Author Portrait Generation

Vedanth S(2018103620)
Niranjan K(2018103569)
Praveen RS(2018103577)
Akash E(2018103005)

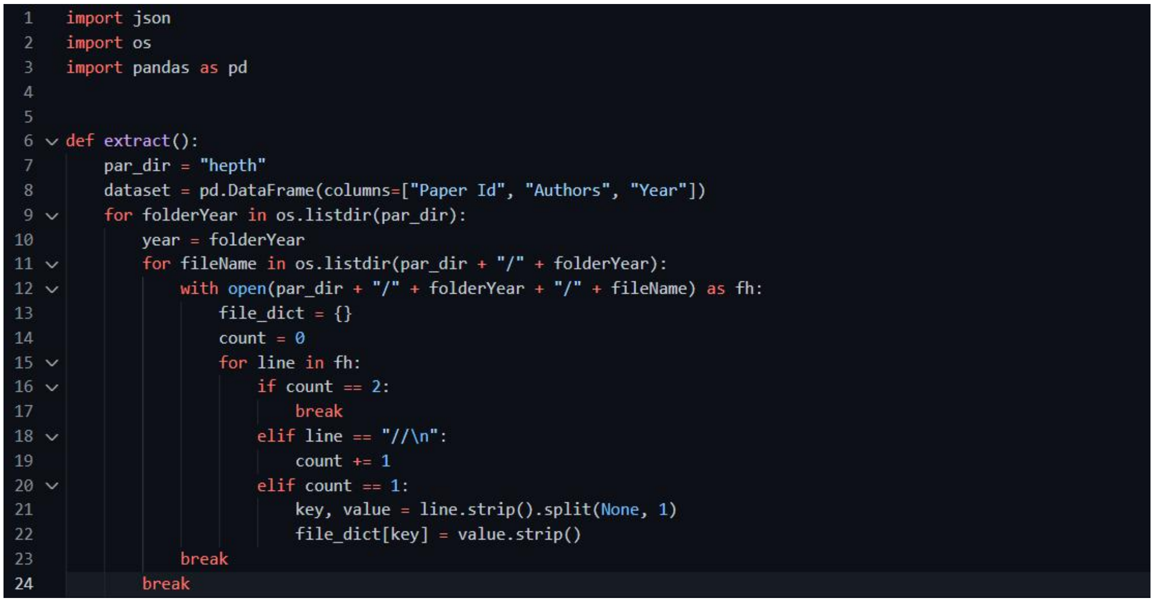
Data Mining - NLP Project
Author Portrait Generation using Topic Modelling

1. Objective :

The objective of this project is to generate a portrait for authors who have published papers in the hep-th domain from 1991 to 2003.

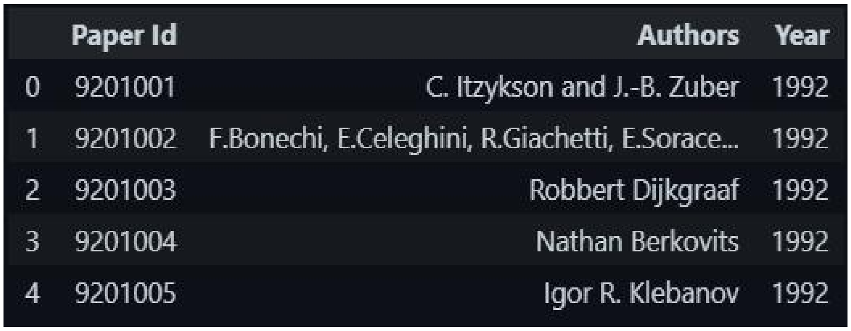
2. File Extraction :

The arxiv dataset we extracted consists of metadata files for published papers in the range 1991 - 2003. The papers are segregated into separate folders based on the year published. An ‘.abs’ file for every paper consists of metadata for the paper. Various details like the title of the paper, authors, year published and abstract are present in the file. The authors need to be extracted from each file. The file is parsed line by line and each line is inserted into a list. The line with the substring ‘Author’ is extracted from the list and the author list is extracted. The subsequent line is also checked for the presence of any author names. The ‘Paper Id’ is the file name and the published year is the parent folder’s name. Finally, a dataframe with the columns Paper Id, Authors, Year is created as shown in the output.



```python
import json
import os
import pandas as pd


def extract():
    par_dir = "hepth"
    dataset = pd.DataFrame(columns=["Paper Id", "Authors", "Year"])
    for folderYear in os.listdir(par_dir):
        year = folderYear
        for fileName in os.listdir(par_dir + "/" + folderYear):
            with open(par_dir + "/" + folderYear + "/" + fileName) as fh:
                file_dict = {}
                count = 0
                for line in fh:
                    if count == 2:
                        break
                    elif line == "//\n":
                        count += 1
                    elif count == 1:
                        key, value = line.strip().split(None, 1)
                        file_dict[key] = value.strip()
            break
        break
```

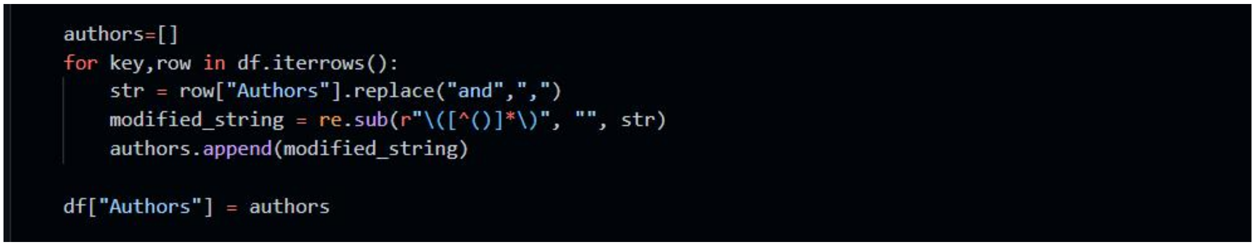
Output :



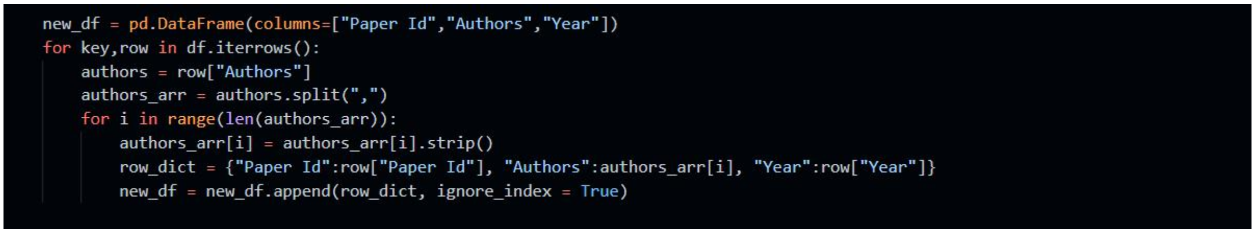
	Paper Id	Authors	Year
0	9201001	C. Itzykson and J.-B. Zuber	1992
1	9201002	F.Bonechi, E.Celeghini, R.Giachetti, E.Sorace...	1992
2	9201003	Robbert Dijkgraaf	1992
3	9201004	Nathan Berkovits	1992
4	9201005	Igor R. Klebanov	1992

3. Data Preprocessing :

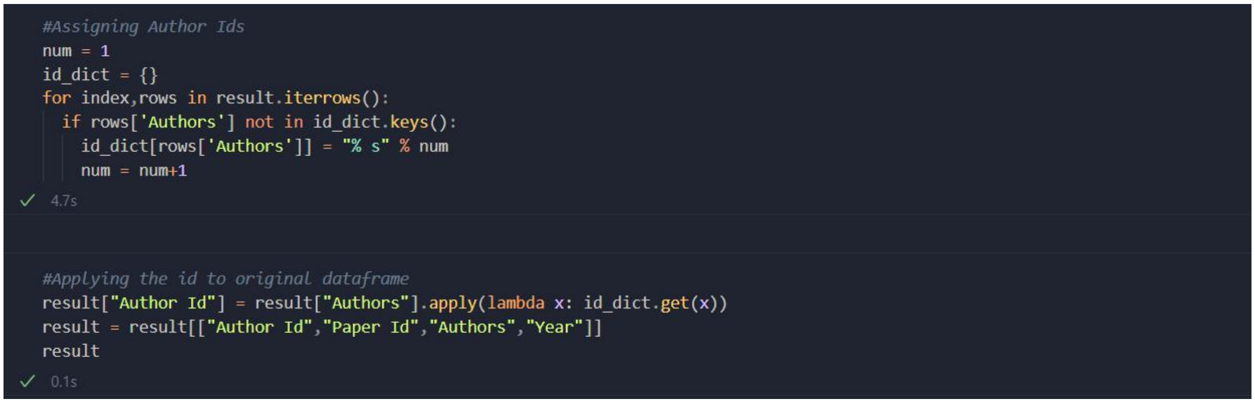
The authors parsed and extracted from the files aren't in a consistent format . Some records have authors separated by `and` whereas some are separated by 'comma'. So, we replace all `and` with 'commas'.Some records have redundant brackets which are replaced with empty strings. All the records in the dataframe are uniquely identified by 'Paper Id' and they contain authors separated by 'comma'. In order to assign a unique id for every author we need records wherein there is a single author for every paper by repeating the entry for each author in the list. Finally we create a dataframe where Paper Id, Authors, Year are present with the Author field containing a single name.



```
authors=[]
for key,row in df.iterrows():
    str = row["Authors"].replace("and",",")
    modified_string = re.sub(r"\([^()]*\)", "", str)
    authors.append(modified_string)

df["Authors"] = authors
```



