

Bilkent University
Computer Engineering Department
CS533 Information Retrieval Systems, Spring 2014

Assignment No: 6

Due Date: until 23:59PM, 31.05.2014

(Please contact ctoraman@cs.bilkent.edu.tr for any assistance and questions.)

Front-page Importance & Novelty Annotation

1. Background

News portals provide huge amounts of news obtained from various sources to news readers (e.g Bilkent News Portal, <http://139.179.21.201/PortalTest/index.php>). Among these news, some of them are displayed in the **front-page** of news portal (i.e main-page). **Important news article** is defined as a news article that should be displayed in the front-page. This decision is upto news reader's interest.

While having important news in front-page, it is also crucial to present news articles that are member of different news topics. A news **topic** is defined as a set of news articles strongly related to a real life event. For instance, "flood disaster in İstanbul" is a news topic. Some example news articles in this topic are titled "consequences of flood disaster in İstanbul", " government mistakes on Istanbul flood disaster", and "emergency relief by social associations".

A news article is defined as **novel**, in this assignment, if it is a member of a new topic that has not been provided by other news articles in the front-page. For instance, let be three documents d_1 , d_2 and d_3 with titles "31 people died by flood disaster in İstanbul", "prime minister of Turkey responds to allegations on flood disaster", "Turkish prime minister plays down threat to Iraq", respectively. d_1 and d_2 are in the same topic- "flood disaster in Istanbul" whereas d_3 is a member of another topic. Note that d_2 and d_3 are in the same category- "politics", which means that *news articles belonging to the same category could be members of different topics*.

2. Tasks

In this assignment, you have two main tasks. You are going to;

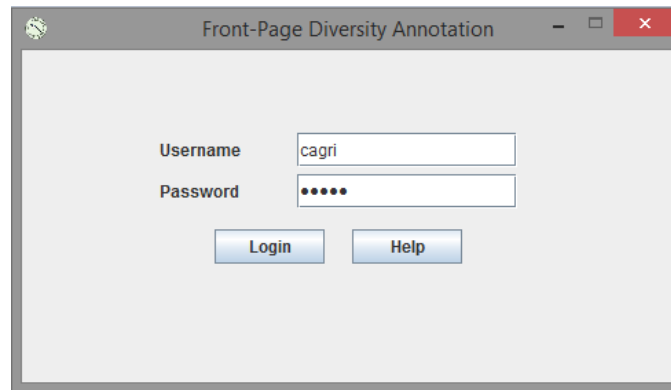
- (i) decide whether given news articles are important or not based on your own interests, and
- (ii) determine which news article belongs to which topic, i.e *novelty* of a front-page. However, there will be no pre-defined topics in this assignment. Instead, you are going to put each news article into a *virtual cluster* whose aim is to collect news articles belonging to the same topic. We call them virtual clusters since they do not have any title that describes the topic.

3. Annotation Program

You are going to complete your tasks via the annotation program that can be downloaded from (<http://cs.bilkent.edu.tr/~ctoraman/annotation.rar>). You need Internet connection to login into the annotation program. **You must connect inside Bilkent Campus or via Bilkent VPN.**

3.1 Login Screen

Each annotator has an username and password to login. Username and passwords will be sent to your Bilkent e-mail accounts. Figure 1 is a sample login screen.



