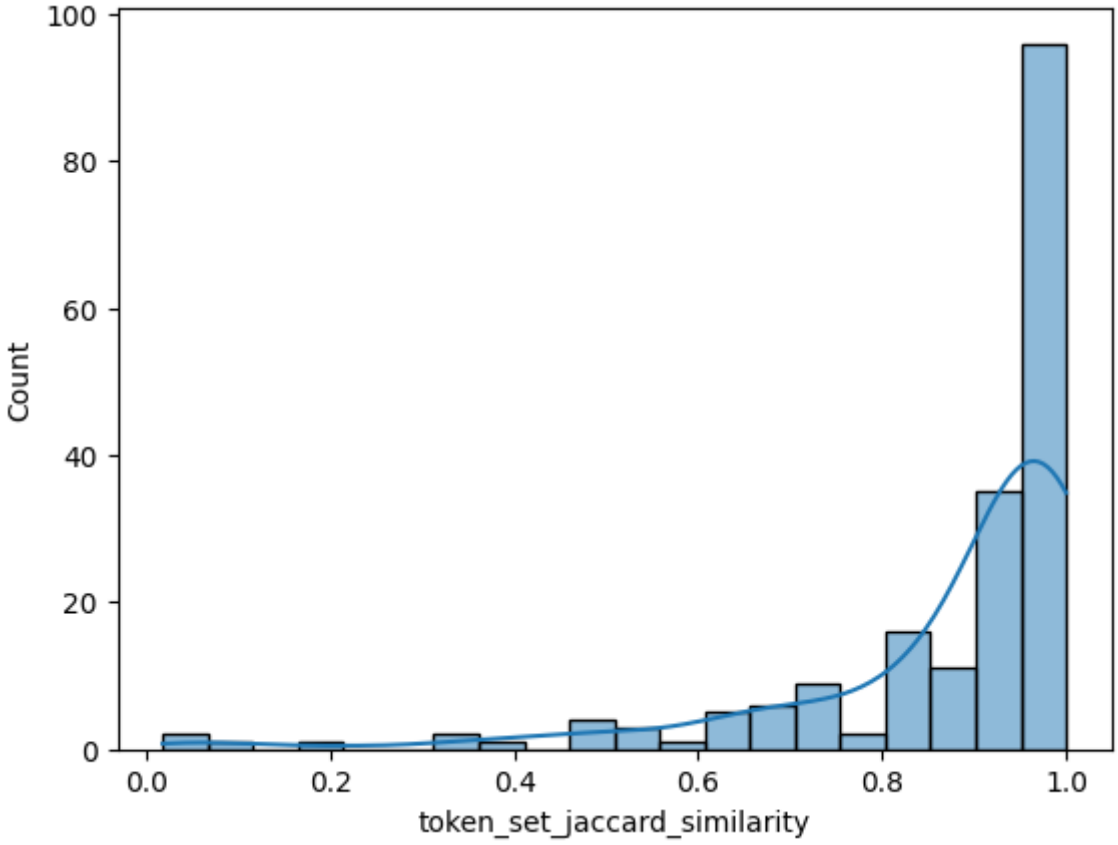
In [ ]:
```
# samples where the graph text set size is greater than the content text set siz
pdf['ratio'] = 1 - pdf['graph_text_clean_set_size'] / pdf['content_text_clean_se
pdf['ratio'] = pdf['ratio'].abs()
for r in pdf[pdf['ratio'] > .5].iterrows():
    print(r[0])
    print(r[1]['url'])
    pprint(r[1]['content_text_clean'])
    print('\n----------------\n')
    pprint(r[1]['graph_text_clean'])
    print(r[1]['ratio'])
    print('\n\n\n\n')
```

Jaccard similarity of tokens set

In [ ]:
```
# Jaccard similarity of tokens set
pdf['token_set_jaccard_similarity'] = pdf.apply(lambda x: jaccard_similarity(x['
```

