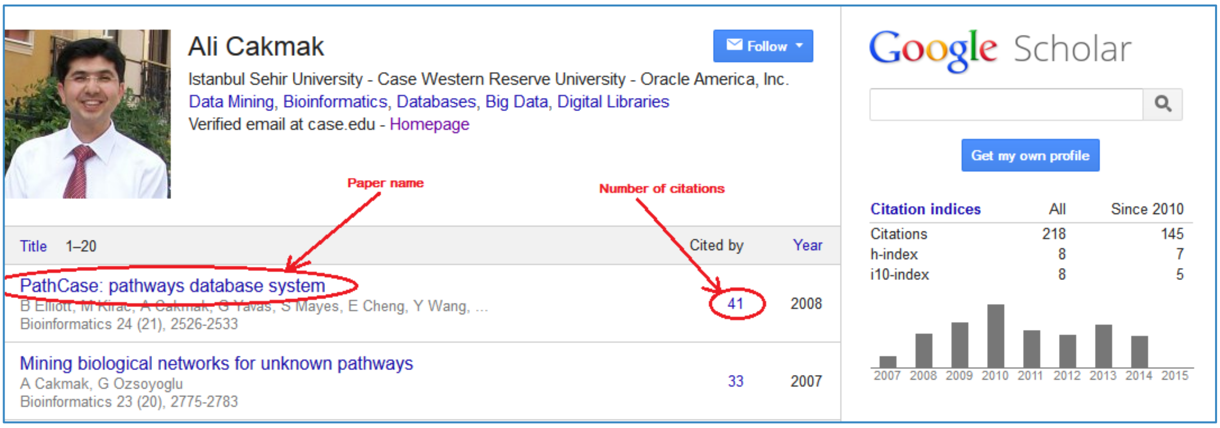
ENGR 212 – PROGRAMMING PRACTICE

SPRING 2015

MINI PROJECT 2

March 27, 2015

Google Scholar allows researchers to create a profile page (e.g., see Dr. Cakmak's Google Scholar profile at http://scholar.google.com.tr/citations?user=mBvj3CMAAAAJ&hl=en). Once a researcher creates a profile page, from that point on, Google Scholar automatically maintains his/her publication list, as well as the citations that his/her works get (see Figure 1).



Ali Cakmak

Istanbul Sehir University - Case Western Reserve University - Oracle America, Inc.

Data Mining, Bioinformatics, Databases, Big Data, Digital Libraries

Verified email at case.edu - Homepage

Title 1–20	Cited by	Year
PathCase: pathways database system B Elliott, M Kirac, A Cakmak, G Yavas, S Mayes, E Cheng, Y Wang, ... Bioinformatics 24 (21), 2526-2533	41	2008
Mining biological networks for unknown pathways A Cakmak, G Ozsoyoglu Bioinformatics 23 (20), 2775-2783	33	2007

Citation indices	All	Since 2010
Citations	218	145
h-index	8	7
i10-index	8	5

Figure 1

In this mini project, you are going to write a Python program that, given URLs for the Google Scholar profiles of a set of researchers, it will fetch information from their Google Scholar profiles, and provide some advanced viewing and clustering functionalities. Here are the detailed specifications for the behavior of your program.

1. Your program's main driver method should be named as "fetch", and it should be placed in a class, called GoogleScholarFetcher. This class's __init__ method should take a URL as input, and assign it to a properly named instance-level attribute. As an example, a sample use of your class may be as follows. You are encouraged having other classes in your program to practice object-oriented programming.

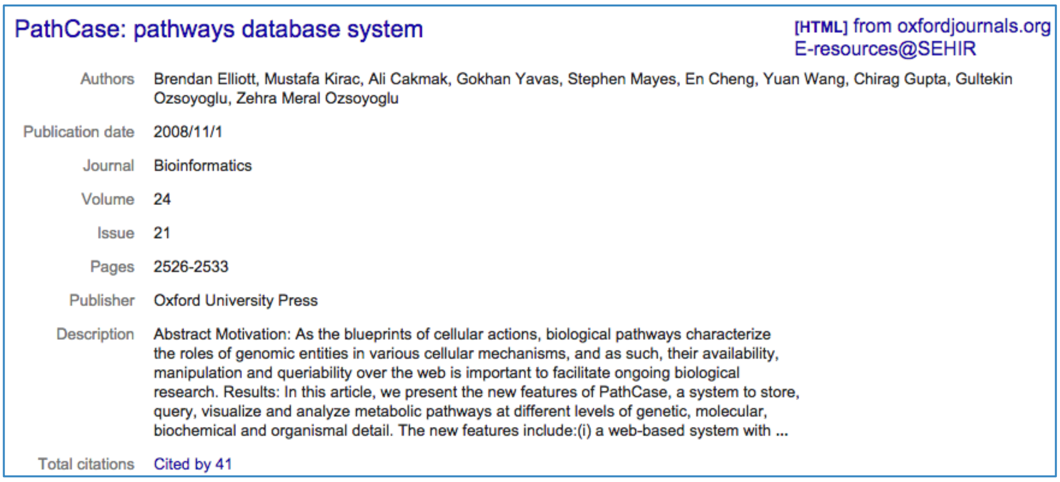
```
fetcher = GoogleScholarFetcher('http://scholar.google.com.tr/citations?user=mBvj3CMAAAAJ&hl=en')

fetcher.fetch()
```

2. Each publication title is linked to the details of that publication. As an example, Figure 2 shows the details of the circled-publication in the above figure, which is a journal publication. Your program should get into the details page of each publication by following its link, and extract each listed field with their values (e.g., authors, publication date, journal, volume, issue, etc.). Please note that fields in details page may be slightly different for different types (e.g., conference papers, journal papers, etc) of papers. See Figure 3 as an example of a detail page for a conference paper. Your program should be able to automatically differentiate between different types of papers (based on available fields), and record paper type as part of other paper information that you are going to fetch from Google Scholar. (More explanation is provided on page 4 on different types of papers)

3. Your class should have another method, called "reformat". This method will output the publications for a researcher in different flavors of list format (read on for details). This method should take two input parameters:
    a. group_by: This parameter will define how the publications should be grouped. It may accept two values: 'year', 'type'. If 'year' is selected, publications will be grouped by year. If 'type' is specified, publications will be grouped by publication type (e.g.,

"Journal Papers" and "Conference Papers"). If the value of this parameter is None, then no grouping should be done. Examples are provided below.

b. sort_by: This parameter will specify the way that the publications will be sorted in the list format. It may accept two values: 'year', 'citation count'. Accordingly, publications will be sorted either by publication date or citation count. Default value should be 'year'. It cannot be None. Sorting is applied locally in each group if group_by parameter is not None. Otherwise (i.e., if group_by parameter is None), sorting is applied globally on the whole list. Examples are provided in Figures 6 through 8.