Front-Page Diversity Annotation

Username: cagri

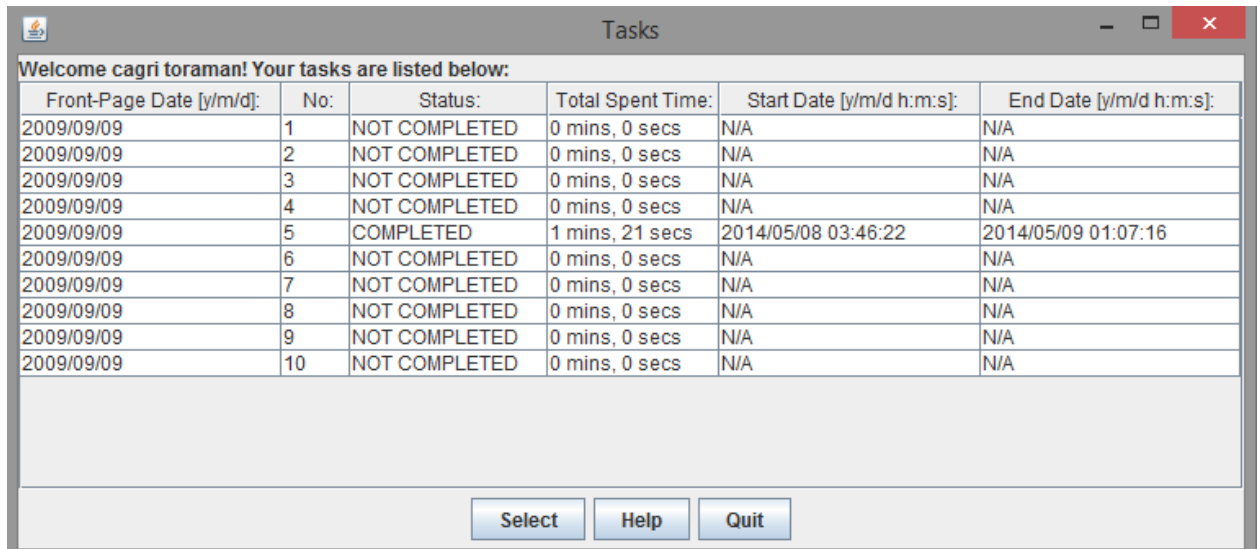
Password:

Login Help

Figure 1. A sample login screen.

3.2 Tasks Screen

Each annotator has a set of tasks to complete. Each task includes importance & novelty annotations of a front-page that has 10 different news articles. Each front-page belongs to a specific date in "year/month/day" format. Each day has different version numbers ("No"). Annotation status is given as "completed"/"not completed". Start and end of annotation process are given in date format. The time spent during annotation process is calculated as well. Note that you can quit your annotation process and continue from the condition you left whenever you want. You can also change your "completed" annotation whenever you want. Figure 2 displays a sample tasks screen.



Welcome cagri toraman! Your tasks are listed below:

Front-Page Date [y/m/d]:	No:	Status:	Total Spent Time:	Start Date [y/m/d h:m:s]:	End Date [y/m/d h:m:s]:
2009/09/09	1	NOT COMPLETED	0 mins, 0 secs	N/A	N/A
2009/09/09	2	NOT COMPLETED	0 mins, 0 secs	N/A	N/A
2009/09/09	3	NOT COMPLETED	0 mins, 0 secs	N/A	N/A
2009/09/09	4	NOT COMPLETED	0 mins, 0 secs	N/A	N/A
2009/09/09	5	COMPLETED	1 mins, 21 secs	2014/05/08 03:46:22	2014/05/09 01:07:16
2009/09/09	6	NOT COMPLETED	0 mins, 0 secs	N/A	N/A
2009/09/09	7	NOT COMPLETED	0 mins, 0 secs	N/A	N/A
2009/09/09	8	NOT COMPLETED	0 mins, 0 secs	N/A	N/A
2009/09/09	9	NOT COMPLETED	0 mins, 0 secs	N/A	N/A
2009/09/09	10	NOT COMPLETED	0 mins, 0 secs	N/A	N/A

Select Help Quit

Figure 2. A sample tasks screen.