In [ ]:
```
sns.histplot(pdf['token_set_jaccard_similarity'], kde=True)
```

Out[ ]:
```
<Axes: xlabel='token_set_jaccard_similarity', ylabel='Count'>
```



In [ ]:
```
#sample with a jaccard similarity under 0.8
for r in pdf[pdf['token_set_jaccard_similarity'] < .8].iterrows():
    print(r[0])
    print(r[1]['url'])
    print('\n----------------Content - Graph----------------')
    pprint(set_difference(r[1]['content_text_clean_set'], r[1]['graph_text_clear
    print('\n----------------Graph - Content----------------')
    pprint(set_difference(r[1]['graph_text_clean_set'], r[1]['content_text_clear
    pprint(r[1]['content_text_clean'])
    print('\n----------------\n')
    pprint(r[1]['graph_text_clean'])
    print(r[1]['token_set_jaccard_similarity'])
    print('\n\n\n\n')
```

Experiments

In [ ]:
```
websegmenter = WebSegmenter(url='https://www.patsnap.com/') #https://www.patsnap
websegmenter.run()
```

In [ ]:
```
pdf = pd.read_csv('PE_relevant_sample_2023_05_23.csv')
pdf = pickle.load(open('PE_relevant_sample_2023_05_23_with_analysis.pkl', 'rb'))
pdf.head()

type(pdf['graph'][0])
```