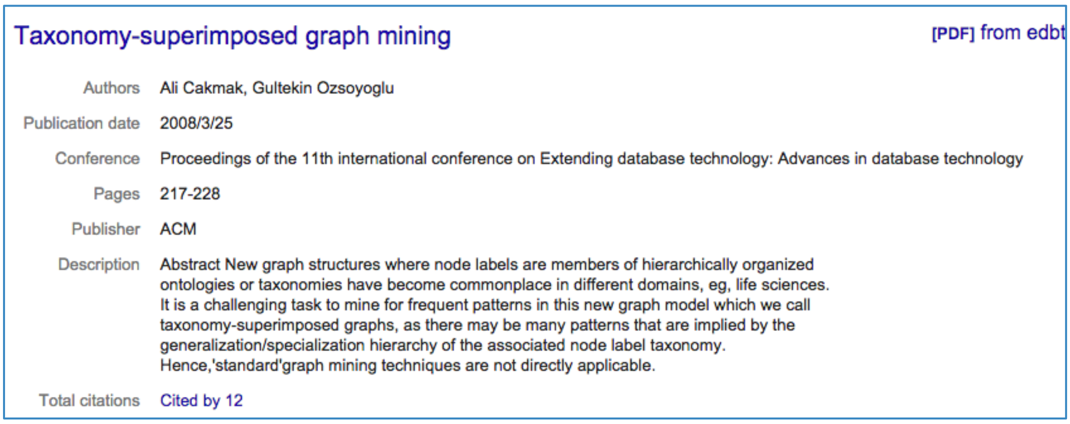


PathCase: pathways database system

[HTML] from oxfordjournals.org
E-resources@SEHIR

Authors	Brendan Elliott, Mustafa Kirac, Ali Cakmak, Gokhan Yavas, Stephen Mayes, En Cheng, Yuan Wang, Chirag Gupta, Gultekin Ozsoyoglu, Zehra Meral Ozsoyoglu
Publication date	2008/11/1
Journal	Bioinformatics
Volume	24
Issue	21
Pages	2526-2533
Publisher	Oxford University Press
Description	Abstract Motivation: As the blueprints of cellular actions, biological pathways characterize the roles of genomic entities in various cellular mechanisms, and as such, their availability, manipulation and queriability over the web is important to facilitate ongoing biological research. Results: In this article, we present the new features of PathCase, a system to store, query, visualize and analyze metabolic pathways at different levels of genetic, molecular, biochemical and organismal detail. The new features include:(i) a web-based system with ...
Total citations	Cited by 41

Figure 2



Taxonomy-superimposed graph mining

[PDF] from edbt

Authors	Ali Cakmak, Gultekin Ozsoyoglu
Publication date	2008/3/25
Conference	Proceedings of the 11th international conference on Extending database technology: Advances in database technology
Pages	217-228
Publisher	ACM
Description	Abstract New graph structures where node labels are members of hierarchically organized ontologies or taxonomies have become commonplace in different domains, eg, life sciences. It is a challenging task to mine for frequent patterns in this new graph model which we call taxonomy-superimposed graphs, as there may be many patterns that are implied by the generalization/specialization hierarchy of the associated node label taxonomy. Hence,'standard'graph mining techniques are not directly applicable.
Total citations	Cited by 12

Figure 3

4. Finally, after downloading the publication list of each researcher, your program should allow to cluster these researchers based on the words that appear in their publication titles and conference/journal names that their publications appear in.