3.3 Annotation Screen

A sample annotation screen corresponding to 2009/09/09 is given in Figure 3. This screen consists of three sub-panels:

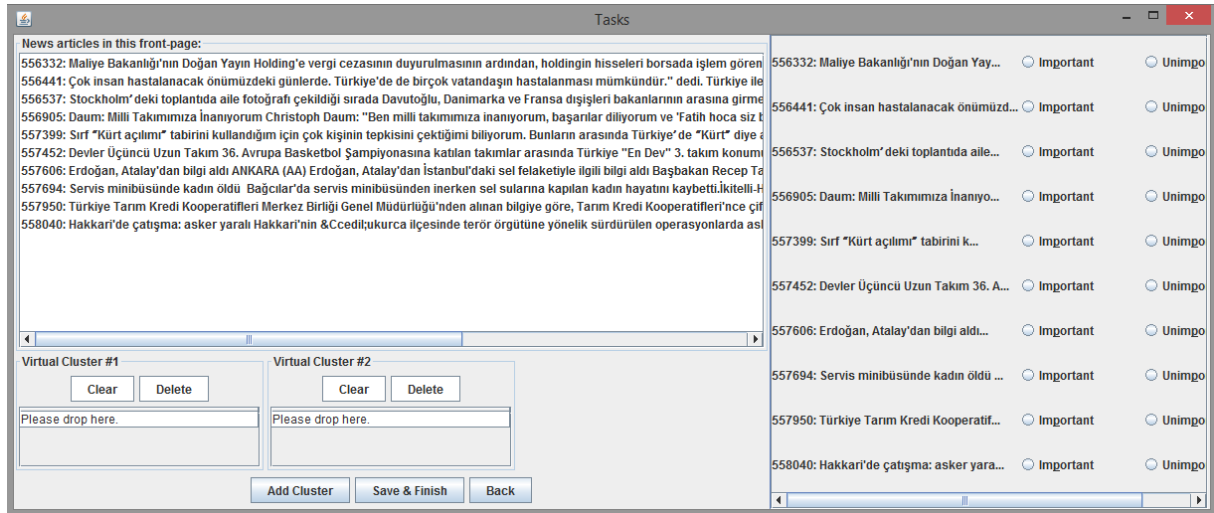


Figure 3. A sample annotation screen.

(i) *Top-left panel- News Articles:*

News articles in front-page are listed in the top left screen. News snippets (first 200 characters) are given starting with the news id in this list. If you double-click on a news snippet, then full news content will be displayed.

(ii) *Right panel- Importance Annotation (Task 1):*

Your first task, importance annotation, will be done at the right panel. Each line corresponds to a news article as in the same order with top-left panel. When you click on a news article in the top-left panel, then its corresponding line at the right panel will be colored as yellow. You have to decide if a news article is important or unimportant. **Recall that important news article is defined as a news article that should be displayed in the front-page and this decision is upto news reader's interest. Length of news article does not affect news importance. Also there may be noise in some news, please neglect such noise and consider true content before deciding.**

(iii) *Bottom-left panel- Novelty Annotation (Task 2):*

Your second task, novelty annotation, will be done in the bottom-left panel. Here, you will see virtual clusters with corresponding ids, clear and delete buttons. Clear button make the cluster empty while delete button removes it from screen. In the given sample screen, there are two empty virtual clusters.

Recall that your task is to determine which news article belongs to which virtual cluster (topic). You can select a news article in the top-left panel by left click on its snippet, then *drag* it over screen, and then *drop* it onto any virtual cluster's first line that says "Please drop here." **Note that you have to first determine a news article's importance at the right panel before drag-drop process.** When you drop a news article, then it will be removed from the top-left panel and also its corresponding line at the right panel will be colored as dark gray. If you would like to undo your annotation, then you can drag-drop news article into the top-left panel.

At the bottom of this panel, there are three buttons: "Add Cluster" is used for adding new virtual clusters if necessary. If you add a cluster wrongly or change your mind, you can delete it with "Delete" button at the bottom of corresponding cluster. "Save & Finish" button should be used when all news articles are annotated and no empty virtual clusters exist. By clicking this button, your

annotation task will be safely completed; however you can still change your completed annotation from tasks screen. "Back button" cancels current annotations in this front-page and returns to tasks screen.