Out[ ]:
```
networkx.classes.digraph.DiGraph
```

In [ ]:
```
graph_data = list(websegmenter.graph.nodes(data=True))
```

In [ ]:
```
graph_data
```

Out[ ]:

```
[('ba99e5176271ffecb5c59c9ba043e67a7103504e',
  {'element_type': 'div',
   'class': "['fixed', 'top-0', 'left-0', 'w-full', 'bg-white', 'z-50']",
   'payload': None,
   'link_meta': None}),
 ('456b89893d2433db83af1e28c0649f070a6ae5cd',
  {'element_type': 'div',
   'name_count': 3,
   'item_index': 1,
   'class': 'mx-auto_max-w-[1280px]_h-full_flex_items-center_justify-between',
   'payload': None,
   'link_meta': None}),
 ('4e355f4434d7490705b37195439383ec78fe9ffd',
  {'element_type': 'ul',
   'name_count': 0,
   'item_index': 2,
   'class': 'flex_items-center_h-full',
   'payload': None,
   'link_meta': None}),
 ('a82752648b9dcbdd1a5ddb3d567f3974c935818a',
  {'element_type': 'a',
   'name_count': 0,
   'item_index': 3,
   'class': 'mr-12',
   'payload': {'href': 'https://www.patsnap.com/', 'text': ''},
   'link_meta': {'href': 'https://www.patsnap.com/', 'text_payload': []}}),
 ('aec7a625020a73d210b52cba1c42363e628632d9',
  {'element_type': 'li',
   'name_count': 0,
   'item_index': 4,
   'class': 'tit-full-hover-div_h-full_each-li_pr-8',
   'payload': None,
   'link_meta': None}),
 ('2c3200b4e245387170081fc8ccff93e90f2b868d',
  {'element_type': 'p',
   'name_count': 1,
   'item_index': 7,
   'class': '',
   'payload': {'href': None, 'text': 'Products & Services'},
   'link_meta': None}),
 ('588416cf5fc902c4d54341e37c52edf30d3feb3a',
  {'element_type': 'ul',
   'name_count': 1,
   'item_index': 7,
   'class': 'mx-auto_max-w-[1280px]_pt-6_pb-9_flex_justify-between',
   'payload': None,
   'link_meta': None}),
 ('02e76daada6e6a4e531a6c5c1a30ad4aa6b90bd9',
  {'element_type': 'li',
   'name_count': 1,
   'item_index': 8,
   'class': 'w-[220px]',
   'payload': None,
   'link_meta': None}),
 ('67cc86cc54c21463b33ae2a5d2f4c1bcf5ff1775',
  {'element_type': 'p',
   'name_count': 3,
   'item_index': 10,
   'class': '',
   'payload': {'href': None, 'text': 'IP Intelligence'},
```

```
  'link_meta': None}),
('c270bebce17e4ddc7f5659af8b81a62a72445b99',
 {'element_type': 'a',
  'name_count': 1,
  'item_index': 11,
  'class': '',
  'payload': {'href': 'https://www.patsnap.com/solutions/ip-intelligence/',
   'text': ''},
  'link_meta': {'href': 'https://www.patsnap.com/solutions/ip-intelligence/',
   'text_payload': ['Patent Analytics']}}),
('36c1f112419e1523aee8a15381ede6a66dabe3ef',
 {'element_type': 'li',
  'name_count': 2,
  'item_index': 9,
  'class': 'w-[220px]',
  'payload': None,
  'link_meta': None}),
('62f28d631d42b17b33a8c845f99f5a1735e58d1f',
 {'element_type': 'p',
  'name_count': 6,
  'item_index': 11,
  'class': '',
  'payload': {'href': None, 'text': 'R&D Intelligence'},
  'link_meta': None}),
('8ba4287d11e62cc72ed20785324fe44a37a277bc',
 {'element_type': 'a',
  'name_count': 2,
  'item_index': 12,
  'class': '',
  'payload': {'href': 'https://www.patsnap.com/solutions/eureka', 'text': ''},
  'link_meta': {'href': 'https://www.patsnap.com/solutions/eureka',
   'text_payload': ['Eureka']}}),
('7c31b9c31a2ae033b1b104a324052fb1798b55a3',
 {'element_type': 'a',
  'name_count': 3,
  'item_index': 13,
  'class': '',
  'payload': {'href': 'https://www.patsnap.com/solutions/discovery/',
   'text': ''},
  'link_meta': {'href': 'https://www.patsnap.com/solutions/discovery/',
   'text_payload': ['Discovery']}}),
('5b3852585187c90e1492b9c8939ab3cc84b39ba7',
 {'element_type': 'li',
  'name_count': 3,
  'item_index': 10,
  'class': 'w-[220px]',
  'payload': None,
  'link_meta': None}),
('bbbc11fb0fb534f5e25751ca0dbd45b6f381b711',
 {'element_type': 'p',
  'name_count': 10,
  'item_index': 12,
  'class': '',
  'payload': {'href': None, 'text': 'Life Sciences Intelligence'},
  'link_meta': None}),