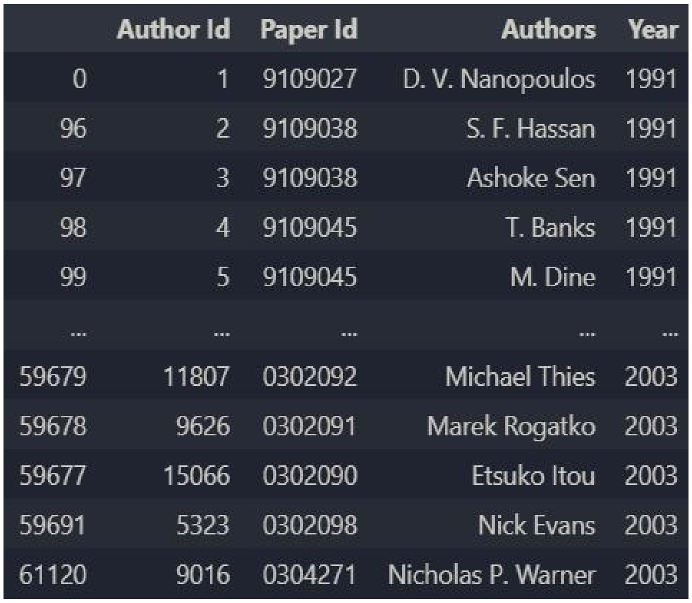
```
new_df = pd.DataFrame(columns=["Paper Id","Authors","Year"])
for key,row in df.iterrows():
    authors = row["Authors"]
    authors_arr = authors.split(",")
    for i in range(len(authors_arr)):
        authors_arr[i] = authors_arr[i].strip()
        row_dict = {"Paper Id":row["Paper Id"], "Authors":authors_arr[i], "Year":row["Year"]}
        new_df = new_df.append(row_dict, ignore_index = True)
```



```
#Assigning Author Ids
num = 1
id_dict = {}
for index,rows in result.iterrows():
  if rows['Authors'] not in id_dict.keys():
    id_dict[rows['Authors']] = "% s" % num
    num = num+1
```

```
#Applying the id to original dataframe
result["Author Id"] = result["Authors"].apply(lambda x: id_dict.get(x))
result = result[["Author Id","Paper Id","Authors","Year"]]
result
```

Output :

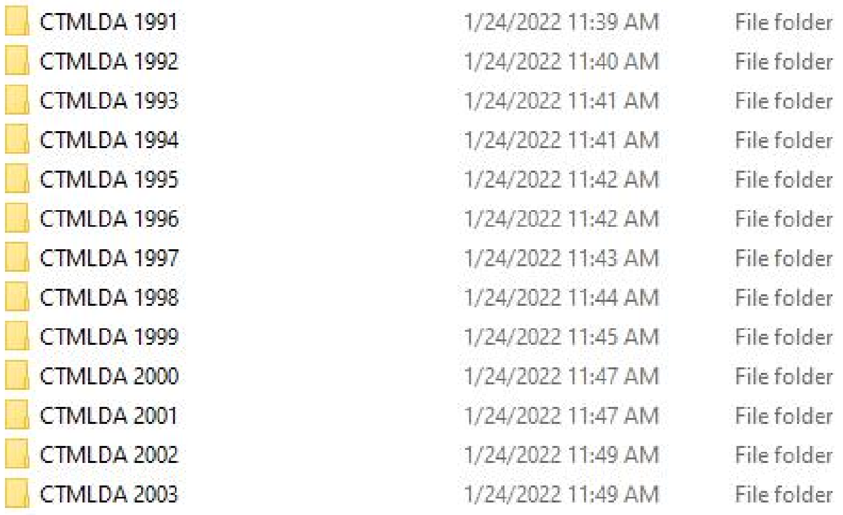


	Author Id	Paper Id	Authors	Year
0	1	9109027	D. V. Nanopoulos	1991
96	2	9109038	S. F. Hassan	1991
97	3	9109038	Ashoke Sen	1991
98	4	9109045	T. Banks	1991
99	5	9109045	M. Dine	1991
...	...	...	...	...
59679	11807	0302092	Michael Thies	2003
59678	9626	0302091	Marek Rogatko	2003
59677	15066	0302090	Etsuko Itou	2003
59691	5323	0302098	Nick Evans	2003
61120	9016	0304271	Nicholas P. Warner	2003

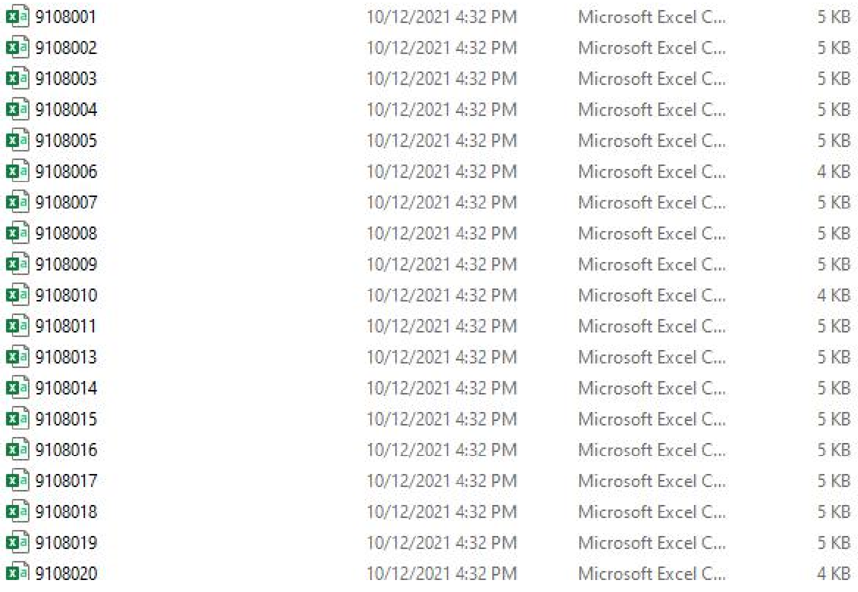
4. Topic Modeling for Articles :

Topic modeling is an unsupervised machine learning technique that's capable of scanning a set of documents, detecting word and phrase patterns within them, and automatically clustering word groups and similar expressions/patterns that best characterize a set of documents.

By detecting patterns such as word frequency and distance between words, a topic model clusters feedback that is similar, and words and expressions that appear most often. With this information, you can quickly deduce what each set of texts are talking about.

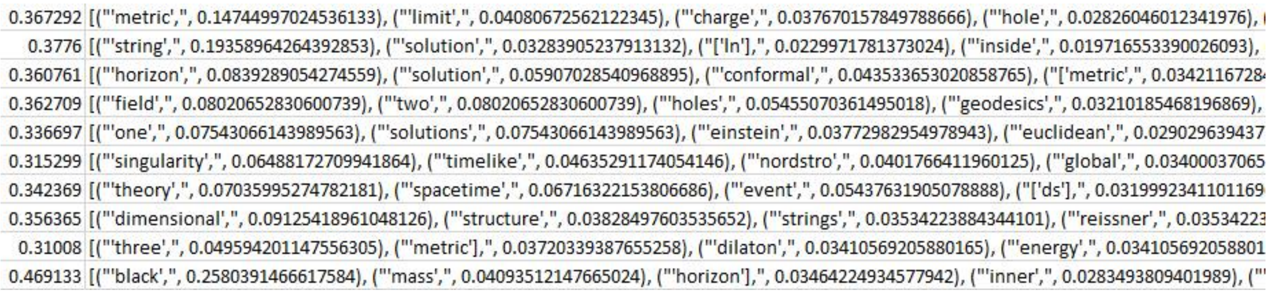


CTMLDA 1991	1/24/2022 11:39 AM	File folder
CTMLDA 1992	1/24/2022 11:40 AM	File folder
CTMLDA 1993	1/24/2022 11:41 AM	File folder
CTMLDA 1994	1/24/2022 11:41 AM	File folder
CTMLDA 1995	1/24/2022 11:42 AM	File folder
CTMLDA 1996	1/24/2022 11:42 AM	File folder
CTMLDA 1997	1/24/2022 11:43 AM	File folder
CTMLDA 1998	1/24/2022 11:44 AM	File folder
CTMLDA 1999	1/24/2022 11:45 AM	File folder
CTMLDA 2000	1/24/2022 11:47 AM	File folder
CTMLDA 2001	1/24/2022 11:47 AM	File folder
CTMLDA 2002	1/24/2022 11:49 AM	File folder
CTMLDA 2003	1/24/2022 11:49 AM	File folder



9108001	10/12/2021 4:32 PM	Microsoft Excel C...	5 KB
9108002	10/12/2021 4:32 PM	Microsoft Excel C...	5 KB
9108003	10/12/2021 4:32 PM	Microsoft Excel C...	5 KB
9108004	10/12/2021 4:32 PM	Microsoft Excel C...	5 KB
9108005	10/12/2021 4:32 PM	Microsoft Excel C...	5 KB
9108006	10/12/2021 4:32 PM	Microsoft Excel C...	4 KB
9108007	10/12/2021 4:32 PM	Microsoft Excel C...	5 KB
9108008	10/12/2021 4:32 PM	Microsoft Excel C...	5 KB
9108009	10/12/2021 4:32 PM	Microsoft Excel C...	5 KB
9108010	10/12/2021 4:32 PM	Microsoft Excel C...	4 KB
9108011	10/12/2021 4:32 PM	Microsoft Excel C...	5 KB
9108013	10/12/2021 4:32 PM	Microsoft Excel C...	5 KB
9108014	10/12/2021 4:32 PM	Microsoft Excel C...	5 KB
9108015	10/12/2021 4:32 PM	Microsoft Excel C...	5 KB
9108016	10/12/2021 4:32 PM	Microsoft Excel C...	5 KB
9108017	10/12/2021 4:32 PM	Microsoft Excel C...	5 KB
9108018	10/12/2021 4:32 PM	Microsoft Excel C...	5 KB
9108019	10/12/2021 4:32 PM	Microsoft Excel C...	5 KB
9108020	10/12/2021 4:32 PM	Microsoft Excel C...	4 KB

This is the corresponding topic modeled words for a paper id. There are 10 groups in total for each paper and each word in a group is associated with a certain probability indicating the importance of that word in that corresponding document.