3.4 Time Length Screen:

Total time spent during your annotation is asked in this screen. Please do not consider your breaks when annotation screen is open. You can give an approximate number in mins using the slider. If it takes more than 10 mins, please select 10 mins.

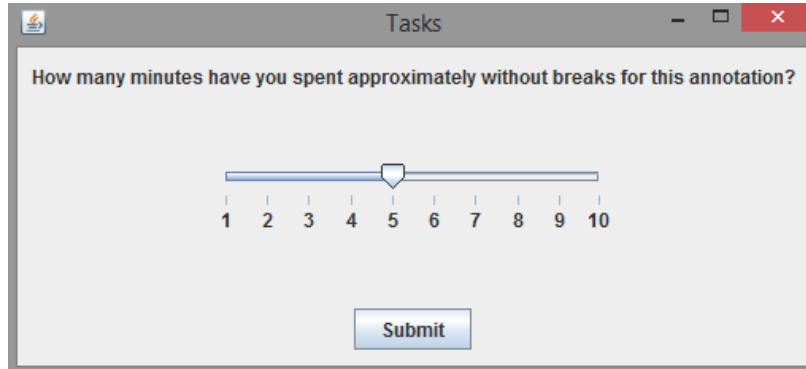


Figure 4. Annotation time length screen.

3.5 Example Annotation:

Let's do a sample annotation for the given sample front-page in Figure 3. I will annotate first news article in the top-left panel in three separate steps. Then, other news will be annotated without showing their steps explicitly.

Step 1 - Read news in the top-left panel: First news article in the top-left list is clearly about punishment of Doğan Yayın Holding for tax evasion. You can read full content by double-click on snippet to decide aboutness of news content (Figure 5). **Length does not affect news importance.** Also there may be noise in some news, please neglect such noise and consider true content before deciding.

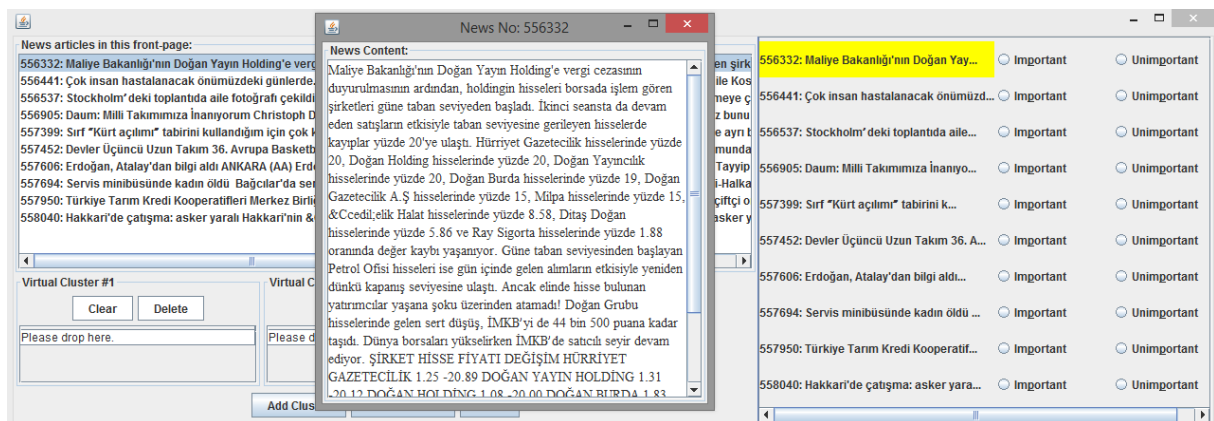


Figure 5. Double click on news snippet to read full news content.

Step 2 - Decide importance at the right panel: Before deciding on its virtual cluster, we have to determine its importance at the right panel. As a news reader who is interested in politics and media, I think "punishment of Doğan Yayın" is important, then I select important button for this news article at the right panel (Figure 6).