('11e62eb0e9aeadb37369b6a5c38fbcb02046dbb8',
 {'element_type': 'a',
  'name_count': 4,
  'item_index': 13,
  'class': '',
```

```
   'payload': {'href': 'https://www.patsnap.com/solutions/synapse/',
    'text': ''},
   'link_meta': {'href': 'https://www.patsnap.com/solutions/synapse/',
    'text_payload': ['Synapse']}}),
 ('3f470a4d4d19ec2f614c86f9cef61f88cd74412d',
  {'element_type': 'a',
   'name_count': 5,
   'item_index': 14,
   'class': '',
   'payload': {'href': 'https://www.patsnap.com/solutions/bio/', 'text': ''},
   'link_meta': {'href': 'https://www.patsnap.com/solutions/bio/',
    'text_payload': ['Bio']}}),
 ('bf391bfac0f269bcc296d745e65806cfbce46b99',
  {'element_type': 'a',
   'name_count': 6,
   'item_index': 15,
   'class': '',
   'payload': {'href': 'https://www.patsnap.com/solutions/chemical/',
    'text': ''},
   'link_meta': {'href': 'https://www.patsnap.com/solutions/chemical/',
    'text_payload': ['Chemical']}}),
 ('dd88981d2ba7b3104b0d4b938c1ee1a9e1cbbe75',
  {'element_type': 'li',
   'name_count': 4,
   'item_index': 11,
   'class': 'w-[220px]',
   'payload': None,
   'link_meta': None}),
 ('b757390b9809a88529317b6b0806ea29bbcf228e',
  {'element_type': 'p',
   'name_count': 15,
   'item_index': 13,
   'class': '',
   'payload': {'href': None, 'text': 'Business Solutions'},
   'link_meta': None}),
 ('f82b1b17c0c36322c1bc658a857360543e3cd14f',
  {'element_type': 'a',
   'name_count': 7,
   'item_index': 14,
   'class': '',
   'payload': {'href': 'https://www.patsnap.com/solutions/business-solutions/ip
-data-as-a-service/',
    'text': ''},
   'link_meta': {'href': 'https://www.patsnap.com/solutions/business-solutions/
ip-data-as-a-service/',
    'text_payload': ['IP data-as-a-service']}}),
 ('98a510fa2748b1ee1a2f360b066eba3ef5d4ca5c',
  {'element_type': 'a',
   'name_count': 8,
   'item_index': 15,
   'class': '',
   'payload': {'href': 'https://www.patsnap.com/solutions/business-solutions/in
vestor-solutions/',
    'text': ''},
   'link_meta': {'href': 'https://www.patsnap.com/solutions/business-solutions/
investor-solutions/',
    'text_payload': ['Investor solutions']}}),
 ('46de72d5ac7a6a0e3c596be97a85f9793b5764f6',
  {'element_type': 'a',
   'name_count': 9,
```

```
   'item_index': 16,
   'class': '',
   'payload': {'href': 'https://www.patsnap.com/solutions/business-solutions/st
rategy-and-business/',
    'text': ''},
   'link_meta': {'href': 'https://www.patsnap.com/solutions/business-solutions/
strategy-and-business/',
    'text_payload': ['Strategy and business']}}),
 ('22ffabf1a180701b6149bf5b49c9d75cad1a7adc',
  {'element_type': 'li',
   'name_count': 5,
   'item_index': 12,
   'class': 'w-[220px]',
   'payload': None,
   'link_meta': None}),
 ('cae301f2d1a957f9f1df8a211b4d4b8603d34dcb',
  {'element_type': 'p',
   'name_count': 20,
   'item_index': 14,
   'class': '',
   'payload': {'href': None, 'text': 'Other Services'},
   'link_meta': None}),
 ('f4f180ba63b4d839558dad1999fb7514b96da61d',
  {'element_type': 'a',
   'name_count': 10,
   'item_index': 15,
   'class': '',
   'payload': {'href': 'https://www.patsnap.com/solutions/professional-service
s/',
    'text': ''},
   'link_meta': {'href': 'https://www.patsnap.com/solutions/professional-servic
es/',
    'text_payload': ['Professional Services']}}),
 ('efc03fef93beb1a8ae4759314815021e007c1d94',
  {'element_type': 'a',
   'name_count': 11,
   'item_index': 16,
   'class': '',
   'payload': {'href': 'https://www.patsnap.com/solutions/professional-service
s/platform-services/',
    'text': ''},
   'link_meta': {'href': 'https://www.patsnap.com/solutions/professional-servic
es/platform-services/',
    'text_payload': ['Platform Services']}}),
 ('ce9f3a63d0304d53d4cdf01cff7376bda0c8f274',
  {'element_type': 'a',
   'name_count': 12,
   'item_index': 17,