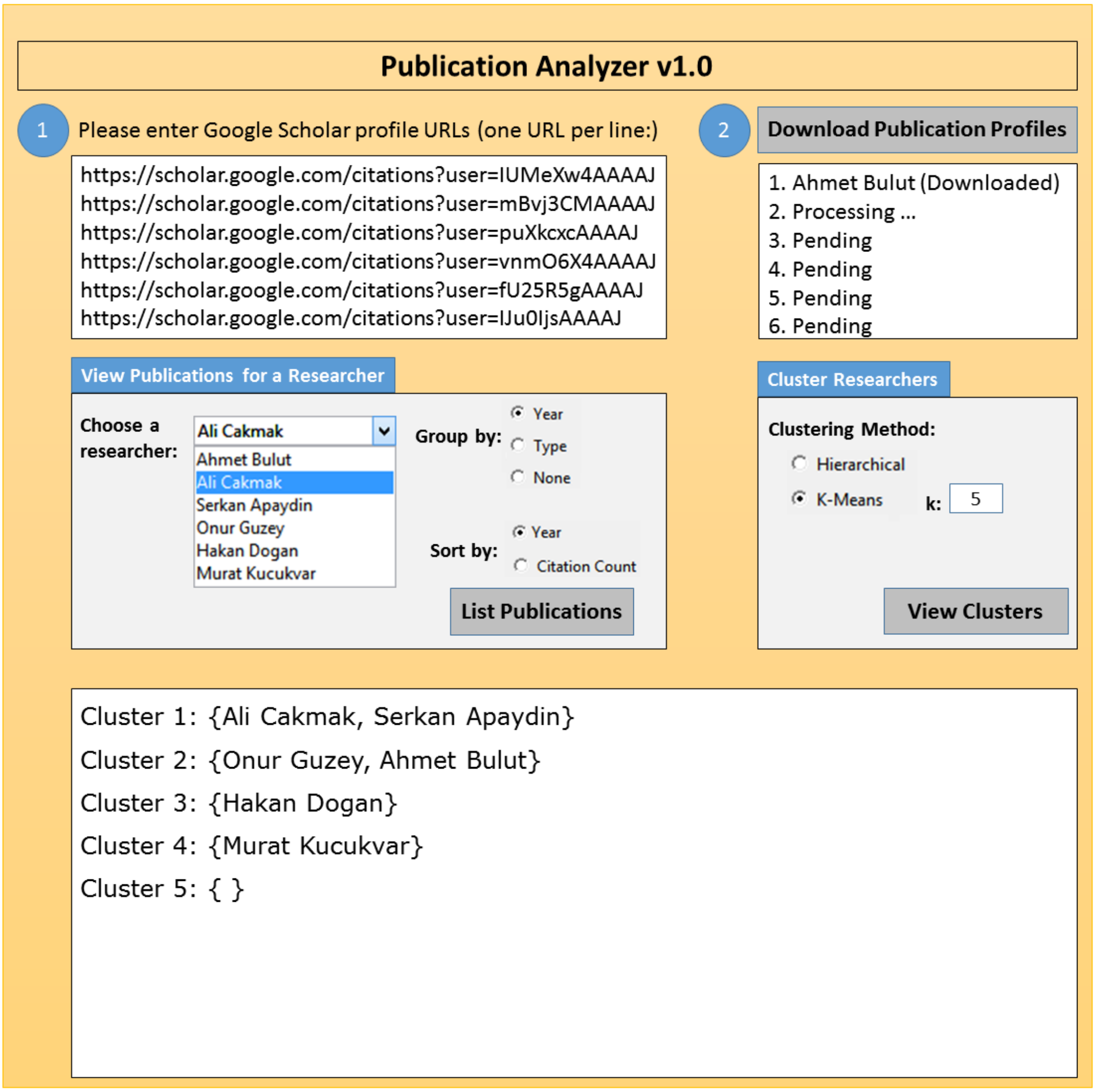
GUI

Your program will have a graphical user interface, and users will interact with your program through widgets on this interface. Figure 4 shows how the GUI of your program should be organized like (colorings are optional, but we expect you to place widgets in a similar way on GUI). Please note that results shown on Figures 1 through 8 are not from real runs. Hence, do not consider them as the correct results that your program should produce. Just take them as general guides.

- First, users should provide a set of URLs (one per line) each of which belongs to the Google Scholar page of a researcher.
- Then, the user clicks on a button labeled as "Download Publication Profiles". Once this button is clicked, your program should download and process each page using urllib2 and BeautifulSoup. Your program should let the user know about its progress in the box below the download button. As seen in Figure 4, for each URL, progress is reported with the name of the person which is extracted from the downloaded profile page. In Figure 4, it is shown that the

first URL that belongs to Ahmet Bulut has been downloaded, and it is currently processing the second URL, while others are still in the queue (i.e., pending). As your program progress, it should update this list (note that there is one entry per URL in this progress list). If the URL list in step 1 is empty, then your program should show an error message ("Please provide a few URLs") in this progress reporting area when download button is clicked. Note that in this example, 6 URLs are provided, but you should not assume any fixed value for the number of URLs. User may provide any number of URLs (e.g., 1, 5, 100, etc.)

- Once downloading of all pages is completed, then the user can do two things: (i) cluster researchers based on their publication information, and (ii) list publications of a selected researcher in different orderings and groupings. Details are provided below.



Publication Analyzer v1.0

1 Please enter Google Scholar profile URLs (one URL per line:)

```
https://scholar.google.com/citations?user=IUMeXw4AAAAJ
https://scholar.google.com/citations?user=mBvj3CMAAAAJ
https://scholar.google.com/citations?user=puXkcxcAAAAJ
https://scholar.google.com/citations?user=vnmO6X4AAAAJ
https://scholar.google.com/citations?user=fU25R5gAAAAJ
https://scholar.google.com/citations?user=IJu0ljsAAAAJ
```