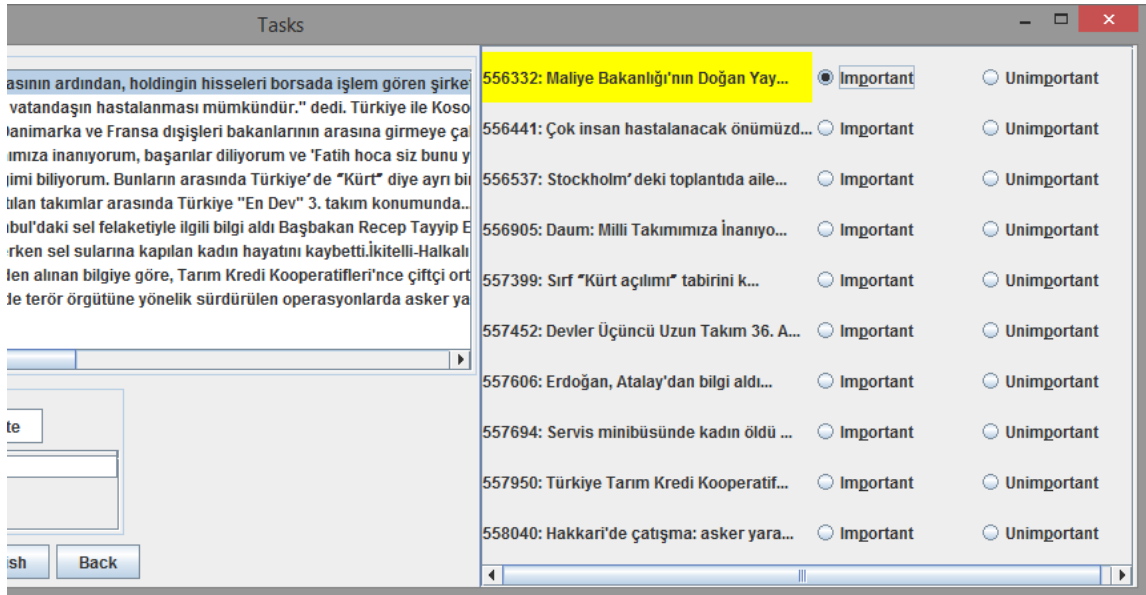


Figure 6. Determine news importance at the right panel.

Step 3 - Decide novelty in the bottom-left panel: Since there are no previously annotated news in the down-left virtual clusters, we can put this news article into first topic cluster by drag-drop from top-left panel to bottom-left panel (Figure 7). Since we annotate its both importance (right) and novelty (bottom-left), it is removed from the top-left list and also colored as dark gray at the right panel.

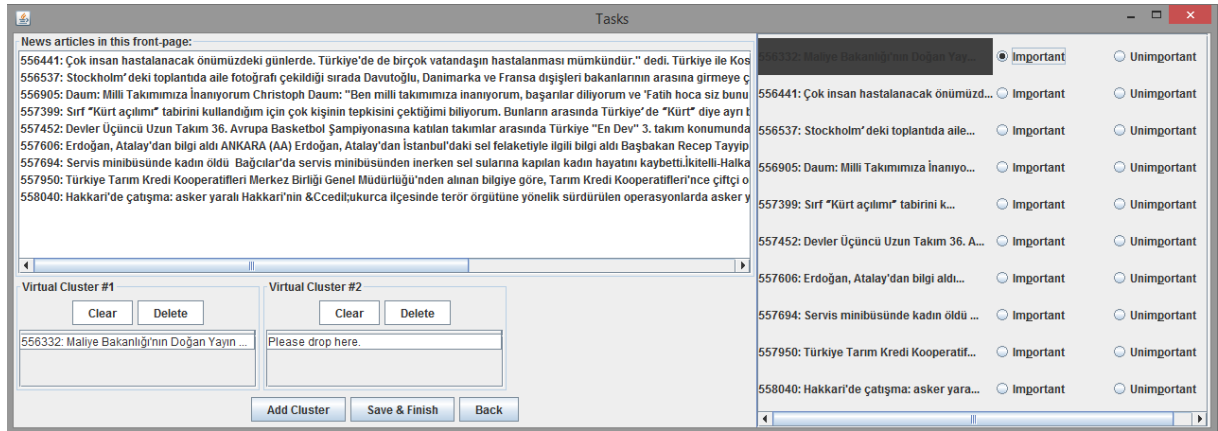


Figure 7. Determine novelty in the bottom-left panel.

