First of all; location, organization and person files are opened and written into a list with their first letter capitalized and others are lowercased.

Then input file is separated according to new line character to indicate the line count.

Then for every line, words are evaluated according to the following ways:

1. Any kind of school type (İlk, Orta,Yüksek Okul or Lise) with including punctuation and possible adding. ORGANIZATION
2. Looking for the word “yıl” to match year integers with 4 digits DATE
3. Checking for academic title Dr. PERSON
4. Looking for the word “başkanı” to find Başkanı, Cumhurbaşkanı, xxbaşkanı yybaşkanı etc. PERSON.

Matches for PERSON type also returns the title of the PERSON.

For example: “Cumhurbaşkanımız Ali Veli ALABIYIK” will be matched as a whole including the word “Cumhurbaşkanımız”

1. Other PERSON types such as “bakanı, vali, kaymakam, müdür” works the same. Bakanı can also find Başbakanı etc (“Ekonomi Bakanı” is not included it needs to be concatenated to the string without space)
2. “Üniversitesi,Derneği,Takımı, etc…” is captured within one line to match the ORGANIZATIONs
3. “Caddesi,Sokağı etc…” is written in a similar way with Üniversitesi,Derneği etc… they capture the LOCATION data.
4. Hoca, Ağabey, Hanım, etc… is matching for PERSON (Ekin bey, Ayşe hanım…)
5. “Sayın, Sevgili” words are directly looked with if because these words are not expected have (çekim eki, yapım eki). If they exist within the input then program looks for a match otherwise it does not unnecessarily looks for a match.
6. “LTD, ŞTİ, AŞ” is evaluated in the same way and if they exist within the input with different period combinations (A.Ş., AŞ., LTD., …)
7. There are multiple date finders within the code #date-1 : Looking for the types of “12 Şubat 2000” where day number is optional and there might be (çekim eki, yapım eki) at the end optionally. #date-2: Looking for the type of 10.12.2000 where there might be “/ - . ” can be the separator. #date-3: Looking for hour format where hour is shown as 19:45 or 19.45 #date-4: Looking for “Aralık ayında” where letters and apostrophe after “ayı” is optional #date-5: Looking for days of the week; Pazartesi, Salı, … #date-6: Looking for “ 2000’de ” where years have (çekim eki, yapım eki) and apostrophe optionally
8. Every word is checked whether their first letter is capital, if so, they are checked whether they exist within any list (organizations, locations, person). If exact match is found, then they are printed according to which list it is found at. A word may be present in more than one list, but lists are checked in the order of “PERSON, LOCATION, ORGANIZATION” and if a match is found within any list further list searches are discarded to reduce the workload and multiple printing.

Because of that approach, a word (for instance “Mehmet”) can be captured by both the words that are emphasizing that it is a PERSON (for instance bey, cumhurbaşkanı, etc…) and for loop that is searching for a match within the lists.

Side Notes:

a)

if word[0] != ' '

If condition is there to prevent empty printings such as

Line 9: PERSON

b) Adding each print to an array prevents multiple prints of the exact same value for each category. If “Ali Osman” is printed once for PERSON category, then it will not be printed again to not overwhelm the screen.

c) Lines are counted according to enter (\n) command. If input file is written without new line character, then every line will be written with “Line 1”

d) Couple of regular expressions are found online and they have been cited with comments.

e) Comments at the bottom of the code are the previous or not working versions of the codes above.

f) There might be multiple info regarding the same entity. That is because there are more than one “indicator” has been used for a specific entity.

For instance, if the line “Cumhurbaşkanımız Ekin bey” is in the input then 2 lines will be printed:

-Cumhurbaşkanımız Ekin bey 🡪 Because of “Cumhurbaşkanı” 🡪 PERSON

-Cumhurbaşkanımız Ekin bey 🡪 Because of “bey” 🡪 PERSON (First word is taken because it is uppercase)

(If “C” of Cumhurbaşkanı is lowercase than we will not observe that word in the “bey” and output will be as follows:)

-Cumhurbaşkanımız Ekin bey 🡪 Because of “Cumhurbaşkanı” 🡪 PERSON

-Ekin bey 🡪 Because of “bey” 🡪 PERSON

g) Project is designed in a way to be run through terminal with “python ner.py input.txt > output.txt” code. Since my environment has some issues regarding the encoding/decoding of the Turkish characters, I was unable to use that command. Instead, I used “python ner.py input.txt” to print the output into terminal and asked for TA to check if he is getting any charmap error. He stated that code has no charmap error.