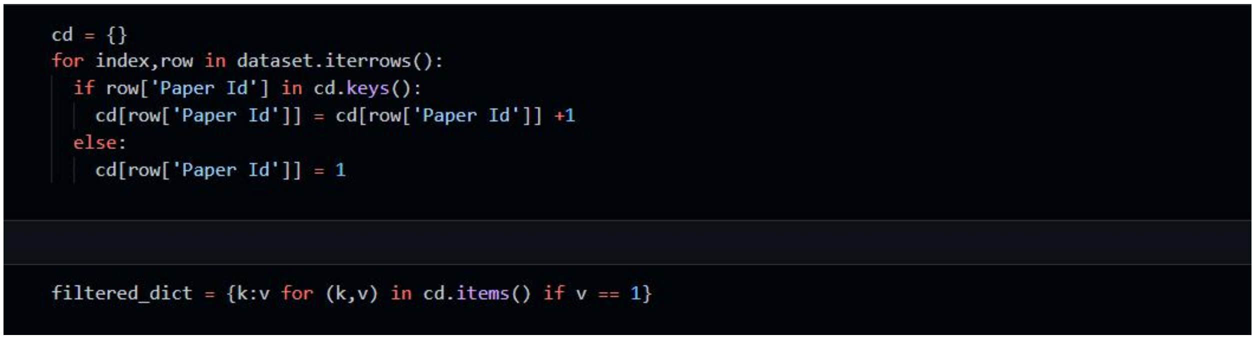
0.367292	[("'metric',", 0.14744997024536133), ("'limit',", 0.04080672562122345), ("'charge',", 0.037670157849788666), ("'hole',", 0.02826046012341976),
0.3776	[("'string',", 0.19358964264392853), ("'solution',", 0.03283905237913132), ("['ln'],", 0.0229971781373024), ("'inside',", 0.019716553390026093),
0.360761	[("'horizon',", 0.0839289054274559), ("'solution',", 0.05907028540968895), ("'conformal',", 0.043533653020858765), ("['metric',", 0.0342116728
0.362709	[("'field',", 0.08020652830600739), ("'two',", 0.08020652830600739), ("'holes',", 0.05455070361495018), ("'geodesics',", 0.03210185468196869),
0.336697	[("'one',", 0.07543066143989563), ("'solutions',", 0.07543066143989563), ("'einstein',", 0.03772982954978943), ("'euclidean',", 0.029029639437
0.315299	[("'singularity',", 0.06488172709941864), ("'timelike',", 0.04635291174054146), ("'nordstro',", 0.0401766411960125), ("'global',", 0.03400037065
0.342369	[("'theory',", 0.07035995274782181), ("'spacetime',", 0.06716322153806686), ("'event',", 0.05437631905078888), ("['ds'],", 0.0319992341101169
0.356365	[("'dimensional',", 0.09125418961048126), ("'structure',", 0.03828497603535652), ("'strings',", 0.03534223884344101), ("'reissner',", 0.03534223
0.31008	[("'three',", 0.049594201147556305), ("'metric'],", 0.03720339387655258), ("'dilaton',", 0.03410569205880165), ("'energy',", 0.034105692058801
0.469133	[("'black',", 0.2580391466617584), ("'mass',", 0.04093512147665024), ("'horizon'],", 0.03464224934577942), ("'inner',", 0.0283493809401989), ("

5. Single Author Portrait Generation :

Let us consider an author R_0 who has authored ‘n’ single authored papers. We have to compute the single author history of R_0 which is also known as Author Topic Model [ATM]. It is represented by an ATM[R_0]. . This is obtained by aggregating the Document Topic Model (DTM) of every single author publication of R_0. DTM is obtained by generating a topic model for the corresponding research article. Therefore, if R_0 has authored ‘n’ research articles as sole author, then,

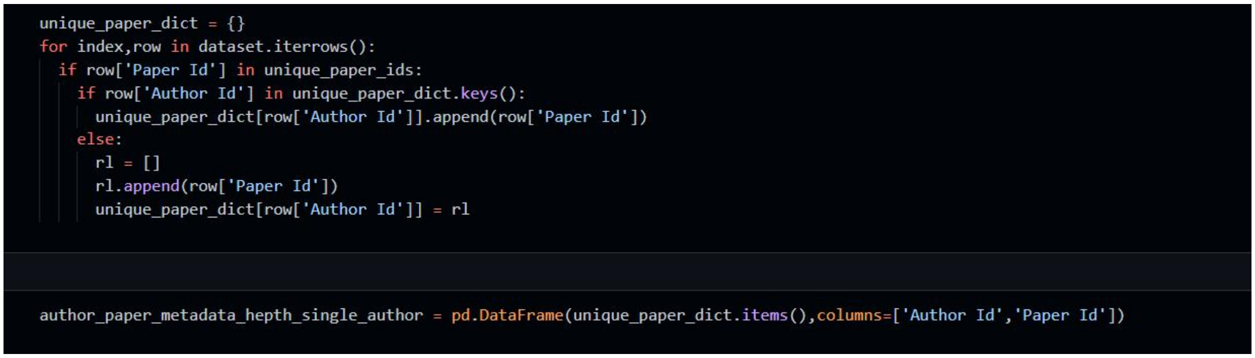
$$ATM(R_0) = DTM(S_0) + DTM(S_1) + ... + DTM(S_{n-1})$$

Aggregation of topics of sole author publications means, taking the maximum of the respective probabilities. Let us consider, an author has published articles on the topic ‘deep learning’. When this topic appearing in one article with probability 0.45 is aggregated with the occurrences of ‘deep learning’ from other two articles written by the same author with topic probability 0.7, 0.89 respectively, the aggregated topic probability of ‘deep learning’ is 0.89. This indicates that the author has a proficiency of 89% related to ‘deep learning’ .



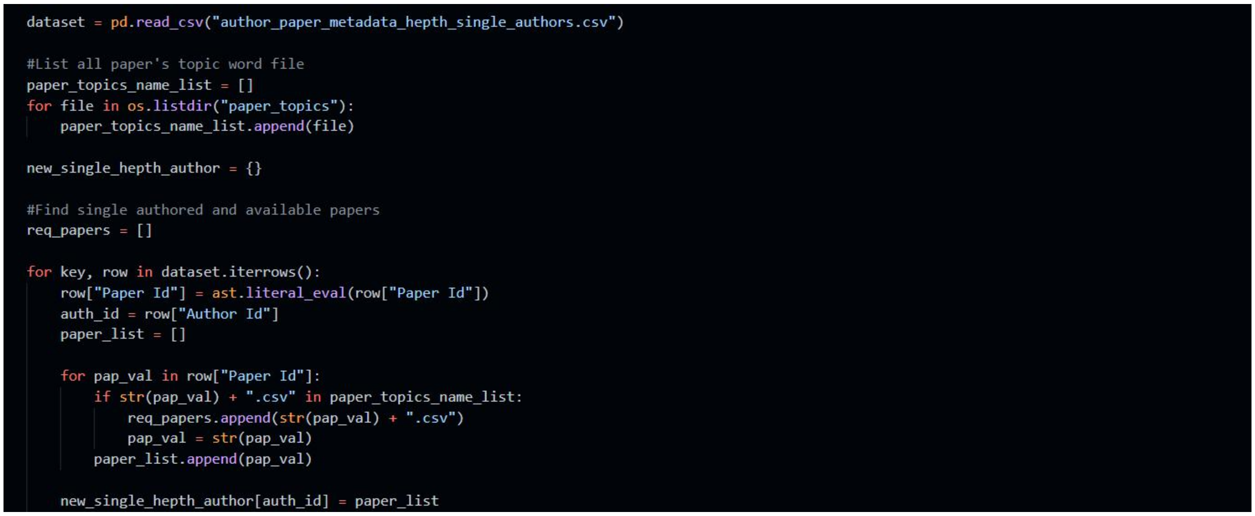
```
cd = {}
for index,row in dataset.iterrows():
  if row['Paper Id'] in cd.keys():
    cd[row['Paper Id']] = cd[row['Paper Id']] +1
  else:
    cd[row['Paper Id']] = 1
```

```
filtered_dict = {k:v for (k,v) in cd.items() if v == 1}
```

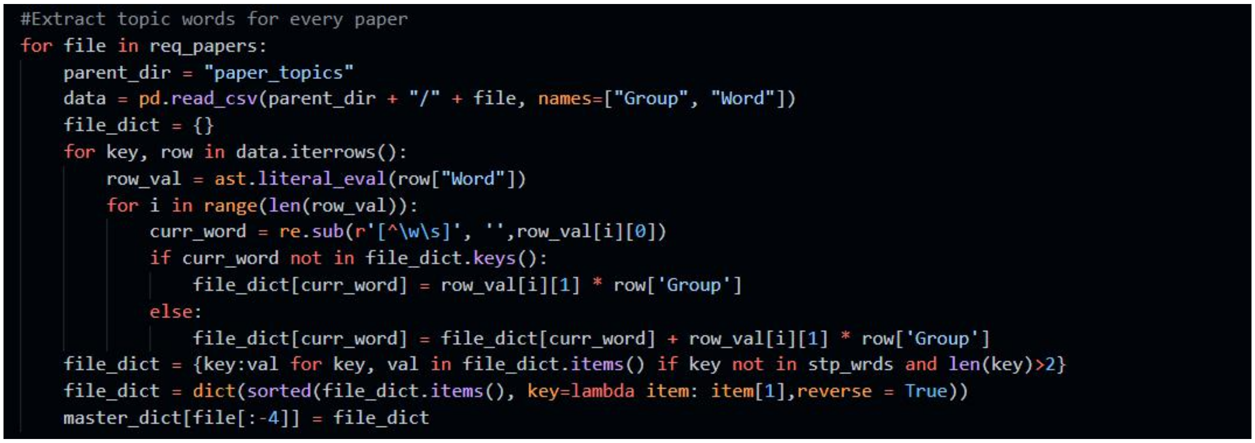


```
unique_paper_dict = {}
for index,row in dataset.iterrows():
  if row['Paper Id'] in unique_paper_ids:
    if row['Author Id'] in unique_paper_dict.keys():
      unique_paper_dict[row['Author Id']].append(row['Paper Id'])
    else:
      rl = []
      rl.append(row['Paper Id'])
      unique_paper_dict[row['Author Id']] = rl
```