After annotating first news article about "Doğan Yayın punishment", we can continue with second news, which leads currently in the top-left list. This news article is about swine flu and I think it is important. Since we previously put news about "Doğan Yayın punishment" into first virtual cluster, we can not put swine flu news into the same cluster. Instead, we put it into the next empty cluster (Figure 8).

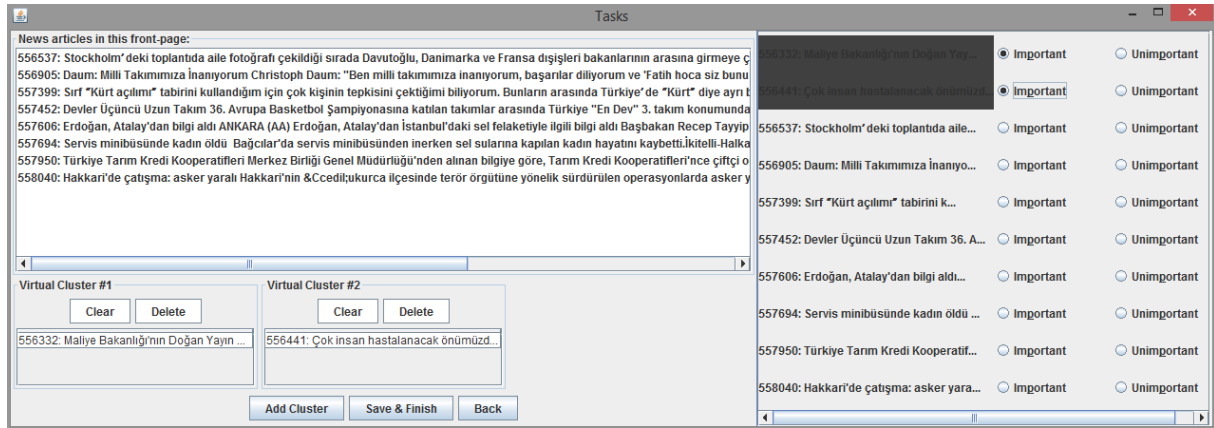


Figure 8. After annotation of second news article.

Third news article is about relations between Turkey and EU. I think its content is not so much important; but still can be put into front-page; thus, I select important. First two clusters are about other topics, then it should be dropped into another virtual cluster. You can add new virtual cluster by clicking "Add Cluster" button at the bottom of screen (Figure 9).

