IN1140, H2021 – Oblig 1b
Regulære Uttrykk
Tokenisering med Regulære Uttrykk i Python

Innleveringsfrist: 06.10 kl. 23.59

Innlevering av oppgaven skjer i Devilry. Se emnesiden for mer informasjon om reglement rundt innlevering samt bruk av Devilry. Registrer svarene dine i en fil som angir brukernavnet ditt slik:

`oblig1b_brukernavn.py`

En perfekt løsning av denne oppgaven er verdt 100 poeng. Koden må kunne kjøres på IFIs maskiner og må inneholde kommentarer som forklarer hva koden gjør. Løsningen på oppgave 1, 2, 3 og 4 skal skrives som kommentar i innleveringsfilen.

NB! Vi ber deg også om å levere inputfilen `in01.txt` fra oblig 1a som vi skal bruke også i denne oppgaven. Dette gjør det enklere for oss å teste koden din ved retting.

1 Regulære uttrykk (eksamen H2018) (10 poeng)

Hvilket av alternativene kan **ikke** gjenkjennes med det følgende regulære uttrykket:

```
[1-9][0-9]*\s((cent(s)?)|(dollar(s)?\s+([1-9][0-9]*\scent(s)?)))
```

Velg ett alternativ:

1. `1 dollar 35 cents`
2. `35 dollars`
3. `99 cents`
4. `99 dollars 1 cent`

2 Regulære uttrykk for verb (inspirert av eksamen H2017) (20 poeng)

Skriv et regulært uttrykk som gjenkjenner formene imperativ, infinitiv, presens, og preteritum av følgende norske verb: spise, kjøpe, tenke, rope. Bruk tabell 1 som referanse.

Infinitiv	Imperativ	Presens	Preteritum
spise	spis	spiser	spiste
kjøpe	kjøp	kjøper	kjøpte
tenke	tenk	tenker	tenkte
rope	rop	roper	ropte

Table 1: Formene imperativ, infinitiv, presens, og preteritum av spise, kjøpe, tenke, rope

3 Regulært utrykk for filnavn (20 poeng)

Skriv et regulært uttrykk som gjenkjenner filnavn som ender med:

- `.gif`
- `.jpg`
- `.png`
- `.jpeg`
- `.pdf`
- `.rtf`

4 Tokenisering med regulære uttrykk (50 poeng)

I denne oppgaven skal vi jobbe med å forbedre tokeniseringskoden fra **Oblig 1a**, basert på feilanalysen fra oppgave 5 (i oblig 1a!). Du skal her forsøke å rette noen av de feilene du fant under feilanalysen din.

1. Vi starter med å formulere et regulært uttrykk som definerer et gyldig norsk ord. Her bør du bruke disjunksjon for å beskrive forskjellige typer ord. For eksempel, bør det regulære uttrykket ditt skille mellom bokstavsekvenser og tegnsetting, da et gyldig norsk ord ikke inkluderer tegnsetting og ulike typer tegnsetting skal skilles ut som et eget token. Eksempelvis er `Løp` et gyldig norsk ord/token, mens `Løp!` ikke er det.
2. Benytt deg av Python's `re.findall` for å hente ut ordene som matcher uttrykket ditt fra linjene med tekst fra filen `in01.txt` som ble utdelt med oblig 1a. Denne metoden tar et regulært uttrykk og en streng og returnerer en liste med alle treff. F.eks. vil `re.findall("([0-9])","in1140")` gi ut listen ['1', '1', '4', '0'].
3. Skriv ut og inspisér resultatet av den forbedrede tokeniseringen. Gjenstår det noen feil? Gi i såfall noen eksempler.