   'class': '',
   'payload': {'href': 'https://www.patsnap.com/solutions/professional-service
s/research-and-analytics-services/',
    'text': ''},
   'link_meta': {'href': 'https://www.patsnap.com/solutions/professional-servic
es/research-and-analytics-services/',
    'text_payload': ['Research Services']}}),
 ('2a4f14aefb4e92b1bf77ae50f081700f3c404e91',
  {'element_type': 'a',
   'name_count': 13,
   'item_index': 18,
   'class': '',
```

```
   'payload': {'href': 'https://www.patsnap.com/solutions/professional-service
s/search-services/',
    'text': ''},
   'link_meta': {'href': 'https://www.patsnap.com/solutions/professional-servic
es/search-services/',
    'text_payload': ['Search Services']}}),
 ('e5fde3c3d720b982a05520f9aa5617ed8f7b6315',
  {'element_type': 'li',
   'name_count': 6,
   'item_index': 5,
   'class': 'tit-hover_h-full_each-li_relative_pr-8',
   'payload': None,
   'link_meta': None}),
 ('98f322795d363a5d858db429a8f0cd2d242aa964',
  {'element_type': 'p',
   'name_count': 26,
   'item_index': 8,
   'class': '',
   'payload': {'href': None, 'text': 'Resources'},
   'link_meta': None}),
 ('378fa50658f34fa4eb95d9bf65dea5f487d190fe',
  {'element_type': 'ul',
   'name_count': 2,
   'item_index': 8,
   'class': 'px-6_py-7_rounded',
   'payload': None,
   'link_meta': None}),
 ('14d9cb70fb5b3dbc5c4f95e35485f048fdfd11d1',
  {'element_type': 'a',
   'name_count': 14,
   'item_index': 10,
   'class': '',
   'payload': {'href': 'https://www.patsnap.com/resources/', 'text': ''},
   'link_meta': {'href': 'https://www.patsnap.com/resources/',
    'text_payload': ['Resources']}}),
 ('088135cbed41a802a3759aa9c63cfa5e0dcc7099',
  {'element_type': 'a',
   'name_count': 15,
   'item_index': 11,
   'class': '',
   'payload': {'href': 'https://www.patsnap.com/customers/', 'text': ''},
   'link_meta': {'href': 'https://www.patsnap.com/customers/',
    'text_payload': ['Success Stories']}}),
 ('1da5b439cfef5148227000b2fa9f2b15a262ec0d',
  {'element_type': 'a',
   'name_count': 16,
   'item_index': 12,
   'class': '',
   'payload': {'href': 'https://www.patsnap.com/cii-newsletter/', 'text': ''},
   'link_meta': {'href': 'https://www.patsnap.com/cii-newsletter/',
    'text_payload': ['Newsletter']}}),
 ('4b61dd49bf615a27a905a3159d365d1bbf18331e',
  {'element_type': 'a',
   'name_count': 17,
   'item_index': 13,
   'class': '',
   'payload': {'href': 'https://academy.patsnap.com/', 'text': ''},
   'link_meta': {'href': 'https://academy.patsnap.com/',
    'text_payload': ['Innovation Academy']}}),
 ('7473f25ce423943698519a07c74feca627edf031',
```

```
 {'element_type': 'a',
  'name_count': 18,
  'item_index': 14,
  'class': '',
  'payload': {'href': 'https://www.patsnap.com/glossary/', 'text': ''},
  'link_meta': {'href': 'https://www.patsnap.com/glossary/',
   'text_payload': ['Glossary']}}),
('b7614f60eeef6a42fa2a64e1a62f00c065bc5788',
 {'element_type': 'a',
  'name_count': 19,
  'item_index': 15,
  'class': '',
  'payload': {'href': 'https://www.patsnap.com/data', 'text': ''},
  'link_meta': {'href': 'https://www.patsnap.com/data',
   'text_payload': ['Data Coverage']}}),
('292c7da51c2827a0cec978c9b25c27cebb25a69c',
 {'element_type': 'li',
  'name_count': 13,
  'item_index': 6,
  'class': 'tit-hover_h-full_each-li_relative_pr-8',
  'payload': None,
  'link_meta': None}),
('438c93f46f42d81d2b9a6ace7d9d7a982f7d2aaa',
 {'element_type': 'p',
  'name_count': 34,
  'item_index': 9,
  'class': '',
  'payload': {'href': None, 'text': 'Why Patsnap'},
  'link_meta': None}),
('22f77dd0a51a9b817b03a5171cafb666a8eefa31',
 {'element_type': 'ul',
  'name_count': 3,
  'item_index': 9,
  'class': 'px-6_py-7_rounded',
  'payload': None,
  'link_meta': None}),
('c7a4cea8e7c6c2c51338e543bca7f867b7503a50',
 {'element_type': 'a',
  'name_count': 20,
  'item_index': 11,