2 Download Publication Profiles

1. Ahmet Bulut (Downloaded)
2. Processing ...
3. Pending
4. Pending
5. Pending
6. Pending

View Publications for a Researcher

Choose a researcher: Ali Cakmak

Group by: ● Year ○ Type ○ None

Sort by: ● Year ○ Citation Count

Cluster Researchers

Clustering Method: ○ Hierarchical ● K-Means k: 5

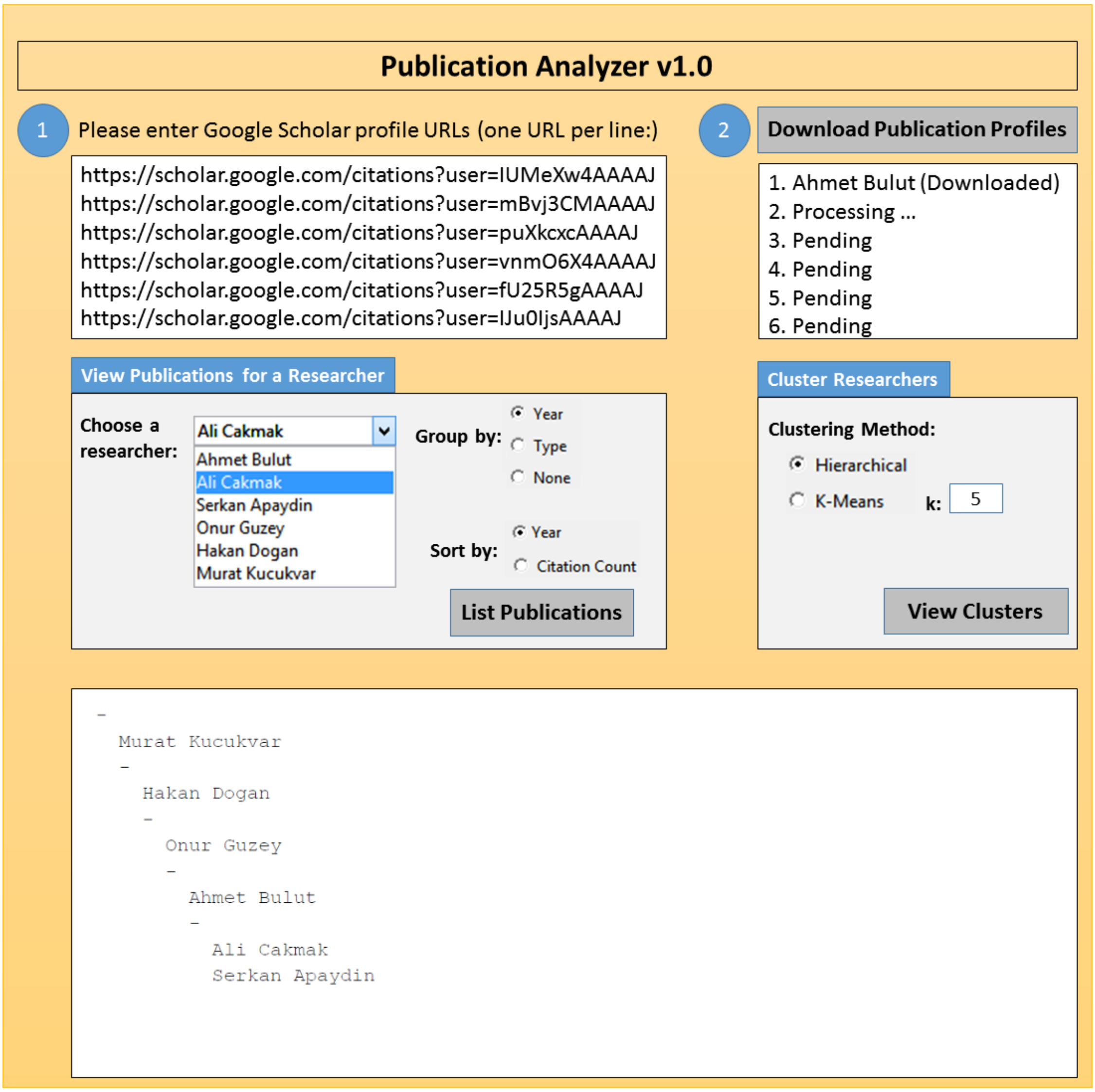
Cluster 1: {Ali Cakmak, Serkan Apaydin}
Cluster 2: {Onur Guzey, Ahmet Bulut}
Cluster 3: {Hakan Dogan}
Cluster 4: {Murat Kucukvar}
Cluster 5: { }

Figure 4

- **Clustering Researchers**: On the right-hand side of the middle row in Figure 4, a set of controls are provided for the users to perform clustering on researchers.
  - Users may choose hierarchical or k-means clustering. In case of k-means clustering, users should provide the value of k as well (by default it will be 5).
  - When "view clusters" button is clicked, then users should see the result of the clustering at the bottom part. For hierarchical clustering, results should be shown as a string dendogram (use *printclust* in clusters module) (Figure 5 shows an example), while for k-means clustering, the results will be shown in text form by listing each member of clusters (see Figure 4 for an example) as shown in lecture slides.
  - As for the clustering data, you are going to prepare a matrix similar to the one in blogdata.txt where rows will be researchers and columns will be words that appear in publication titles and publication venue names (i.e., conference names, journal names,

    etc.) where those publications appeared. The body/data in the matrix will be the frequencies of the corresponding words in publication titles and venue names for each researcher (see the use of feeds.py in lecture and book examples).

    - For the implementation of all the above clustering features, you are going to import *clusters* and *feeds* modules (provided in lecture codes), and call functions from these modules as required similar to the examples that we have seen in the class.



Publication Analyzer v1.0

1 Please enter Google Scholar profile URLs (one URL per line:)

```
https://scholar.google.com/citations?user=IUMeXw4AAAAJ
https://scholar.google.com/citations?user=mBvj3CMAAAAJ
https://scholar.google.com/citations?user=puXkcxcAAAAJ
https://scholar.google.com/citations?user=vnmO6X4AAAAJ
https://scholar.google.com/citations?user=fU25R5gAAAAJ
https://scholar.google.com/citations?user=lJu0ljsAAAAJ
```

2 Download Publication Profiles

1. Ahmet Bulut (Downloaded)
2. Processing ...
3. Pending
4. Pending
5. Pending
6. Pending

View Publications for a Researcher

Choose a researcher: Ali Cakmak

Ahmet Bulut
Ali Cakmak
Serkan Apaydin
Onur Guzey
Hakan Dogan
Murat Kucukvar

Group by: ● Year ○ Type ○ None

Sort by: ● Year ○ Citation Count

Cluster Researchers

Clustering Method: ● Hierarchical ○ K-Means k: 5

```
-
 Murat Kucukvar
 -
   Hakan Dogan
   -
     Onur Guzey
     -
       Ahmet Bulut
       -
         Ali Cakmak
         Serkan Apaydin
```

