Introduction to Natural Language Processing (NLP)

Semester Project - NER

Hi, this is Dr. Asan Abdullah with the course Introduction to Natural Language Processing and this is the semester project which is based upon the Named Entity Recognition. The videos for Named Entity Recognition and the corresponding material is available on Canvas and is also available on my YouTube channel. Okay. So, what is the project about?

1 sec.

So, these are the details. So, the project can be a group project if you feel that more people are required, but not more than three students per group. Now, the project covers the list of 10 domains. Of course, you will only work on one domain. So, that list of domains is given on the next page. So, you select three domains out of those 10 and email to the instructor the domain names and the SIS ID

31 sec.

of the group members if applicable. So the instructor will confirm final domain assignments so that the domains are distributed across the class. So Phase-I of the project is due on the 21st of February, 50% weight, and implement the semester project as shown in the video. Phase-I is as shown in the video, which is towards the end of this video. And for NER, you can preferably use a model.

1 min. 1 sec.

okay but you can also use dictionary or rules or their combinations they are acceptable but the accuracy should be good then is the phase two of the semester project which is due on the 14th of march carries 20 percent weight so in addition to Phase-I this phase i.e. Phase-II you will add an open button right to read a file 1200 to 1500 words that could be in MS word format

1 min. 31 sec.

File could be PDF format and display the contents of the file in the text box as was the case in Phase-I. And the NER results, okay, because now it's a big file, so there should be another text box in which the NER tagged file is displayed. And there's a save button. Pressing on that button saves this NER file, okay, name entity recognized file as a MS Word file back to the disk. All right.

2 min.

So then is the Phase-III, which is due towards the end of the semester. So in addition to Phase-II, what we need to do is calculate the vital statistics. Vital statistics. So the list of those vital statistics is given on the last page of this presentation. And the details are given on Canvas. Already there. And when this stat button is pressed, those statistics should be displayed. Okay.

2 min. 29 sec.

and submit all these three phases right as per the MS word template which is already communicated to you and the file name of the template should start with say for example Phase-I id phase one phase two phase three and so on all right so let's move ahead so these are the project domains so for example you select say for example you select this one and you select this one and you select this one right and some other person all right

3 min. 1 sec.

Some other persons may have a similar selection. Some person may select this one and may also select this one, right? And select this one. So, we'll try to ensure that most of the domains are covered, okay? And the choice and your choices are accommodated. Try to do that, but cannot be guaranteed, right? So, this is why there are three choices. All right. So, after then you communicate the three domains and one is assigned to you.

3 min. 30 sec.

And the Phase-III is the key document statistics. So out of these nine, you can work with any five of them. Okay, any five of them. For example, entity count and distribution. So entity count, as you see, there are five entities. Distribution is that entity one occurs ten times, entity two occurs five times or whatever. And the entity type distribution, entity type distribution is that how many occurrences of this type of entity and this type of entity and so on.

4 min.

So this detail is already given their definitions on canvas. So that's all I have for the semester project and now the video for phase one. Thank you very much. So this is the Phase-I work which is expected from you. Of course, you can make certain changes in the display, but this display is very nice. So you are welcome to use Spacy, okay? And which has pre-trained models also.

4 min. 28 sec.

But of course, what you will be doing will be as per the domain you have selected. All right. And corresponding to the domain, you will have the corresponding entities here. For example, if it's healthcare, then over here that there would be an entity by the name of medication, entity by the name of treatment, and so on. And if this is, say, for example, legal, then of course, you will have the corresponding entities here. All right.

4 min. 57 sec.

so let me paste some text here okay so this is the text this is all phase one so if I click on this icon so it has not done any tagging why because no entity was selected right unless I forget so you have to just do it for English language so let me select an entity I select geographical location and you see that it was tagged and a color was also assigned

5 min. 26 sec.

right and then I select the entity of cardinal and you can see that it was tagged color assigned there's an error because this is a phone number but it was broken into two pieces and this piece was not selected and of course there are other entities which are listed here but they are not present here for example if I select on date so there is no date here if I select on time there is no time here okay so now uh let's move ahead and I select person

5 min. 55 sec.

Now when I select on person, you can see that John and Gail Robbins Center for Academic Excellence and Innovation has been broken and John is considered to be a person while in fact this entire is an organization. So this is an error, right? So try to minimize the errors, right? And of course, if the selection is all is made, then all the entities are activated.

6 min. 24 sec.

and you can see the labeling and if I select again click on it, then everything is gone. So this is Phase-I. So in Phase-II when you load a file that file will be displayed here the contents. Okay with this scroll bar and the tag reserves will be here. There can be a box here which again has a vertical scroll bars and then a button to read the file contents and then the button to write the file as a word file to the disk.

6 min. 53 sec.

and so on so that's all thank you very much

7 min. 23 sec.