  'class': '',
  'payload': {'href': 'https://www.patsnap.com/why-patsnap/', 'text': ''},
  'link_meta': {'href': 'https://www.patsnap.com/why-patsnap/',
   'text_payload': ['Why Patsnap']}}),
('9c423475a98223c9c6018e477645f816cd9f9b67',
 {'element_type': 'a',
  'name_count': 21,
  'item_index': 12,
  'class': '',
  'payload': {'href': 'https://www.patsnap.com/why-patsnap/about-us/',
   'text': ''},
  'link_meta': {'href': 'https://www.patsnap.com/why-patsnap/about-us/',
   'text_payload': ['About us']}}),
('fbeb8d41cd0011939dd03f84b65d333279de78fe',
 {'element_type': 'a',
  'name_count': 22,
  'item_index': 13,
  'class': '',
  'payload': {'href': 'https://www.patsnap.com/contact-us/', 'text': ''},
  'link_meta': {'href': 'https://www.patsnap.com/contact-us/',
```

```
    'text_payload': ['Contact us']}}),
 ('5f75d193bcba92011bfeca3aebe69a745670ad2b',
  {'element_type': 'a',
   'name_count': 23,
   'item_index': 14,
   'class': '',
   'payload': {'href': 'https://www.patsnap.com/careers/', 'text': ''},
   'link_meta': {'href': 'https://www.patsnap.com/careers/',
    'text_payload': ['Careers']}}),
 ('1081155e839f2eaf809b2d8502b4097f65f187ec',
  {'element_type': 'a',
   'name_count': 24,
   'item_index': 15,
   'class': '',
   'payload': {'href': 'https://help.patsnap.com/hc/en-us', 'text': ''},
   'link_meta': {'href': 'https://help.patsnap.com/hc/en-us',
    'text_payload': ['Help center']}}),
 ('60367dd9e4acf9453464e9b730e94c7106e52b87',
  {'element_type': 'ul',
   'name_count': 4,
   'item_index': 3,
   'class': 'flex_items-center_h-full',
   'payload': None,
   'link_meta': None}),
 ('46329983f5523c69dd33a513ab520d7e8bdb7a46',
  {'element_type': 'a',
   'name_count': 25,
   'item_index': 4,
   'class': 'cursor-pointer_h-10_leading-10_w-[169px]_bg-[#0764E9]_rounded_text
-center_text-white',
   'payload': {'href': 'https://www.patsnap.com/request-a-demo/', 'text': ''},
   'link_meta': {'href': 'https://www.patsnap.com/request-a-demo/',
    'text_payload': ['Request a Demo']}}),
 ('0aa3e351bb52c96952f853470e47ce33e68431ff',
  {'element_type': 'div',
   'name_count': 25,
   'item_index': 8,
   'class': 'flex_flex-col',
   'payload': None,
   'link_meta': None}),
 ('4ce1ab26b436ffff25ba918ed43aee8cea4576b8',
  {'element_type': 'a',
   'name_count': 26,
   'item_index': 9,
   'class': 'h-8_px-2_leading-8_hover:bg-[#B3BAC5]_hover:bg-opacity-20',
   'payload': {'href': 'https://www.zhihuiya.com/', 'text': ''},
   'link_meta': {'href': 'https://www.zhihuiya.com/',
    'text_payload': ['简体中文']}}),
 ('798995d2f0a9b75f1e270c5ffa4df67074f6667e',
  {'element_type': 'a',
   'name_count': 27,
   'item_index': 10,
   'class': 'px-2_rounded_flex_justify-between_h-8_leading-8_text-[#0764E9]',
   'payload': {'href': 'javascript:void(0)', 'text': 'English'},
   'link_meta': {'href': 'javascript:void(0)', 'text_payload': ['English']}}),
 ('a6427bb44378065ae9eed7be0c1bda74eccbce2b',
  {'element_type': 'a',
   'name_count': 28,
   'item_index': 11,
   'class': 'h-8_px-2_leading-8_hover:bg-[#B3BAC5]_hover:bg-opacity-20',
```

```
  'payload': {'href': 'javascript:void(0)', 'text': ''},
  'link_meta': {'href': 'javascript:void(0)', 'text_payload': ['日本語']}}),
('d89376a19f7bc8e513dbe6d3c05a79a9f9b80187',
 {'element_type': 'a',
  'name_count': 29,
  'item_index': 12,
  'class': 'h-8_px-2_leading-8_hover:bg-[#B3BAC5]_hover:bg-opacity-20',
  'payload': {'href': 'javascript:void(0)', 'text': ''},
  'link_meta': {'href': 'javascript:void(0)', 'text_payload': ['한국어']}}),
('fafd9059ea4ac36f051d5d8fb8816e109a194e9c',
 {'element_type': 'a',
  'name_count': 30,
  'item_index': 7,
  'class': 'hover:text-[#0764E9]',
  'payload': {'href': 'https://www.patsnap.com/login/', 'text': ''},
  'link_meta': {'href': 'https://www.patsnap.com/login/',
   'text_payload': ['Login']}})]
```