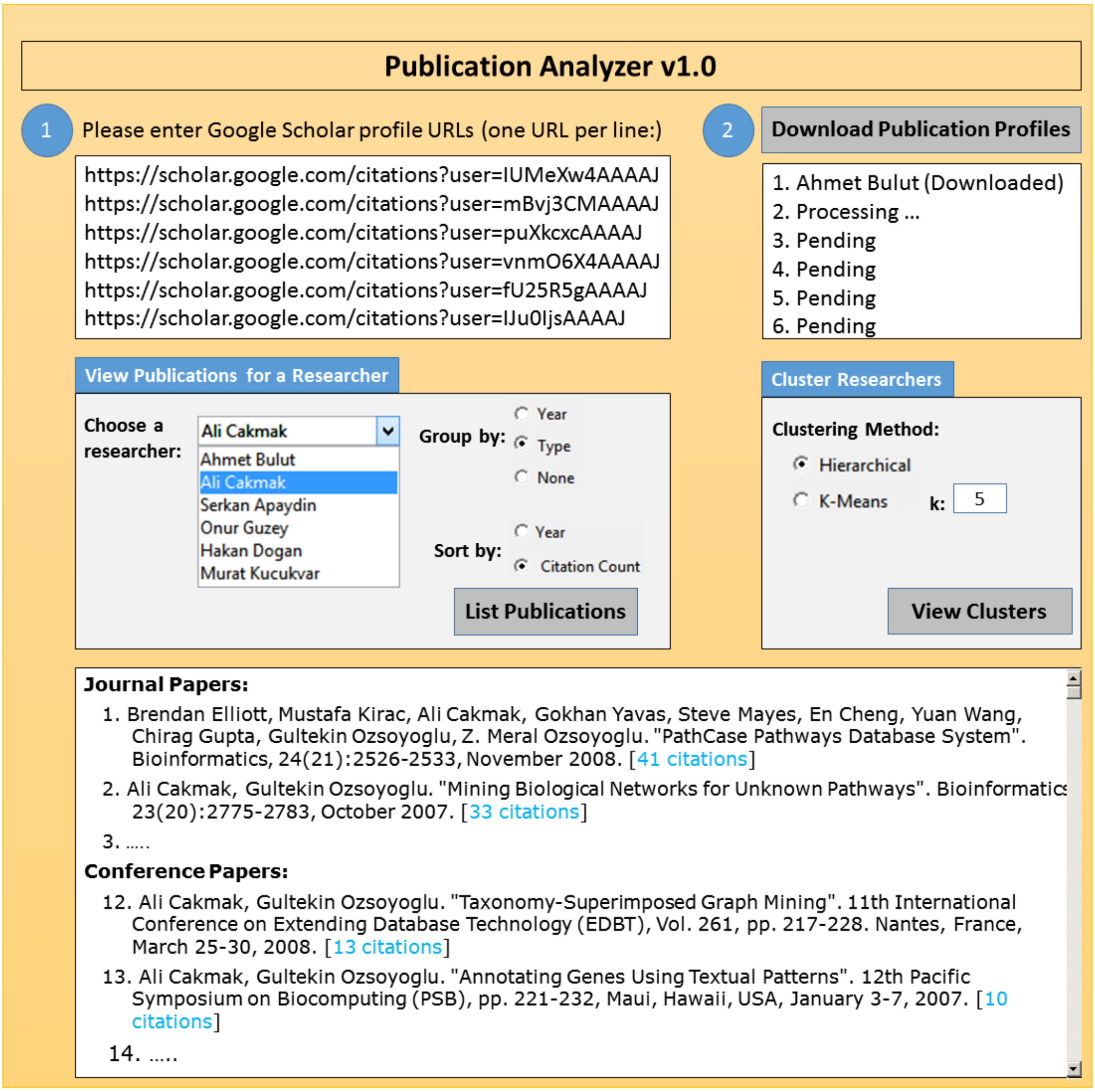
Figure 5

- **Viewing Publication List of a Researcher**: On the left-hand side of the middle row in Figure 4, a set of controls are provided for the users to view the publications of a researcher.
    - First the user selects a researcher from a combo box which is populated automatically by your program after extracting researcher name for each provided URL in step 1 and 2.
    - Users may choose to group publications by year (default value) or publication type. Alternatively, they may omit grouping altogether by choosing None here. You may assume that there may be four types of publications: Journal Papers, Conference Papers, Book Chapters, and Patents. You will identify the type of a publication when you parse the detail page of a publication, and see if there is a ‘Journal’ (Figure 2), ‘Conference’ (Figure 3), ‘Patent office’ (see this example), or ‘Book’ (see this example) field. If you cannot identify such fields in a publication (see this example), assume that its type is ‘Technical Report’, and list it under that category if group by ‘type’ is selected. Publication type-based groups should be listed in the following order: Journal Papers, Conference Papers, Book Chapters, Patents. If a researcher does

not have a publication under a group (say Patents), then that group should be omitted from the list.

- Users may choose to sort publications either by year (default value) or citation count. In the above field, if user choose a grouping other than None, then sorting should be done within each group. Examples are provided below.

- Once the user clicks on “List Publications” button, at the bottom, your program should show the publication list of the selected researcher as grouped and sorted by the user-selected options. Figure 6 provides a sample view when user chooses to view publications grouped by publication type and sorted by citation count. Note that sorting is applied within each group separately, that is, Journal Papers are sorted among themselves, and Conference Papers are sorted among themselves, and so on. Please also note that each publication in the list is numbered, and numbering continues across different groups. Dotted lines in the figure show omitted publications for brevity. Your program will not have dotted lines.


Publication Analyzer v1.0
1
Please enter Google Scholar profile URLs (one URL per line:)
2
Download Publication Profiles
View Publications for a Researcher
Cluster Researchers
List Publications
View Clusters

https://scholar.google.com/citations?user=IUMeXw4AAAAJ
https://scholar.google.com/citations?user=mBvj3CMAAAAJ
https://scholar.google.com/citations?user=puXkcxcAAAAJ
https://scholar.google.com/citations?user=vnmO6X4AAAAJ
https://scholar.google.com/citations?user=fU25R5gAAAAJ
https://scholar.google.com/citations?user=lJu0ljsAAAAJ

1. Ahmet Bulut (Downloaded)
2. Processing ...
3. Pending
4. Pending
5. Pending
6. Pending

Choose a researcher: Ali Cakmak

Ahmet Bulut
Ali Cakmak
Serkan Apaydin
Onur Guzey
Hakan Dogan
Murat Kucukvar

Group by: ○ Year ● Type ○ None

Sort by: ○ Year ● Citation Count

Clustering Method:

● Hierarchical

○ K-Means k: 5

Journal Papers:

1. Brendan Elliott, Mustafa Kirac, Ali Cakmak, Gokhan Yavas, Steve Mayes, En Cheng, Yuan Wang, Chirag Gupta, Gultekin Ozsoyoglu, Z. Meral Ozsoyoglu. "PathCase Pathways Database System". Bioinformatics, 24(21):2526-2533, November 2008. [41 citations]
2. Ali Cakmak, Gultekin Ozsoyoglu. "Mining Biological Networks for Unknown Pathways". Bioinformatic 23(20):2775-2783, October 2007. [33 citations]
3. .....

