ENGR 212 Programming Practice
Mini Project 3
November 14, 2016

In this project, you are going to group together researchers that presented papers in the AAAI (*Association for the Advancement of Artificial Intelligence*) Conference on Artificial Intelligence, in 2014. You will be given a data file listing the accepted publications to the conference. The format of the given file that holds this information is as follows:

title,authors,groups,keywords,topics,abstract

where ***title*** is the title of the paper presented by the researchers, ***authors*** is the list of authors for that publication (if there are two authors their names are separated by 'and'; if there are more than two authors the names are enclosed in double quotation marks (" "), and each one separated by comma except the last one is separated by 'and'), ***groups*** is a list of the categorical, author selected high-level keywords representing the scientific groups the publication belongs to (each group has an abbreviation given after the group text within parenthesis; if there are more than one group listed they are separated by a newline character, and whole list is enclosed in in double quotation marks (" ")), ***keywords*** is a list of author generated free text representing the publication (if there are more than one keyword listed they are separated by a newline character, and whole list is enclosed in double quotation marks (" ")), ***topics*** is another list of free text author selected low-level keywords (if there are more than one topic listed they are separated by a newline character, and whole list is enclosed in in double quotation marks (" ")), and ***abstract*** is the free-text abstract of the paper. Each part is separated by a comma. Note that for each part which contains a comma in itself, the whole part is also enclosed with double quotes. Two consecutive double quotes ("") within a quoted string on the other hand, does not specify the start or end of a part but used to enclose a quoted string within a part (i.e., one of those pair of quotes should be dropped while processing such strings). An example is as follows:

> Lifetime Lexical Variation in Social Media,"Liao Lizi, Jing Jiang, Ying Ding, Heyan Huang and Ee-Peng Lim",NLP and Text Mining (NLPTM), "Generative model Social Networks Age Prediction","AIW: Web personalization and user modeling
> NLPTM: Information Extraction
> NLPTM: Natural Language Processing (General/Other)","As the rapid growth of online social media attracts a large number of Internet users, the large volume of content generated by these users also provides us with an opportunity to study the lexical variations of people of different age. In this paper, we present a latent variable model that jointly models the lexical content of tweets and Twitter users' age. Our model inherently assumes that a topic has not only a word distribution but also an age distribution. We propose a Gibbs-EM algorithm to perform inference on our model. Empirical evaluation shows that our model can generate meaningful age-specific topics such as ""school"" for teenagers and ""health"" for older people. Our model also performs age prediction better than a number of baseline methods."

For the above example, the related fields are:

title : Lifetime Lexical Variation in Social Media
authors : Liao Lizi, Jing Jiang, Ying Ding, Heyan Huang, Ee-Peng Lim
groups : NLP, Text Mining (NLPTM)
keywords : Generative model Social Networks Age Prediction
topics : Web personalization and user modeling
NLPTM: Information Extraction
NLPTM: Natural Language Processing (General/Other)
abstract : As the rapid growth of online social media attracts……. age-specific topics such as "school" for teenagers and "health" for older people. Our model also performs age prediction better than a number of baseline methods.

Here, your program should be able to extract the above field information for each of the accepted publications to the conference. Note that several lists (author list, group lists) contain the "and" word. You should eliminate these "and" words, and store proposition clear texts as part of the field information (i.e., your author lists should NOT contain an author name starting or ending with "and"!).

1. Your program will have a graphical user interface (GUI) which will look like as shown in Figure1. Details about how it should work are provided below.