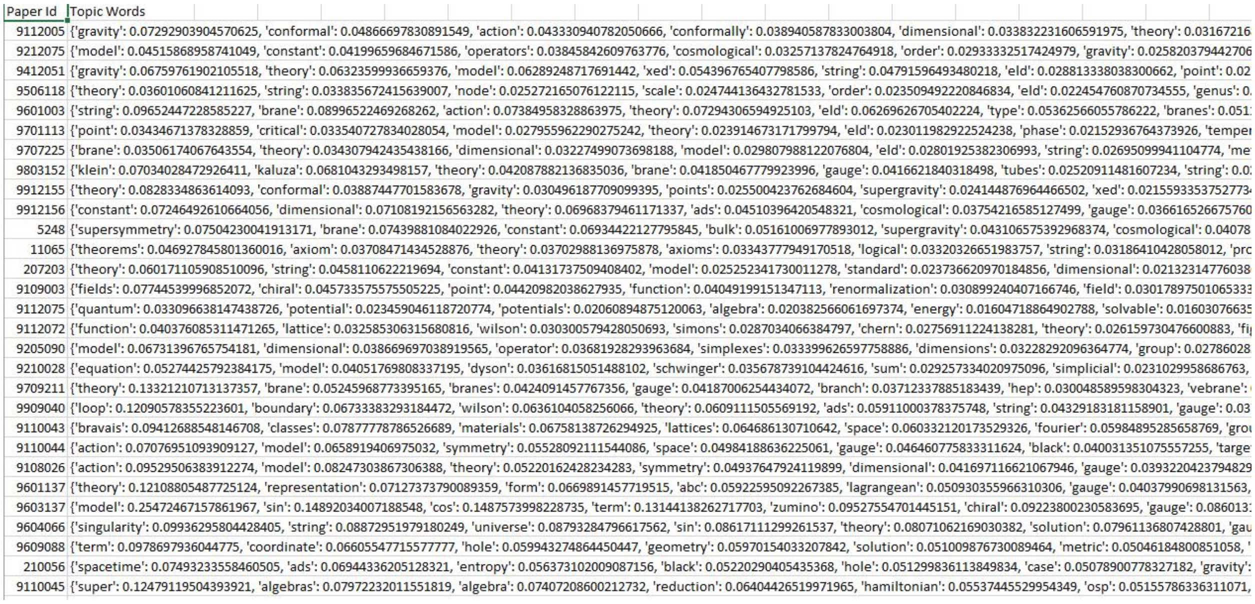
```
author_paper_metadata_hepth_single_author = pd.DataFrame(unique_paper_dict.items(),columns=['Author Id','Paper Id'])
```



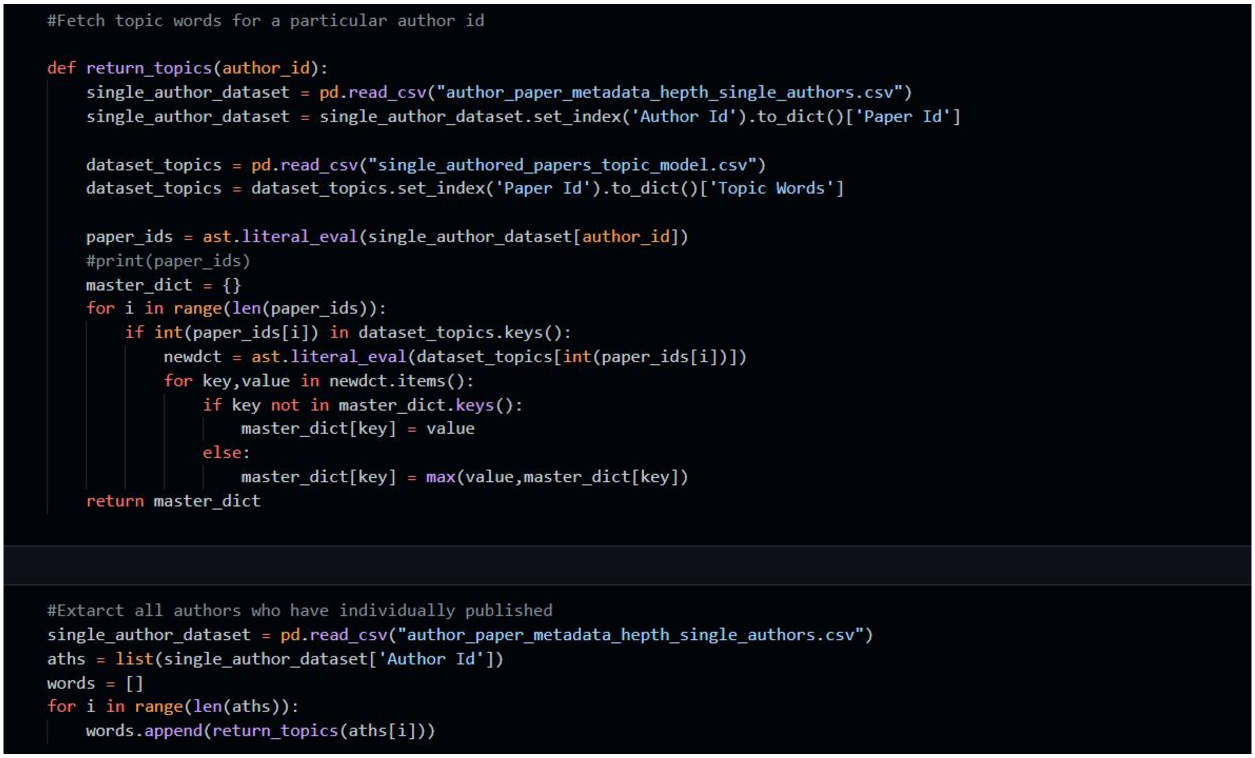
```
dataset = pd.read_csv("author_paper_metadata_hepth_single_authors.csv")

#List all paper's topic word file
paper_topics_name_list = []
for file in os.listdir("paper_topics"):
    paper_topics_name_list.append(file)

new_single_hepth_author = {}

#Find single authored and available papers
req_papers = []

for key, row in dataset.iterrows():
    row["Paper Id"] = ast.literal_eval(row["Paper Id"])
    auth_id = row["Author Id"]
    paper_list = []

    for pap_val in row["Paper Id"]:
        if str(pap_val) + ".csv" in paper_topics_name_list:
            req_papers.append(str(pap_val) + ".csv")
            pap_val = str(pap_val)
        paper_list.append(pap_val)

    new_single_hepth_author[auth_id] = paper_list
```



```
#Extract topic words for every paper
for file in req_papers:
    parent_dir = "paper_topics"
    data = pd.read_csv(parent_dir + "/" + file, names=["Group", "Word"])
    file_dict = {}
    for key, row in data.iterrows():
        row_val = ast.literal_eval(row["Word"])
        for i in range(len(row_val)):
            curr_word = re.sub(r'[^\w\s]', '',row_val[i][0])
            if curr_word not in file_dict.keys():
                file_dict[curr_word] = row_val[i][1] * row['Group']
            else:
                file_dict[curr_word] = file_dict[curr_word] + row_val[i][1] * row['Group']
    file_dict = {key:val for key, val in file_dict.items() if key not in stp_wrds and len(key)>2}
    file_dict = dict(sorted(file_dict.items(), key=lambda item: item[1],reverse = True))
    master_dict[file[:-4]] = file_dict
```