Conference Papers:

12. Ali Cakmak, Gultekin Ozsoyoglu. "Taxonomy-Superimposed Graph Mining". 11th International Conference on Extending Database Technology (EDBT), Vol. 261, pp. 217-228. Nantes, France, March 25-30, 2008. [13 citations]
13. Ali Cakmak, Gultekin Ozsoyoglu. "Annotating Genes Using Textual Patterns". 12th Pacific Symposium on Biocomputing (PSB), pp. 221-232, Maui, Hawaii, USA, January 3-7, 2007. [10 citations]
14. .....

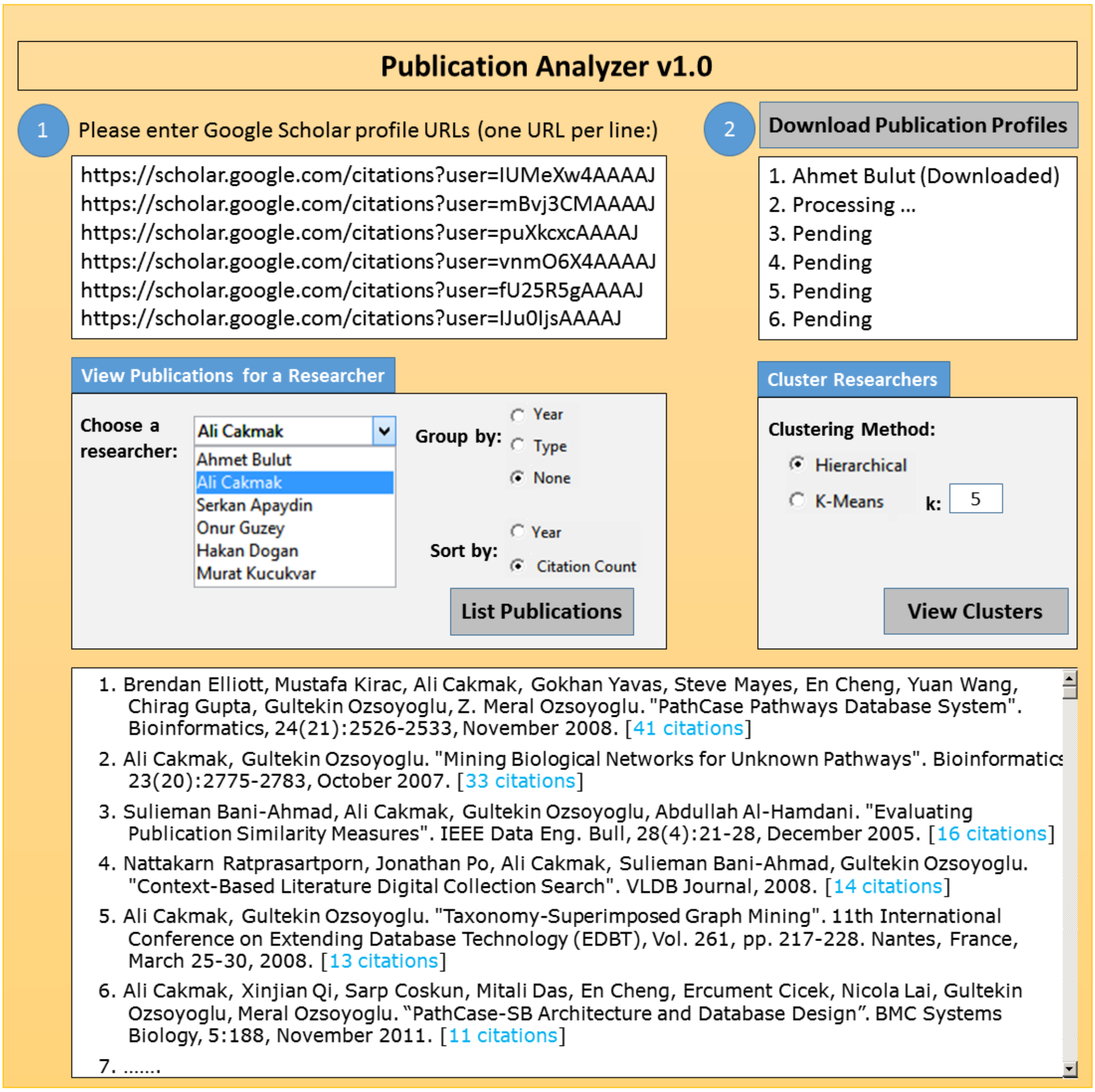
Figure 6

- Publication information should be constructed by your program by combining fields that you extracted from the corresponding publication detail page in the following order:

  {Authors}. {“Publication Title”}. {Publication Venue}, {Volume, if available}, {Issue, if available}{Page numbers, if available}, {Year}. [Citation Count].

  Please see Figures 6 through 8 for examples.

Figure 7 provides a sample view when user chooses to view publications without any grouping and sorted by citation count. Figure 8 provides a sample view when user chooses to view publications grouped by year and sorted by year.


Publication Analyzer v1.0
1
2
Download Publication Profiles
View Publications for a Researcher
Cluster Researchers
List Publications
View Clusters

Please enter Google Scholar profile URLs (one URL per line:)

https://scholar.google.com/citations?user=IUMeXw4AAAAJ
https://scholar.google.com/citations?user=mBvj3CMAAAAJ
https://scholar.google.com/citations?user=puXkcxcAAAAJ
https://scholar.google.com/citations?user=vnmO6X4AAAAJ
https://scholar.google.com/citations?user=fU25R5gAAAAJ
https://scholar.google.com/citations?user=lJu0ljsAAAAJ

1. Ahmet Bulut (Downloaded)
2. Processing ...
3. Pending
4. Pending
5. Pending
6. Pending

Choose a researcher: Ali Cakmak

Ahmet Bulut
Ali Cakmak
Serkan Apaydin
Onur Guzey
Hakan Dogan
Murat Kucukvar

Group by: ○ Year ○ Type ● None

Sort by: ○ Year ● Citation Count

Clustering Method:

● Hierarchical

○ K-Means k: 5

1. Brendan Elliott, Mustafa Kirac, Ali Cakmak, Gokhan Yavas, Steve Mayes, En Cheng, Yuan Wang, Chirag Gupta, Gultekin Ozsoyoglu, Z. Meral Ozsoyoglu. "PathCase Pathways Database System". Bioinformatics, 24(21):2526-2533, November 2008. [41 citations]
2. Ali Cakmak, Gultekin Ozsoyoglu. "Mining Biological Networks for Unknown Pathways". Bioinformatics 23(20):2775-2783, October 2007. [33 citations]
3. Sulieman Bani-Ahmad, Ali Cakmak, Gultekin Ozsoyoglu, Abdullah Al-Hamdani. "Evaluating Publication Similarity Measures". IEEE Data Eng. Bull, 28(4):21-28, December 2005. [16 citations]
4. Nattakarn Ratprasartporn, Jonathan Po, Ali Cakmak, Sulieman Bani-Ahmad, Gultekin Ozsoyoglu. "Context-Based Literature Digital Collection Search". VLDB Journal, 2008. [14 citations]
5. Ali Cakmak, Gultekin Ozsoyoglu. "Taxonomy-Superimposed Graph Mining". 11th International Conference on Extending Database Technology (EDBT), Vol. 261, pp. 217-228. Nantes, France, March 25-30, 2008. [13 citations]
6. Ali Cakmak, Xinjian Qi, Sarp Coskun, Mitali Das, En Cheng, Ercument Cicek, Nicola Lai, Gultekin Ozsoyoglu, Meral Ozsoyoglu. "PathCase-SB Architecture and Database Design". BMC Systems Biology, 5:188, November 2011. [11 citations]
7. ....

Figure 7

Can you provide any further pointers that may be helpful? :

- As for the GUI, you **should use** Tkinter that we have covered last semester (see ENGR 211 last week's slides). If you do not like Tkinter, you may use PyQt (we have not covered that in ENGR 211), but in any case, you are **not** allowed to use a designer or any other GUI module (other than the above ones).
- In order to be able to place widgets as shown in Figures, you should heavily use frames. Please see ENGR 211, last week's slides for creating row, col, and grid frames, and their example uses with Swampy.
- You may use the Text widget of Tkinter (slides 11, 12 in ENGR 211.Week 15) to take the list of URLs from the user in Step 1.
- In step 2, to show the progress to the user, you may the listbox widget of Tkinter. An example is provided below:

```
from swampy.Gui import *

g = Gui()
```

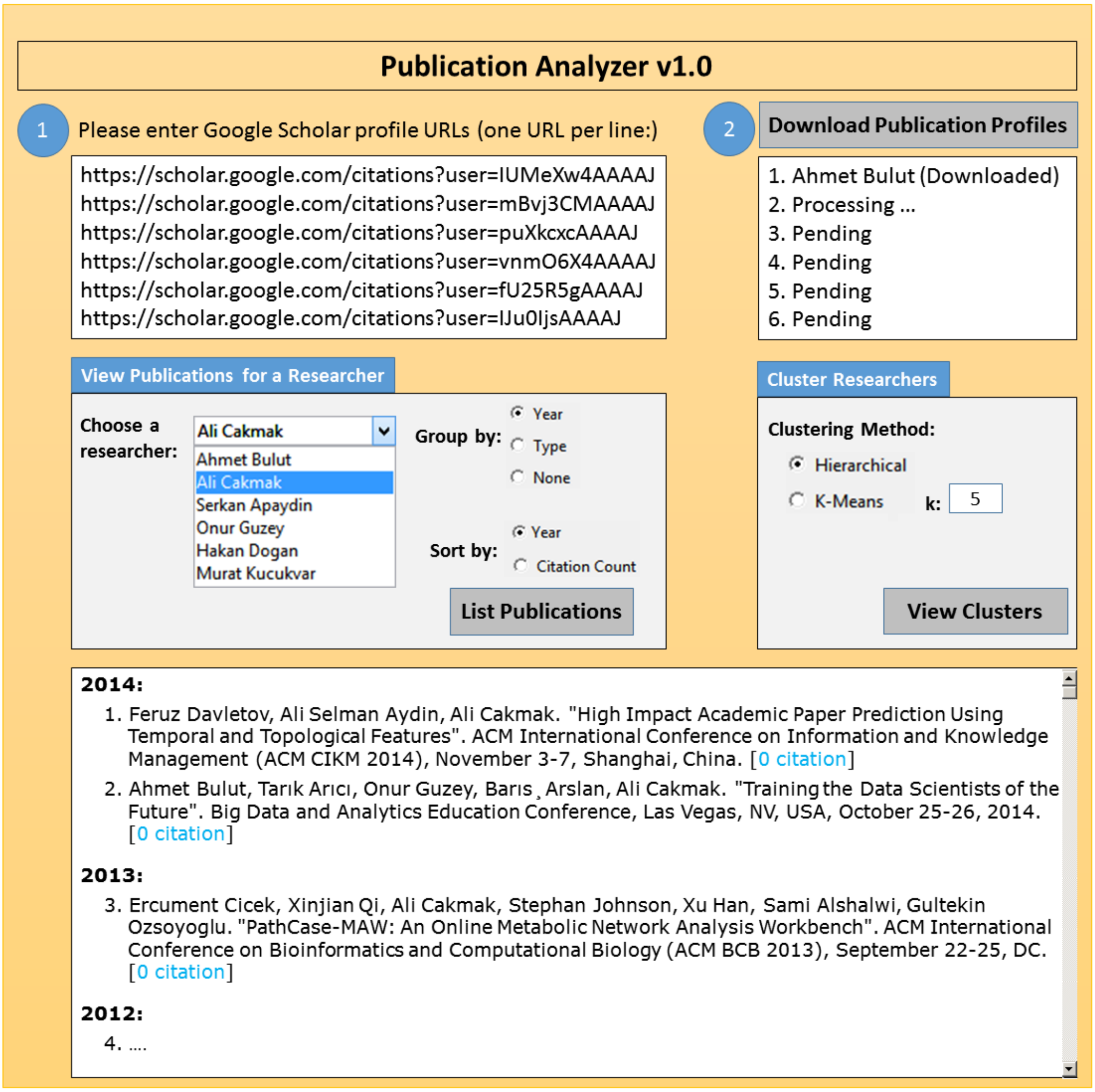
```
lb = g.lb()
for item in ["one", "two", "three", "four"]:
    lb.insert(END, item)
```

- For the group by, sort by, and clustering method options, you may use the RadioButton widget of Tkinter. The following example code pieces may help:

  http://effbot.org/tkinterbook/radiobutton.htm
  http://www.python-course.eu/tkinter_radiobuttons.php

- For researcher name selection in publication list viewing feature, you may use the ComboBox widget of Tkinter. The following example code piece may help:
  http://stackoverflow.com/questions/17757451/simple-ttk-combobox-demo

- For the output area at the bottom, again, you may use the Text widget.