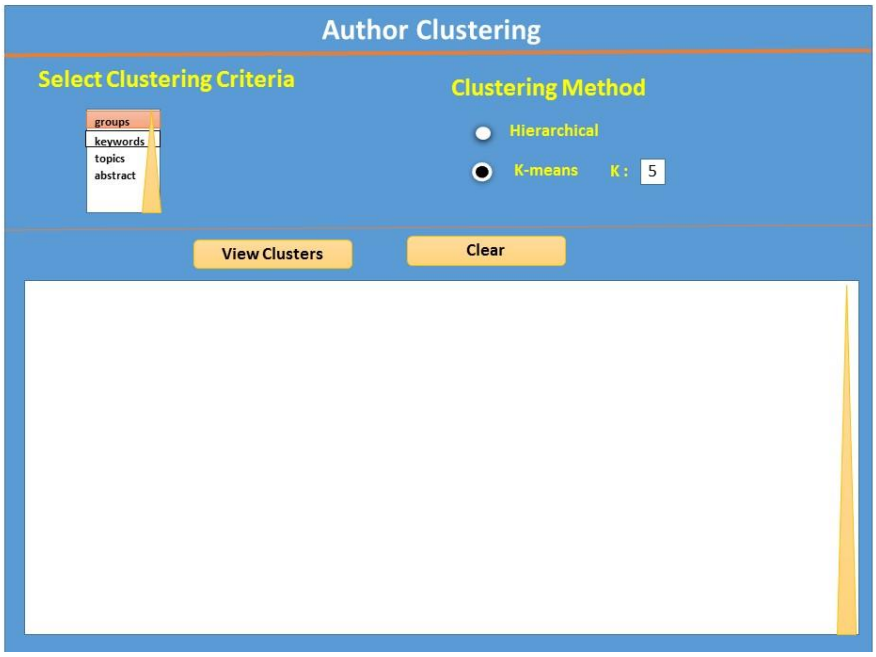

Figure 1

- Your program should load the input file mentioned above upon startup, and be ready to process it according to the specifications and user selections given below.

- In the next section on the GUI, there should be a combo-box widget where the user can select the criteria on which the clustering shall be based. There are four options; clustering based on groups, keywords, topics, or abstract. Selection of each one should generate the appropriate dataset suitable for clustering. You should prepare a matrix similar to the ones in seen in class, "blogdata.txt", or "zebo.txt"; rows should be author names, columns depend on the criteria chosen. Please pay special attention to categorical data such as groups as opposed to free-text data for keywords, topics, and abstract. You should prepare your dataset suitable for clustering appropriately.

- After creating the data matrix, write it to a file with name "groupsData.txt",

- "keywordsData.txt", "topicsData.txt", or "abstractData.txt", depending on the criteria chosen.

- Then the user should have the option of selecting between Hierarchical Clustering, and k-means clustering. In case of k-means clustering, users should provide the value of k, as well (by default it will be 5). In any case, the similarity metric to be used should be carefully selected

you depending on the criteria selected by the user (Hint: you should use appropriate similarity metrics for category-based, and free-text clustering options)

- When "view clusters" button is clicked, then user **should** see the result of the clustering at the bottom part. For hierarchical clustering, results should be shown as a string dendogram (Figure 2 shows an example. Here, you are going to create a similar visualization you obtained by using the "printclust" function in clusters module, but you should modify the function in order to print the results onto the text widget), while for k-means clustering, the results will be shown in text form by listing each member of each clusters (see Figure 3 for an example). Your test box should have horizontal and vertical scrollbars.
- As a validation for your solution, all co-authors of a publication should always be in the same cluster, irrespective of the criteria or clustering method selected. They can be clustered with other people, of course.
- When the user presses the "Clear" button, the text area should be cleared from any text present.