In [ ]:
```
websegmenter.search('Benefits')
```

Out[ ]:
```
[('86f3108012c47c5730cf14e9e72d725c27da507a',
  {'label': 'Benefits',
   'score': 17,
   'to_filter': False,
   'item_class': 'has_grid'})]
```

In [ ]:
```
node_id = '86f3108012c47c5730cf14e9e72d725c27da507a'
node_summary = websegmenter.summarize(node_id=node_id)
pprint(node_summary,sort_dicts=False)
```

```
{'id': '86f3108012c47c5730cf14e9e72d725c27da507a',
 'content': [{'type': 'grid',
              'payload': [{'type': 'grid_item',
                           'payload': [{'type': 'atom',
                                        'payload': 'We automate all major '
                                                   'steps in the ML lifecycle '
                                                   'from raw data ingestion to '
                                                   'sustained production '
                                                   'deployment'},
                                       {'type': 'atom',
                                        'payload': 'End to end predictive '
                                                   'analytics automation'}],
                           'index': 19},
                          {'type': 'grid_item',
                           'payload': [{'type': 'atom',
                                        'payload': 'Query the future in our '
                                                   '‘Predictive Querying’ '
                                                   'language that is as easy '
                                                   'to use as SQL'},
                                       {'type': 'atom',
                                        'payload': 'No ML experience '
                                                   'required'}],
                           'index': 20},
                          {'type': 'grid_item',
                           'payload': [{'type': 'atom',
                                        'payload': 'Cutting edge AI'},
                                       {'type': 'atom',
                                        'payload': 'Leverage automated machine '
                                                   'learning with state of the '
                                                   'art'},
                                       {'type': 'atom',
                                        'payload': 'to drive higher accuracy '
                                                   'even with less data'}],
                           'index': 21},
                          {'type': 'grid_item',
                           'payload': [{'type': 'atom',
                                        'payload': 'Deliver more predictions '
                                                   'more quickly across every '
                                                   'team, enabling your entire '
                                                   'enterprise to more '
                                                   'proactively choose the '
                                                   'future you want'},
                                       {'type': 'atom',
                                        'payload': 'Transform your '
                                                   'decision-making'}],
                           'index': 22}],
              'index': 13},
             {'type': 'atom', 'payload': 'Benefits', 'index': 18}],
 'links': [{'href': 'capabilities#inner-section-7',
            'text_payload': ['Graph Neural Network technology']}]}
```

In [ ]:

```python
def get_text_from_graph(graph_data):
    txt = ''
    for el in graph_data:
        if 'payload' in el[1]:
            if el[1]['payload']:
                txt += el[1]['payload'].get('text')
                if not txt.endswith('\n'):
                      txt += '\n'
```

```
        if 'link_meta' in el[1]:
            if el[1]['link_meta'] is not None:
                if 'text_payload' in el[1]['link_meta'] and el[1]['link_meta'].g
                    for txt_tmp in el[1]['link_meta']['text_payload']:
                        if not txt_tmp.endswith('\n'):
                            txt_tmp += '\n'
                        if not txt.endswith(txt_tmp):
                            txt += txt_tmp
    return txt
```

In [ ]:
```
txt = ''
for el in graph_data:

    # if 'text' in el[1]:
    #     if el[1]['text']:
    #         txt += el[1]['text']
    #         txt += '\n'

    if 'payload' in el[1]:
        if el[1]['payload']:
            txt += el[1]['payload'].get('text')
            txt += '\n'

    if 'link_meta' in el[1]:
        if el[1]['link_meta'] is not None:
            if 'text_payload' in el[1]['link_meta'] and el[1]['link_meta'].get('
                txt_tmp = ' '.join(el[1]['link_meta']['text_payload'])
                txt_tmp += '\n'
                if not txt.endswith(txt_tmp):
                    txt += txt_tmp
```

In [ ]:
```
print(get_text_from_graph(graph_data))
```

In [ ]:
```
pdf[1].apply(lambda x: x['class'])
```

In [ ]:
```

```