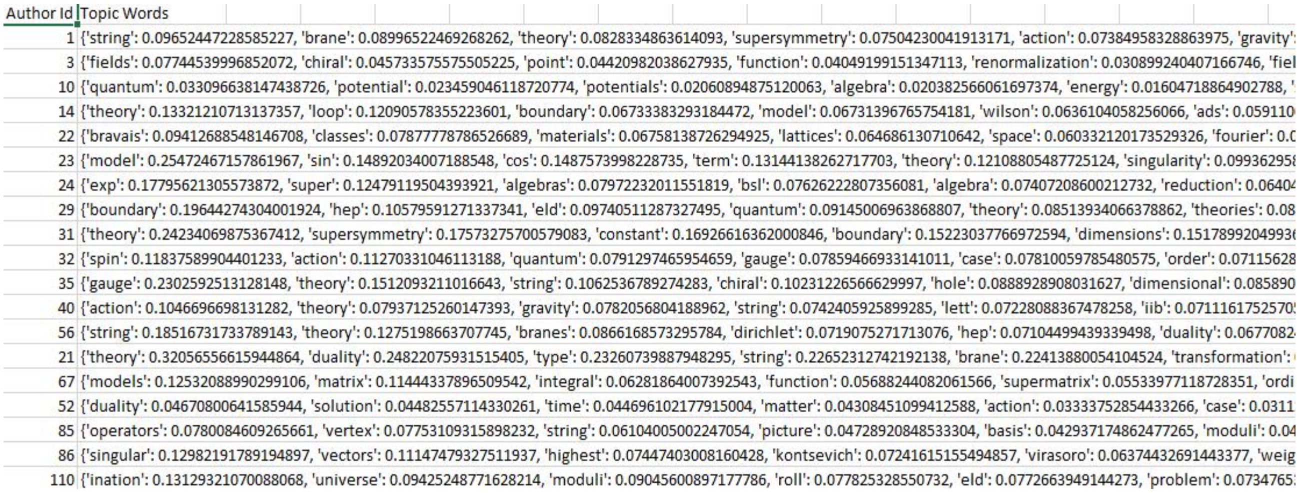


Paper Id	Topic Words
9112005	{'gravity': 0.07292903904570625, 'conformal': 0.04866697830891549, 'action': 0.043330940782050666, 'conformally': 0.038940587833003804, 'dimensional': 0.033832231606591975, 'theory': 0.03167216
9212075	{'model': 0.04515868958741049, 'constant': 0.04199659684671586, 'operators': 0.03845842609763776, 'cosmological': 0.03257137824764918, 'order': 0.02933332517424979, 'gravity': 0.025820379442706
9412051	{'gravity': 0.06759761902105518, 'theory': 0.06323599936659376, 'model': 0.06289248717691442, 'xed': 0.054396765407798586, 'string': 0.04791596493480218, 'eld': 0.028813338038300662, 'point': 0.02
9506118	{'theory': 0.03601060841211625, 'string': 0.033835672415639007, 'node': 0.025272165076122115, 'scale': 0.024744136432781533, 'order': 0.023509492220846834, 'eld': 0.022454760870734555, 'genus': 0.
9601003	{'string': 0.09652447228585227, 'brane': 0.08996522469268262, 'action': 0.07384958328863975, 'theory': 0.07294306594925103, 'eld': 0.06269626705402224, 'type': 0.05362566055786222, 'branes': 0.051
9701113	{'point': 0.03434671378328859, 'critical': 0.033540727834028054, 'model': 0.027955962290275242, 'theory': 0.023914673171799794, 'eld': 0.023011982922524238, 'phase': 0.02152936764373926, 'temper
9707225	{'brane': 0.03506174067643554, 'theory': 0.034307942435438166, 'dimensional': 0.03227499073698188, 'model': 0.029807988122076804, 'eld': 0.02801925382306993, 'string': 0.02695099941104774, 'me
9803152	{'klein': 0.07034028472926411, 'kaluza': 0.0681043293498157, 'theory': 0.042087882136835036, 'brane': 0.041850467779923996, 'gauge': 0.0416621840318498, 'tubes': 0.02520911481607234, 'string': 0.0
9912155	{'theory': 0.0828334863614093, 'conformal': 0.03887447701583678, 'gravity': 0.030496187709099395, 'points': 0.025500423762684604, 'supergravity': 0.024144876964466502, 'xed': 0.0215593353752773
9912156	{'constant': 0.07246492610664056, 'dimensional': 0.07108192156563282, 'theory': 0.06968379461171337, 'ads': 0.04510396420548321, 'cosmological': 0.03754216585127499, 'gauge': 0.036616526675760
5248	{'supersymmetry': 0.07504230041913171, 'brane': 0.07439881084022926, 'constant': 0.06934422127795845, 'bulk': 0.05161006977893012, 'supergravity': 0.043106575392968374, 'cosmological': 0.04078
11065	{'theorems': 0.046927845801360016, 'axiom': 0.03708471434528876, 'theory': 0.03702988136975878, 'axioms': 0.03343777949170518, 'logical': 0.03320326651983757, 'string': 0.03186410428058012, 'prc
207203	{'theory': 0.060171105908510096, 'string': 0.0458110622219694, 'constant': 0.04131737509408402, 'model': 0.025252341730011278, 'standard': 0.023736620970184856, 'dimensional': 0.02132314776038
9109003	{'fields': 0.07744539996852072, 'chiral': 0.045733575575505225, 'point': 0.04420982038627935, 'function': 0.04049199151347113, 'renormalization': 0.030899240407166746, 'field': 0.0301789750106533
9112075	{'quantum': 0.033096638147438726, 'potential': 0.023459046118720774, 'potentials': 0.020608948751 20063, 'algebra': 0.020382566061697374, 'energy': 0.01604718864902788, 'solvable': 0.01603076635
9112072	{'function': 0.040376085311471265, 'lattice': 0.032585306315680816, 'wilson': 0.030300579428050693, 'simons': 0.0287034066384797, 'chern': 0.02756911224138281, 'theory': 0.026159730476600883, 'fi
9205090	{'model': 0.06731396765754181, 'dimensional': 0.038669697038919565, 'operator': 0.03681928293963684, 'simplexes': 0.033399626597758886, 'dimensions': 0.03228292096364774, 'group': 0.02786028
9210028	{'equation': 0.05274425792384175, 'model': 0.04051769808337195, 'dyson': 0.03616815051488102, 'schwinger': 0.035678739104424616, 'sum': 0.029257334020975096, 'simplicial': 0.0231029958686763,
9709211	{'theory': 0.13321210713137357, 'brane': 0.05245968773395165, 'branes': 0.0424091457767356, 'gauge': 0.04187006254434072, 'branch': 0.03712337885183439, 'hep': 0.030048589598304323, 'vebrane':
9909040	{'loop': 0.12090578355223601, 'boundary': 0.06733383293184472, 'wilson': 0.0636104058256066, 'theory': 0.0609111505569192, 'ads': 0.05911000378375748, 'string': 0.04329183181158901, 'gauge': 0.03
9110043	{'bravais': 0.09412688548146708, 'classes': 0.07877778786526689, 'materials': 0.06758138726294925, 'lattices': 0.064686130710642, 'space': 0.060332120173529326, 'fourier': 0.05984895285658769, 'grou
9110044	{'action': 0.07076951093909127, 'model': 0.0658919406975032, 'symmetry': 0.05528092111544086, 'space': 0.04984188636225061, 'gauge': 0.046460775833311624, 'black': 0.040031351075557255, 'targe
9108026	{'action': 0.09529506383912274, 'model': 0.08247303867306388, 'theory': 0.05220162428234283, 'symmetry': 0.04937647924119899, 'dimensional': 0.041697116621067946, 'gauge': 0.0393220423794829
9601137	{'theory': 0.12108805487725124, 'representation': 0.07127373790089359, 'form': 0.0669891457719515, 'abc': 0.05922595092267385, 'lagrangean': 0.050930355966310306, 'gauge': 0.04037990698131563,
9603137	{'model': 0.25472467157861967, 'sin': 0.14892034007188548, 'cos': 0.1487573998228735, 'term': 0.13144138262717703, 'zumino': 0.09527554701445151, 'chiral': 0.09223800230583695, 'gauge': 0.086013
9604066	{'singularity': 0.09936295804428405, 'string': 0.08872951979180249, 'universe': 0.08793284796617562, 'sin': 0.08617111299261537, 'theory': 0.08071062169030382, 'solution': 0.07961136807428801, 'gau
9609088	{'term': 0.0978697936044775, 'coordinate': 0.06605547715577777, 'hole': 0.059943274864450447, 'geometry': 0.05970154033207842, 'solution': 0.051009876730089464, 'metric': 0.05046184800851058, '
210056	{'spacetime': 0.07493233558460505, 'ads': 0.06944336205128321, 'entropy': 0.056373102009087156, 'black': 0.05220290405435368, 'hole': 0.051299836113849834, 'case': 0.05078900778327182, 'gravity':
9110045	{'super': 0.12479119504393921, 'algebras': 0.07972232011551819, 'algebra': 0.07407208600212732, 'reduction': 0.06404426519971965, 'hamiltonian': 0.05537445529954349, 'osp': 0.05155786336311071,



```python
#Fetch topic words for a particular author id

def return_topics(author_id):
    single_author_dataset = pd.read_csv("author_paper_metadata_hepth_single_authors.csv")
    single_author_dataset = single_author_dataset.set_index('Author Id').to_dict()['Paper Id']

    dataset_topics = pd.read_csv("single_authored_papers_topic_model.csv")
    dataset_topics = dataset_topics.set_index('Paper Id').to_dict()['Topic Words']

    paper_ids = ast.literal_eval(single_author_dataset[author_id])
    #print(paper_ids)
    master_dict = {}
    for i in range(len(paper_ids)):
        if int(paper_ids[i]) in dataset_topics.keys():
            newdct = ast.literal_eval(dataset_topics[int(paper_ids[i])])
            for key,value in newdct.items():
                if key not in master_dict.keys():
                    master_dict[key] = value
                else:
                    master_dict[key] = max(value,master_dict[key])
    return master_dict
```

```python
#Extarct all authors who have individually published
single_author_dataset = pd.read_csv("author_paper_metadata_hepth_single_authors.csv")
aths = list(single_author_dataset['Author Id'])
words = []
for i in range(len(aths)):
    words.append(return_topics(aths[i]))
```



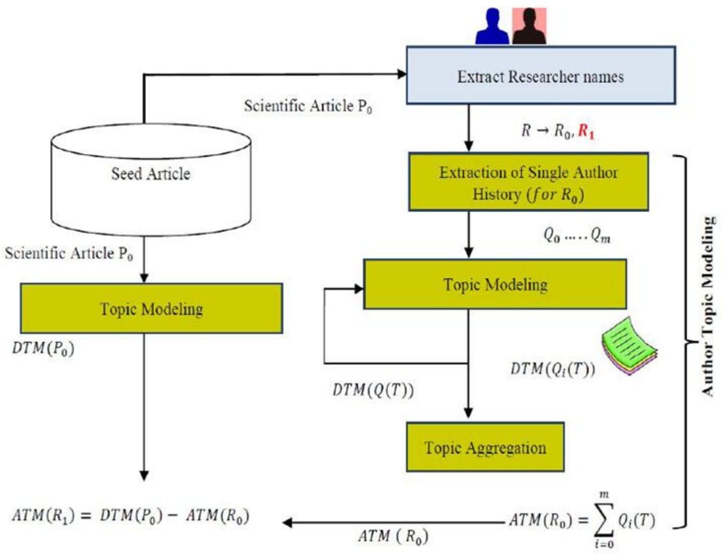
Author Id	Topic Words
1	{'string': 0.09652447228585227, 'brane': 0.08996522469268262, 'theory': 0.0828334863614093, 'supersymmetry': 0.07504230041913171, 'action': 0.07384958328863975, 'gravity':
3	{'fields': 0.07744539996852072, 'chiral': 0.045733575575505225, 'point': 0.04420982038627935, 'function': 0.04049199151347113, 'renormalization': 0.030899240407166746, 'fiel
10	{'quantum': 0.033096638147438726, 'potential': 0.023459046118720774, 'potentials': 0.02060894875120063, 'algebra': 0.020382566061697374, 'energy': 0.01604718864902788, '
14	{'theory': 0.13321210713137357, 'loop': 0.12090578355223601, 'boundary': 0.06733383293184472, 'model': 0.06731396765754181, 'wilson': 0.0636104058256066, 'ads': 0.059110
22	{'bravais': 0.09412688548146708, 'classes': 0.07877778786526689, 'materials': 0.06758138726294925, 'lattices': 0.064686130710642, 'space': 0.060332120173529326, 'fourier': 0.0
23	{'model': 0.25472467157861967, 'sin': 0.1489203400718854, 'cos': 0.1487573998228735, 'term': 0.13144138262717703, 'theory': 0.12108805487725124, 'singularity': 0.0993629
24	{'exp': 0.17795621305573872, 'super': 0.12479119504393921, 'algebras': 0.07972232011551819, 'bsl': 0.07626222807356081, 'algebra': 0.07407208600212732, 'reduction': 0.0640
29	{'boundary': 0.19644274304001924, 'hep': 0.10579591271337341, 'eld': 0.09740511287327495, 'quantum': 0.09145006963868807, 'theory': 0.08513934066378862, 'theories': 0.08
31	{'theory': 0.24234069875367412, 'supersymmetry': 0.17573275700579083, 'constant': 0.16926616362000846, 'boundary': 0.15223037766972594, 'dimensions': 0.151789920499
32	{'spin': 0.11837589904401233, 'action': 0.11270331046113188, 'quantum': 0.0791297465954659, 'gauge': 0.07859466933141011, 'case': 0.07810059785480575, 'order': 0.07115628
35	{'gauge': 0.2302592513128148, 'theory': 0.1512093211016643, 'string': 0.1062536789274283, 'chiral': 0.10231226566629997, 'hole': 0.0888928908031627, 'dimensional': 0.085890
40	{'action': 0.1046696698131282, 'theory': 0.07937125260147393, 'gravity': 0.0782056804188962, 'string': 0.0742405925899285, 'lett': 0.07228088367478258, 'iib': 0.0711161752570
56	{'string': 0.18516731733789143, 'theory': 0.1275198663707745, 'branes': 0.0866168573295784, 'dirichlet': 0.0719075271713076, 'hep': 0.07104499439339498, 'duality': 0.067708
21	{'theory': 0.32056556615944864, 'duality': 0.24822075931515405, 'type': 0.23260739887948295, 'string': 0.22652312742192138, 'brane': 0.22413880054104524, 'transformation':
67	{'models': 0.12532088990299106, 'matrix': 0.11444337896509542, 'integral': 0.06281864007392543, 'function': 0.05688244082061566, 'supermatrix': 0.05533977118728351, 'ordi
52	{'duality': 0.04670800641585944, 'solution': 0.04482557114330261, 'time': 0.044696102177915004, 'matter': 0.04308451099412588, 'action': 0.03333752854433266, 'case': 0.0311
85	{'operators': 0.0780084609265661, 'vertex': 0.07753109315898232, 'string': 0.06104005002247054, 'picture': 0.04728920848533304, 'basis': 0.042937174862477265, 'moduli': 0.04
86	{'singular': 0.12982191789194897, 'vectors': 0.11147479327511937, 'highest': 0.07447403008160428, 'kontsevich': 0.07241615155494857, 'virasoro': 0.06374432691443377, 'weig
110	{'ination': 0.13129321070088068, 'universe': 0.09425248771628214, 'moduli': 0.09045600897177786, 'roll': 0.077825328550732, 'eld': 0.0772663949144273, 'problem': 0.0734765

6. Multi Author Portrait Generation :

In the previous step, we would have obtained the topic words for authors who have published papers on their own without any co-authors. We use this result to derive the topic words for the rest of the authors who are a part of the dataset.