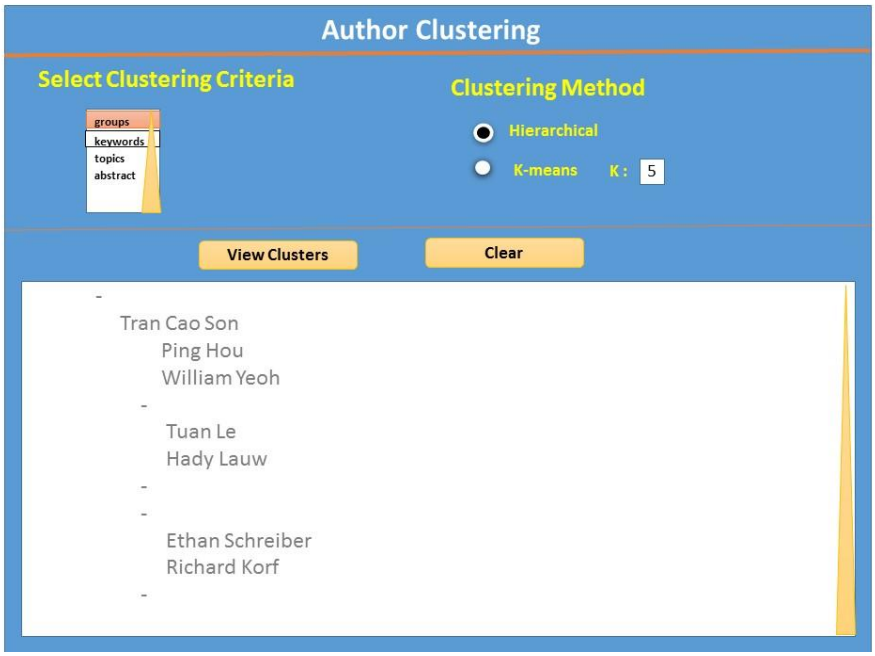

Figure 2

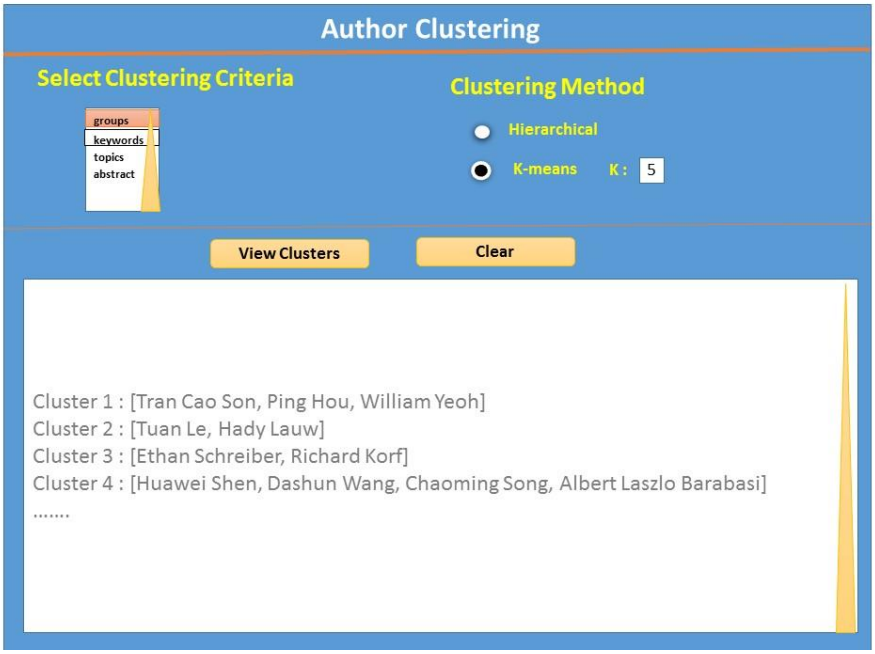

Figure 3

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

- For clustering criteria selection you can use the ComboBox widget of Tkinter. The following example code piece may help:
  http://stackoverflow.com/questions/17757451/simple-ttk-combobox-demo

- For the clustering method selection, 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

Warnings:

- **Do not** talk to your classmates on project topics when you are implementing your projects. **Do not** show or email your code to others. If you need help, talk to your TAs or myself, not to your classmates. If somebody asks you for help, explain them the lecture slides, but do not explain any project related topic or solution. Any similarity in your source codes will have **serious** consequences for both parties.
- Carefully read the project document, and pay special attention to sentences that involve "**should**", "**should not**", "**do not**", and other underlined/bold font statements.
- If you use code from a resource (web site, book, etc.), make sure that you reference those resource at the top of your source code file in the form of comments. You should give details of which part of your code is from what resource. Failing to do so **may result in** plagiarism investigation.
- Even if you work as a group of two students, each member of the team should know every line of the code well. Hence, it is **important** to understand all the details in your submitted code. You may be interviewed about any part of your code.

How and when do I submit my project? :

- Projects may be done individually or as a small group of two students (doing it individually is recommended for best learning experience). 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 clusterys.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.
  - Those who **do not** follow the above naming conventions **will get -5 off** of their project grade.
- Submit it online on LMS (Go to the Assignments Tab) by **17:00 on November 27, 2016**.

  Late Submission Policy:

  - -10%: Submissions between 17:01 – 18:00 on the due date
  - -20%: Submissions between 18:01 – midnight (00:00) on the due date
  - -30%: 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? :

Code Organization			Functionality					
Meaningful variable names (%5)	Classes and objects used (%5)	Sufficient commenting (%5)	Compiles % Runs? (25)	GUI Design (10)	Reading dataset / Populating matrix data (30)	Selection of appropriate similarity metrics (20)	Printing hiearhical clustering properly (15)	Printing k-means clustering properly (15)

- Interview evaluation (your grade from interview will be between 0 and 1, and it will be used as a coefficient to compute your final grade. For instance, if your initial grade was 80 before the interview, and your interview grade is 0.5, then your final grade will be 80*0.5 = 40). Not showing up for the interview appointment will **result in** grade 0.

Have further questions? :

- Contact your TA's during their office hours, and please do NOT hesitate to ask the instructor (me) also (by e-mail or in-person), about anything related to the project.

Good Luck!