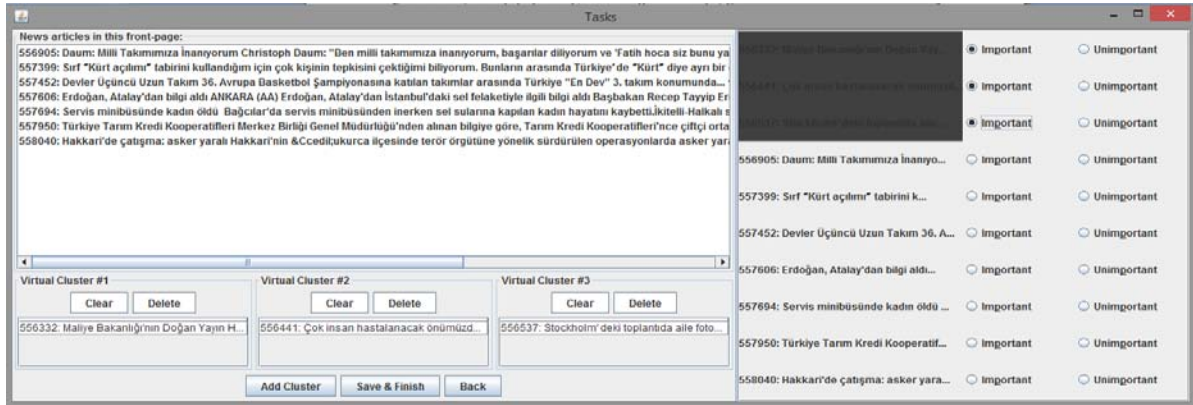


Figure 9. After annotation of third news article.

When other news articles are read, it is clearly seen that only news articles with id 557606 and 557694 should be in the same virtual cluster since they are both about flood disaster in İstanbul. Thus, there are 9 virtual clusters at the end of annotation process. Also 6 of 10 news articles are selected as important at the end of example annotation (Figure 10). Now, we can safely click "Save & Finish" button to finish our annotation. If you do not click on that button and exit, then your annotations will not be saved.

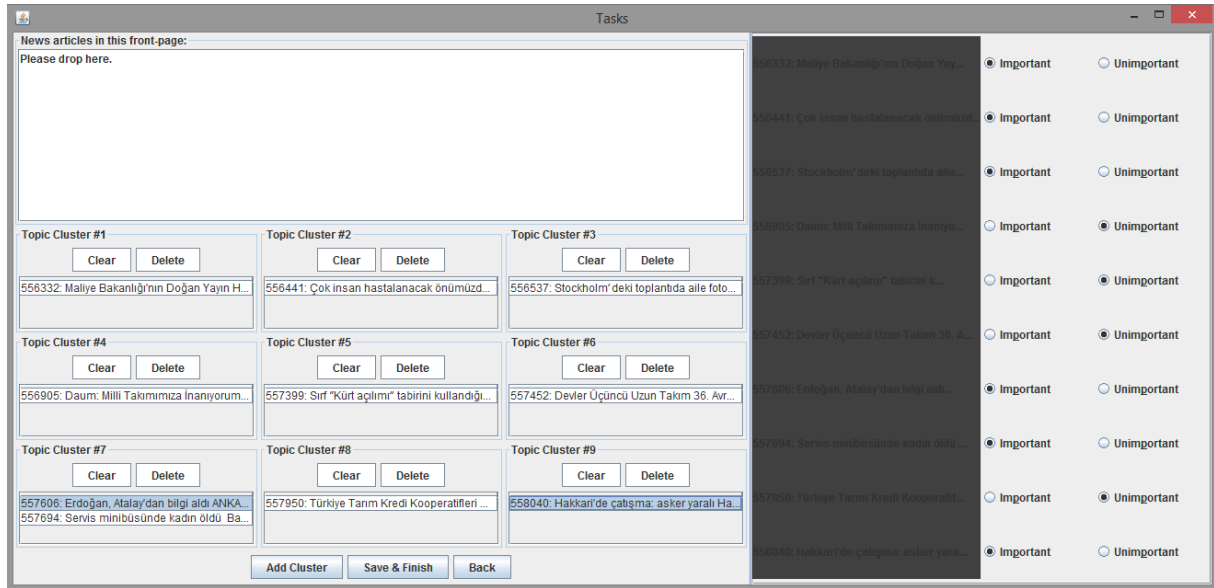


Figure 10. There are 9 clusters and 6 of 10 news are selected as important at the end of example annotation.

4. Conclusion

This assignment aims to explain front-page importance and novelty annotation process. **Your assignment will be finished when status of all tasks become "COMPLETED". Accuracy of your annotations (i.e correctness and attention you give) will be scored.** Each annotation task will be repeated by at least two different annotators. Do not remove the annotation program from your computer immediately after finishing, since there may be additional annotations.

Thanks for supporting our research!