First, we obtain the topic words for each unique paper id by cross mapping between the parent dataset which has the paper id as the primary key and the topic modeled result for these papers. A new dataframe of the format (paper_id, author_ids) is created.

Author Topic Model (ATM) is generated by statistical means. Let us consider there is a research article (P_0) authored by two authors (R_0, R_1). Let us assume that the single author history of research articles is available only for (R_0). In other words, the single author history of R_0 is also known as ATM(R_0). This is obtained by aggregating the Document Topic Model (DTM) of every single author publication of $R_{0.}$



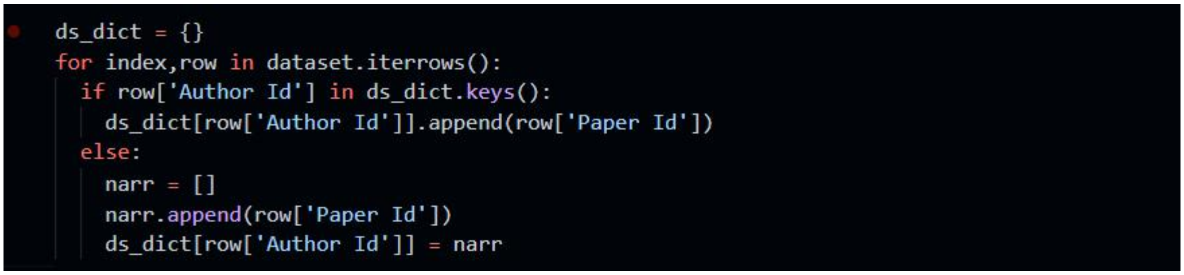
$$ATM(R_1) = DTM(P_0) - ATM(R_0)$$

This process is known as ATM by subtraction. Consider articles P_0 authored by R_0, R_1 and P_2 authored by R_0, R_1, R_2. If only one single author history is available, i.e. ATM(R0). Then, first the ATM of R_1 shall be formed by subtraction. Following this, the ATM of R_2 is found using the following equation.

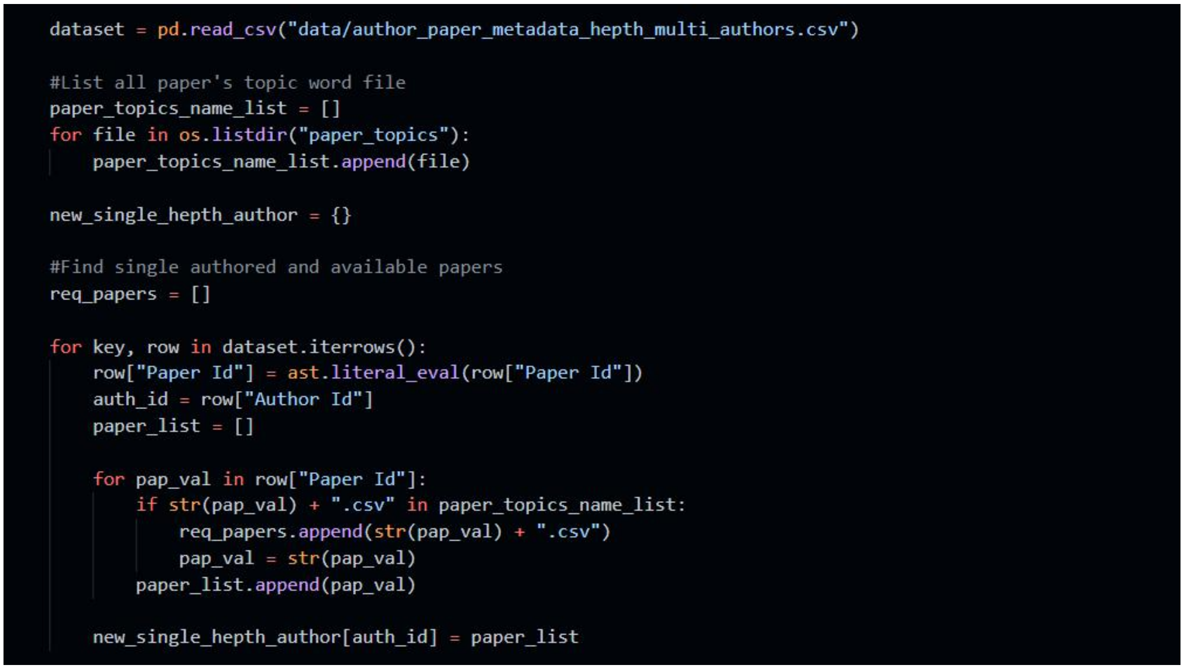
$$ATM(R_2) = DTM(P_3) - (ATM(R_0) \cup ATM(R_1))$$

Article #	Authors	Single Author History (availability)	Method
P_0	R_0, R_1	R_0	$ATM(R_1) = DTM(P_0) - ATM(R_0)$
P_0 P_1	R_0, R_1 R_0, R_1, R_2	R_0, R_1	$ATM(R_2) = DTM(P_1) - DTM(P_0)$ $ATM(R_2) = DTM(P_1) - (ATM(R_0) \cup ATM(R_1))$
P_1 P_2	R_0, R_1, R_2 R_2, R_3, R_4	NOT AVAILABLE	$ATM(R_2) = DTM(P_1) \cap DTM(P_2)$
P_0 P_3	R_0, R_1 R_0, R_1, R_2	R_0	$ATM(R_1) = DTM(P_0) - ATM(R_0)$ $ATM(R_2) = DTM(P_3) - (ATM(R_0) \cup ATM(R_1))$

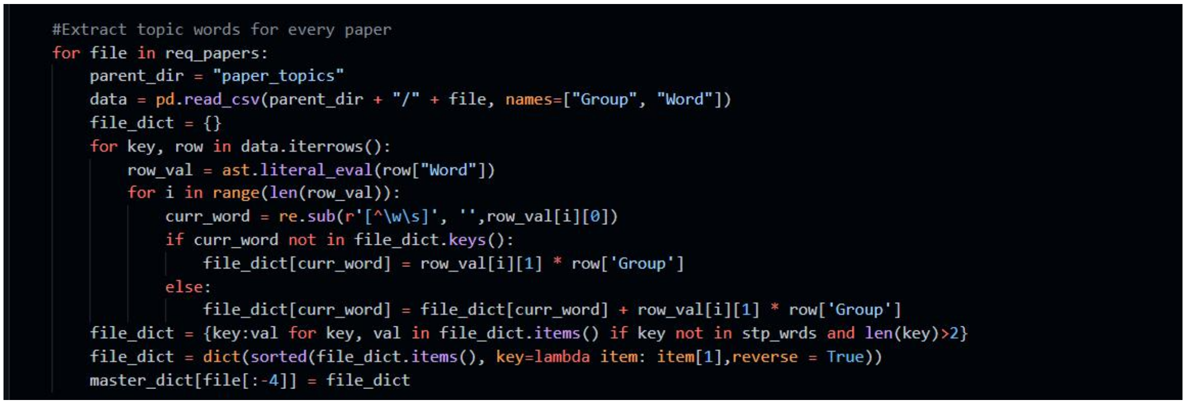
Code :



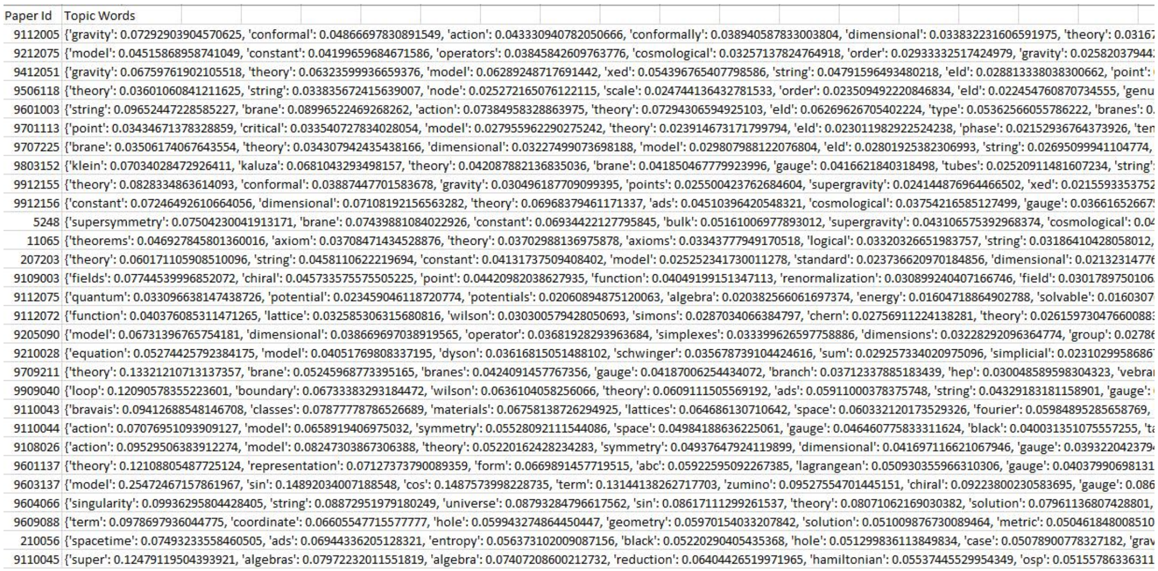
```
ds_dict = {}
for index,row in dataset.iterrows():
  if row['Author Id'] in ds_dict.keys():
    ds_dict[row['Author Id']].append(row['Paper Id'])
  else:
    narr = []
    narr.append(row['Paper Id'])
    ds_dict[row['Author Id']] = narr
```



