Bil334: Formal Languages and Automata

Homework 3

Assigned: 12/11/2021

Due: 21/11/2021, 23:59

In this homework, you are expected to parse a text file in Python **using regular expressions**.

Questions

1. We need to preprocess the text data consisting of Turkish tweets. Write a Python script named **preprocess_text.py** that will clean (remove non-letter characters) text file in the given input text file. Your program should read **tweets.txt** to preprocess text, the program should be run as:

$ python preprocess_text.py

(Note that the program is not getting any input.)

The program must produce the output file **preprocessed_tweets.txt**. The program should read the input file line by line. The output should be consisting of only letters. Punctuation marks and numbers before and at the end of the word should be deleted.

If there is a punctuation mark or/and number in the middle of the word, that word should be ignored. The words starting with "@" should be also ignored. (Example: https://t.co/7eCzBgVOhs and @bil334 should be ignored)

Note: Since the file with sample tweets consists of Turkish characters, read and write files using UTF-8.

Example Input File

@asirihaklii: Psikoloji der ki: "Bulunduğun anı tekrar yaşayamacağının bilincinde olursan hayat sana daha anlamlı gelir"

@madrigalofc: gece sabaha kadar burda oturup hayatı sorgulamak istiyorum https://t.co/b5n1dHungE

@spotifysozleri: "Bu yol nereye gider bilmem ama yürüyorum işte."

@picvsco: Şunlara bayıldım https://t.co/7eCzBgVOhs

Example Output File

Psikoloji der ki Bulunduğun anı tekrar yaşayamacağının bilincinde olursan hayat sana daha anlamlı gelir

gece sabaha kadar burda oturup hayatı sorgulamak istiyorum

Bu yol nereye gider bilmem ama yürüyorum işte

Şunlara bayıldım

2. Write a Python script named **validate_sentences.py** that will validate sentences in the given input text file. Your program should read **sentences.txt** to get sentences, the program should be run as:

$ python validate_sentences.py

(Note that the program is not getting any input.)

The program must produce the output file **sentences_output.txt**. The program should read the input file line by line. The program will decide that a sentence is valid or not valid. The program will write 'valid' to output file for each sentence that is valid and will write 'not valid' to output file for each sentence that is not valid.

Rules

- In each line, there must be only one sentence. Multiple sentences should not be accepted.
- Every sentence must begin with a capital letter.
- A word should not consist of both numbers and letters, they can be found separately. (Example: "4ever" should not be accepted)
- A sentence cannot end with a comma. At the end of the sentence, exclamation point, ellipsis, question mark and dot can be used. Also, sentences can end with quotes and quotation marks. (Example: This was first said by Shakespeare: "To thine own self be true.")
- Words can start with capital letters or words can be uppercase, but no middle or trailing capital letters. (NFA is acceptable, but NfA is not.)
- Sentences can contain punctuation marks such as quotes, quotation marks, colons; and the quotation marks and quotes must match. There may be sentences between quotation marks and quotes.
- You can assume that there are no contracted words. (Example: She'll-She will)
- You can assume that there will be no spaces before any punctuation mark. (Example: you, me is valid but you , me is not valid.)
- You can assume that there are no extra punctuation marks other than those found in the examples.

Example Input File

How can you say that to me?

As he looked at his reflection in the mirror, he took a deep breath.

He nodded at himself and, feeling braver, he stepped outside the bathroom. He bumped straight into the extremely tall man, who was waiting by the door.

David said ‘Oh, sorry!’.

The happy pair discussed their future life 2gether and shared sweet words of admiration.

We will not stop you; I promise!

Come here ASAP!

He pushed his chair back and went to the kitchen at 2 pM.

I do not know…

The main character in the movie said: "Play hard. Work harder."

Example Output File

valid

valid

not valid

valid

not valid

valid

valid

not valid

valid

valid