Publication Analyzer v1.0
1
2
Download Publication Profiles
View Publications for a Researcher
Cluster Researchers
List Publications
View Clusters

Please enter Google Scholar profile URLs (one URL per line:)

https://scholar.google.com/citations?user=IUMeXw4AAAAJ
https://scholar.google.com/citations?user=mBvj3CMAAAAJ
https://scholar.google.com/citations?user=puXkcxcAAAAJ
https://scholar.google.com/citations?user=vnmO6X4AAAAJ
https://scholar.google.com/citations?user=fU25R5gAAAAJ
https://scholar.google.com/citations?user=lJu0ljsAAAAJ

1. Ahmet Bulut (Downloaded)
2. Processing ...
3. Pending
4. Pending
5. Pending
6. Pending

Choose a researcher: Ali Cakmak
Ahmet Bulut
Ali Cakmak
Serkan Apaydin
Onur Guzey
Hakan Dogan
Murat Kucukvar

Group by: ● Year ○ Type ○ None

Sort by: ● Year ○ Citation Count

Clustering Method:
● Hierarchical
○ K-Means k: 5

2014:

1. Feruz Davletov, Ali Selman Aydin, Ali Cakmak. "High Impact Academic Paper Prediction Using Temporal and Topological Features". ACM International Conference on Information and Knowledge Management (ACM CIKM 2014), November 3-7, Shanghai, China. [0 citation]
2. Ahmet Bulut, Tarık Arıcı, Onur Guzey, Barıs¸ Arslan, Ali Cakmak. "Training the Data Scientists of the Future". Big Data and Analytics Education Conference, Las Vegas, NV, USA, October 25-26, 2014. [0 citation]

2013:

3. Ercument Cicek, Xinjian Qi, Ali Cakmak, Stephan Johnson, Xu Han, Sami Alshalwi, Gultekin Ozsoyoglu. "PathCase-MAW: An Online Metabolic Network Analysis Workbench". ACM International Conference on Bioinformatics and Computational Biology (ACM BCB 2013), September 22-25, DC. [0 citation]

2012:

4. ....

Figure 8

Test URLs:

- The original URLs shown in the above figures are listed below. However, do not use these URLs until you make sure that your program is working fine. Otherwise, Google may block you for excessive access after a while. Instead, use the test links provided in the next bullet item during your development phase.

  https://scholar.google.com/citations?user=IUMeXw4AAAAJ
  https://scholar.google.com/citations?user=mBvj3CMAAAAJ

https://scholar.google.com/citations?user=puXkcxcAAAAJ
https://scholar.google.com/citations?user=vnmO6X4AAAAJ
https://scholar.google.com/citations?user=fU25R5gAAAAJ
https://scholar.google.com/citations?user=IJu0IjsAAAAJ

- Alternative URLs to be used during the initial development phase. These are exact copies of the above links. The below links are available at http://engr212.byethost10.com/

  http://engr212.byethost10.com/Scholar%20-%20Ahmet%20Bulut/scholar.google.com/citations2085.html
  http://engr212.byethost10.com/Scholar%20-%20Ali%20Cakmak/scholar.google.com.tr/citationsdd65.html
  http://engr212.byethost10.com/Scholar%20-%20Hakan%20Dogan/scholar.google.com/citations36e2.html
  http://engr212.byethost10.com/Scholar%20-%20Mehmet%20Serkan%20Apaydin/scholar.google.com/citationsb7b2.html
  http://engr212.byethost10.com/Scholar%20-%20Murat%20Kucukvar/scholar.google.com/citations1173.html
  http://engr212.byethost10.com/Scholar%20-%20Onur%20Guzey/scholar.google.com/citations748c.html

How and when do I submit my project? :

- Projects may be done individually or as a small group of two students. If you are doing it as a group, only **one** of the members should submit the project. File name will tell us group members (Please see the next item for details).
- Submit your own code in a **single** Python file (Do **not** include clusters.py or feeds.py that you import). Name it with your and your partner's first and last names (see below for naming).
  - If your team members are Deniz Barış and Ahmet Çalışkan, then name your code file as deniz_baris_ahmet_caliskan.py (Do **not** use any Turkish characters in file name).
  - If you are doing the project alone, then name it with your name and last name similar to the above naming scheme.
- Do **not** copy/paste code from clusters.py or feeds.py into your own code file. Anything that you need from clusters.py or feeds.py should be called with proper dot notation after importing that module.
- Do **not** use any external module other than BeautifulSoup and Swampy, which are not included in standard Python installation.
- Do **not** use Python 3.x. Use Python 2.7.x.
- Submit it online on LMS (Go to the Assignments Tab) by **17:00 on April 16 (Thursday)**.

  Late Submission Policy:

  - -20%: Submissions between 17:01 – midnight (00:00) on the due.
  - -40%: Submissions which are 24 hour late.
  - -50%: Submissions which are 48 hours late.
  - Submission more than 48 hours late will not be accepted.

Grading Criteria? :

- Does it run? (Submissions that do not run will get some partial credit which will not exceed 30% of the overall project grade).
- Does it implement all the features according to the specifications and produce correct results?
- Code organization (Meaningful names, sufficient and appropriate comments, proper organization into functions and classes, clean and understandable, etc.)?
- Interview evaluation.

Have further questions? :

- Please contact your TAs (Mehmet Aytimur and Muhammed Esad Unal) if you have further questions. If you need help with anything, **please use the office hours** of your TAs and the instructor to get help. If office hours are not suitable, please **get** an appointment through email before walking in your TAs offices.