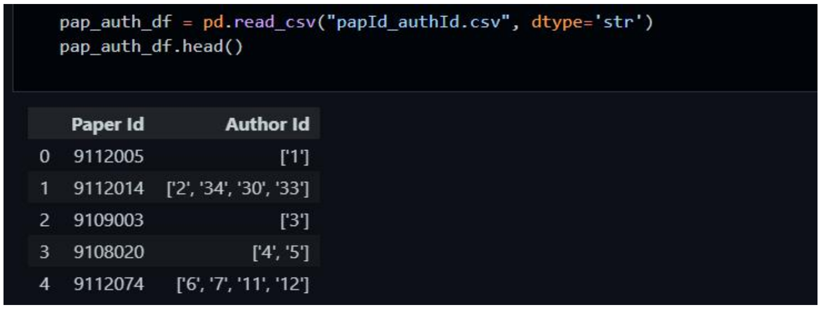
```
dataset = pd.read_csv("data/author_paper_metadata_hepth_multi_authors.csv")

#List all paper's topic word file
paper_topics_name_list = []
for file in os.listdir("paper_topics"):
    paper_topics_name_list.append(file)

new_single_hepth_author = {}

#Find single authored and available papers
req_papers = []

for key, row in dataset.iterrows():
    row["Paper Id"] = ast.literal_eval(row["Paper Id"])
    auth_id = row["Author Id"]
    paper_list = []

    for pap_val in row["Paper Id"]:
        if str(pap_val) + ".csv" in paper_topics_name_list:
            req_papers.append(str(pap_val) + ".csv")
            pap_val = str(pap_val)
        paper_list.append(pap_val)

    new_single_hepth_author[auth_id] = paper_list
```



```
#Extract topic words for every paper
for file in req_papers:
    parent_dir = "paper_topics"
    data = pd.read_csv(parent_dir + "/" + file, names=["Group", "Word"])
    file_dict = {}
    for key, row in data.iterrows():
        row_val = ast.literal_eval(row["Word"])
        for i in range(len(row_val)):
            curr_word = re.sub(r'[^\w\s]', '',row_val[i][0])
            if curr_word not in file_dict.keys():
                file_dict[curr_word] = row_val[i][1] * row['Group']
            else:
                file_dict[curr_word] = file_dict[curr_word] + row_val[i][1] * row['Group']
    file_dict = {key:val for key, val in file_dict.items() if key not in stp_wrds and len(key)>2}
    file_dict = dict(sorted(file_dict.items(), key=lambda item: item[1],reverse = True))
    master_dict[file[:-4]] = file_dict
```

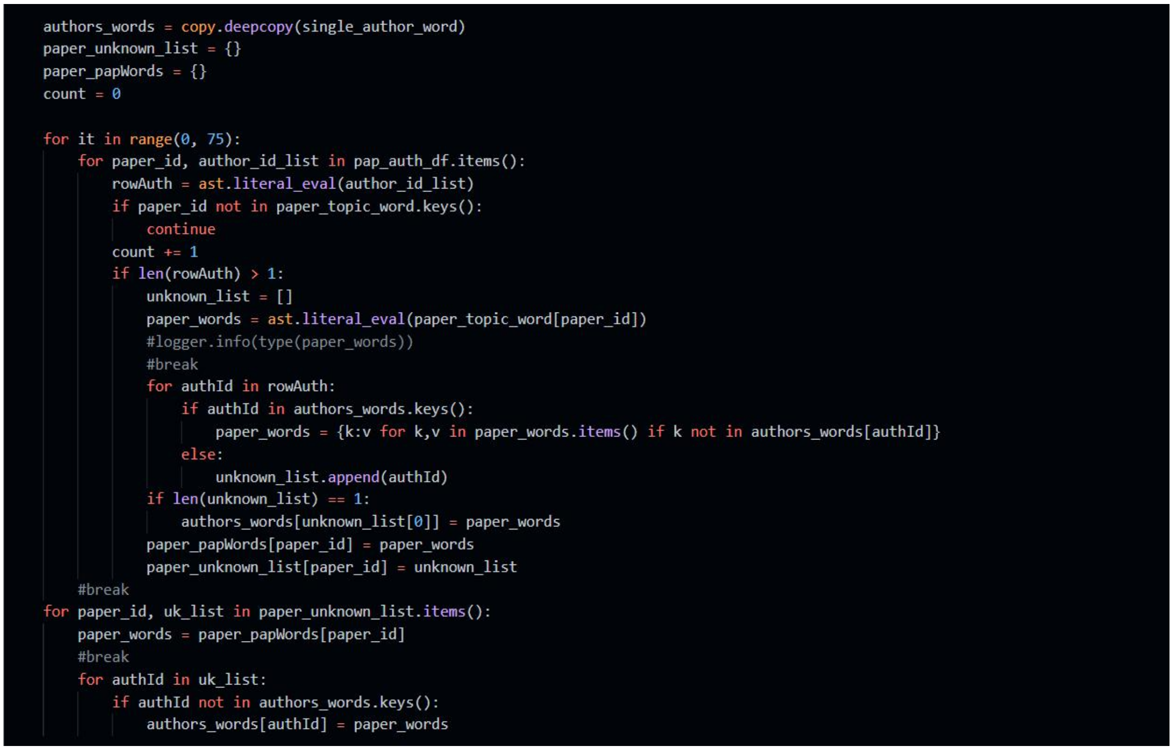


Paper Id	Topic Words
9112005	{'gravity': 0.07292903904570625, 'conformal': 0.04866697830891549, 'action': 0.043330940782050666, 'conformally': 0.038940587833003804, 'dimensional': 0.033832231606591975, 'theory': 0.03167
9212075	{'model': 0.04515868958741049, 'constant': 0.04199659684671586, 'operators': 0.03845842609763776, 'cosmological': 0.03257137824764918, 'order': 0.02933332517424979, 'gravity': 0.02582037944
9412051	{'gravity': 0.06759761902105518, 'theory': 0.06323599936659376, 'model': 0.06289248717691442, 'xed': 0.054396765407798586, 'string': 0.04791596493480218, 'eld': 0.028813338038300662, 'point':
9506118	{'theory': 0.03601060841211625, 'string': 0.033835672415639007, 'node': 0.025272165076122115, 'scale': 0.024744136432781533, 'order': 0.023509492220846834, 'eld': 0.022454760870734555, 'genu
9601003	{'string': 0.09652447228585227, 'brane': 0.08996522469268262, 'action': 0.07384958328863975, 'theory': 0.07294306594925103, 'eld': 0.06269626705402224, 'type': 0.05362566055786222, 'branes': 0.
9701113	{'point': 0.03434671378328859, 'critical': 0.033540727834028054, 'model': 0.027955962290275242, 'theory': 0.023914673171799794, 'eld': 0.023011982922524238, 'phase': 0.02152936764373926, 'ten
9707225	{'brane': 0.03506174067643554, 'theory': 0.034307942435438166, 'dimensional': 0.03227499073698188, 'model': 0.029807988122076804, 'eld': 0.02801925382306993, 'string': 0.02695099941104774,
9803152	{'klein': 0.07034028472926411, 'kaluza': 0.0681043293498157, 'theory': 0.042087882136835036, 'brane': 0.041850467779923996, 'gauge': 0.0416621840318498, 'tubes': 0.02520911481607234, 'string':
9912155	{'theory': 0.0828334863614093, 'conformal': 0.03887447701583678, 'gravity': 0.030496187709099395, 'points': 0.025500423762684604, 'supergravity': 0.024144876964466502, 'xed': 0.0215593353752
9912156	{'constant': 0.07246492610664056, 'dimensional': 0.07108192156563282, 'theory': 0.06968379461171337, 'ads': 0.04510396420548321, 'cosmological': 0.03754216585127499, 'gauge': 0.03661652667
5248	{'supersymmetry': 0.07504230041913171, 'brane': 0.07439881084022926, 'constant': 0.06934422127795845, 'bulk': 0.05161006977893012, 'supergravity': 0.043106575392968374, 'cosmological': 0.04
11065	{'theorems': 0.046927845801360016, 'axiom': 0.03708471434528876, 'theory': 0.03702988136975878, 'axioms': 0.03343777949170518, 'logical': 0.03320326651983757, 'string': 0.03186410428058012,
207203	{'theory': 0.060171105908510096, 'string': 0.0458110622219694, 'constant': 0.04131737509408402, 'model': 0.025252341730011278, 'standard': 0.023736620970184856, 'dimensional': 0.0213231477
9109003	{'fields': 0.07744539996852072, 'chiral': 0.045733575575505225, 'point': 0.04420982038627935, 'function': 0.04049199151347113, 'renormalization': 0.030899240407166746, 'field': 0.0301789750106
9112075	{'quantum': 0.033096638147438726, 'potential': 0.023459046118720774, 'potentials': 0.02060894875120063, 'algebra': 0.020382566061697374, 'energy': 0.01604718864902788, 'solvable': 0.0160307
9112072	{'function': 0.040376085311471265, 'lattice': 0.032585306315680816, 'wilson': 0.030300579428050693, 'simons': 0.0287034066384797, 'chern': 0.02756911224138281, 'theory': 0.02615973047660088
9205090	{'model': 0.06731396765754181, 'dimensional': 0.038669697038919565, 'operator': 0.03681928293963684, 'simplexes': 0.033399626597758886, 'dimensions': 0.03228292096364774, 'group': 0.0278
9210028	{'equation': 0.05274425792384175, 'model': 0.04051769808337195, 'dyson': 0.03616815051488102, 'schwinger': 0.035678739104424616, 'sum': 0.029257334020975096, 'simplicial': 0.0231029958686
9709211	{'theory': 0.13321210713137357, 'brane': 0.05245968773395165, 'branes': 0.0424091457767356, 'gauge': 0.04187006254434072, 'branch': 0.03712337885183439, 'hep': 0.030048589598304323, 'vebra
9909040	{'loop': 0.12090578355223601, 'boundary': 0.06733383293184472, 'wilson': 0.0636104058256066, 'theory': 0.0609111505569192, 'ads': 0.05911000378375748, 'string': 0.04329183181158901, 'gauge':
9110043	{'bravais': 0.09412688548146708, 'classes': 0.07877778786526689, 'materials': 0.06758138726294925, 'lattices': 0.064686130710642, 'space': 0.060332120173529326, 'fourier': 0.05984895285658769,
9110044	{'action': 0.07076951093909127, 'model': 0.0658919406975032, 'symmetry': 0.05528092111544086, 'space': 0.04984188636225061, 'gauge': 0.046460775833311624, 'black': 0.040031351075557255, 'ta
9108026	{'action': 0.09529506383912274, 'model': 0.08247303867306388, 'theory': 0.05220162428234283, 'symmetry': 0.04937647924119899, 'dimensional': 0.041697116621067946, 'gauge': 0.039322042379
9601137	{'theory': 0.12108805487725124, 'representation': 0.07127373790089359, 'form': 0.0669891457719515, 'abc': 0.05922595092267385, 'lagrangean': 0.050930355966310306, 'gauge': 0.04037990698131
9603137	{'model': 0.25472467157861967, 'sin': 0.14892034007188548, 'cos': 0.1487573998228735, 'term': 0.13144138262717703, 'zumino': 0.09527554701445151, 'chiral': 0.09223800230583695, 'gauge': 0.086
9604066	{'singularity': 0.09936295804428405, 'string': 0.08872951979180249, 'universe': 0.08793284796617562, 'sin': 0.08617111299261537, 'theory': 0.08071062169030382, 'solution': 0.07961136807428801,
9609088	{'term': 0.0978697936044775, 'coordinate': 0.06605547715577777, 'hole': 0.059943274864450447, 'geometry': 0.05970154033207842, 'solution': 0.051009876730089464, 'metric': 0.050461848008510
210056	{'spacetime': 0.07493233558460505, 'ads': 0.06944336205128321, 'entropy': 0.056373102009087156, 'black': 0.05220290405435368, 'hole': 0.051299836113849834, 'case': 0.05078900778327182, 'grav
9110045	{'super': 0.12479119504393921, 'algebras': 0.07972232011551819, 'algebra': 0.07407208600212732, 'reduction': 0.06404426519971965, 'hamiltonian': 0.05537445529954349, 'osp': 0.05155786336311



```
pap_auth_df = pd.read_csv("papId_authId.csv", dtype='str')
pap_auth_df.head()
```

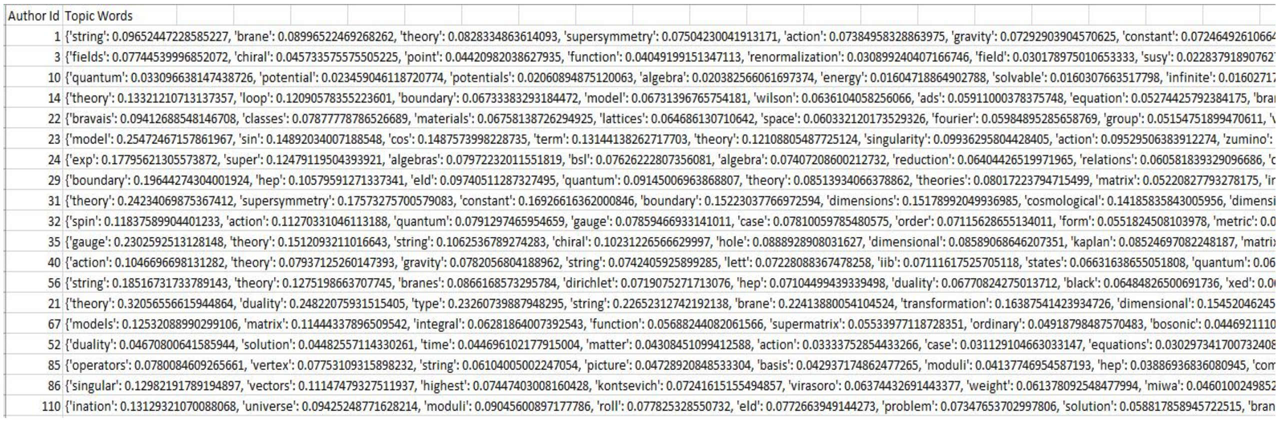
	Paper Id	Author Id
0	9112005	['1']
1	9112014	['2', '34', '30', '33']
2	9109003	['3']
3	9108020	['4', '5']
4	9112074	['6', '7', '11', '12']



```
authors_words = copy.deepcopy(single_author_word)
paper_unknown_list = {}
paper_papWords = {}
count = 0

for it in range(0, 75):
    for paper_id, author_id_list in pap_auth_df.items():
        rowAuth = ast.literal_eval(author_id_list)
        if paper_id not in paper_topic_word.keys():
            continue
        count += 1
        if len(rowAuth) > 1:
            unknown_list = []
            paper_words = ast.literal_eval(paper_topic_word[paper_id])
            #logger.info(type(paper_words))
            #break
            for authId in rowAuth:
                if authId in authors_words.keys():
                    paper_words = {k:v for k,v in paper_words.items() if k not in authors_words[authId]}
                else:
                    unknown_list.append(authId)
            if len(unknown_list) == 1:
                authors_words[unknown_list[0]] = paper_words
            paper_papWords[paper_id] = paper_words
            paper_unknown_list[paper_id] = unknown_list
    #break
for paper_id, uk_list in paper_unknown_list.items():
    paper_words = paper_papWords[paper_id]
    #break
    for authId in uk_list:
        if authId not in authors_words.keys():
            authors_words[authId] = paper_words
```

7. Final output :

Final output consisting of portraits for each author.



Author Id	Topic Words
1	{'string': 0.09652447228585227, 'brane': 0.08996522469268262, 'theory': 0.0828334863614093, 'supersymmetry': 0.07504230041913171, 'action': 0.07384958328863975, 'gravity': 0.07292903904570625, 'constant': 0.0724649261066
3	{'fields': 0.07744539996852072, 'chiral': 0.045733575575505225, 'point': 0.04420982038627935, 'function': 0.04049199151347113, 'renormalization': 0.030899240407166746, 'field': 0.030178975010653333, 'susy': 0.0228379189076
10	{'quantum': 0.033096638147438726, 'potential': 0.023459046118720774, 'potentials': 0.02060894875120063, 'algebra': 0.020382566061697374, 'energy': 0.01604718864902788, 'solvable': 0.0160307663517798, 'infinite': 0.016027
14	{'theory': 0.13321210713137357, 'loop': 0.12090578355223601, 'boundary': 0.06733383293184472, 'model': 0.06731396765754181, 'wilson': 0.0636104058256066, 'ads': 0.05911000378375748, 'equation': 0.05274425792384175, 'bra
22	{'bravais': 0.09412688548146708, 'classes': 0.07877778786526689, 'materials': 0.06758138726294925, 'lattices': 0.06468613071064 2, 'space': 0.060332120173529326, 'fourier': 0.05984895285658769, 'group': 0.05154751899470611, '
23	{'model': 0.2547246715786 1967, 'sin': 0.1489203400718 8548, 'cos': 0.148757399822 8735, 'term': 0.13144138262717703, 'theory': 0.12108805487725124, 'singularity': 0.09936295804428405, 'action': 0.09529506383912274, 'zumino':
24	{'exp': 0.17795621305573872, 'super': 0.12479119504393921, 'algebras': 0.07972232011551819, 'bsl': 0.07626222807356081, 'algebra': 0.07407208600212732, 'reduction': 0.06404426519971965, 'relations': 0.060581839329096686, 'c
29	{'boundary': 0.19644274304001924, 'hep': 0.10579591271337341, 'eld': 0.09740511287327495, 'quantum': 0.09145006963868807, 'theory': 0.08513934066378862, 'theories': 0.08017223794715499, 'matrix': 0.05220827793278175, 'ir
31	{'theory': 0.24234069875367412, 'supersymmetry': 0.17573275700579083, 'constant': 0.16926616362000846, 'boundary': 0.15223037766972594, 'dimensions': 0.15178992049936985, 'cosmological': 0.14185835843005956, 'dimensi
32	{'spin': 0.11837589904401233, 'action': 0.11270331046113188, 'quantum': 0.0791297465954659, 'gauge': 0.07859466933141011, 'case': 0.07810059785480575, 'order': 0.07115628655134011, 'form': 0.0551824508103978, 'metric': 0.0
35	{'gauge': 0.230259251312814 8, 'theory': 0.1512093211016643, 'string': 0.1062536789274283, 'chiral': 0.10231226566629997, 'hole': 0.0888928908031627, 'dimensional': 0.08589068646207351, 'kaplan': 0.08524697082248187, 'matri
40	{'action': 0.1046696698131282, 'theory': 0.07937125260147393, 'gravity': 0.0782056804188962, 'string': 0.07424059258992 85, 'lett': 0.0722808836747825 8, 'iib': 0.07111617525705118, 'states': 0.06631638655051808, 'quantum': 0.06
56	{'string': 0.18516731733789143, 'theory': 0.1275198663707745, 'branes': 0.0866168573295784, 'dirichlet': 0.0719075271713076, 'hep': 0.07104499439339498, 'duality': 0.06770824275013712, 'black': 0.06484826500691736, 'xed': 0.0
21	{'theory': 0.32056556615944864, 'duality': 0.24822075931515405, 'type': 0.23260739887948295, 'string': 0.22652312742192138, 'brane': 0.22413880054104524, 'transformation': 0.16387541423934726, 'dimensional': 0.1545204624 5
67	{'models': 0.12532088990299106, 'matrix': 0.11444337896509542, 'integral': 0.06281864007392543, 'function': 0.05688244082061566, 'supermatrix': 0.05533977118728351, 'ordinary': 0.04918798487570483, 'bosonic': 0.0446921110
52	{'duality': 0.04670800641585944, 'solution': 0.04482557114330261, 'time': 0.04469610217791500 4, 'matter': 0.04308451099412588, 'action': 0.03333752854433266, 'case': 0.031129104663033147, 'equations': 0.030297341700732408
85	{'operators': 0.0780084609265661, 'vertex': 0.07753109315898232, 'string': 0.06104005002247054, 'picture': 0.0472892084853330 4, 'basis': 0.04293717486247726 5, 'moduli': 0.04137746954587193, 'hep': 0.03886936836080945, 'com
86	{'singular': 0.12982191789194897, 'vectors': 0.11147479327511937, 'highest': 0.07447403008160428, 'kontsevich': 0.07241615155494857, 'virasoro': 0.06374432691443377, 'weight': 0.06137809254847799 4, 'miwa': 0.046010024985 2
110	{'ination': 0.13129321070088068, 'universe': 0.09425248771628214, 'moduli': 0.09045600897177786, 'roll': 0.07782532855073 2, 'eld': 0.0772663949144273, 'problem': 0.07347653702997806, 'solution': 0.05881785894572251 5, 'bran