

Visualisering og WordCloud

Baggrund

Vi fortsætter med at bruge datasættet med fag og målepinde til at øve os i databehandling og visualisering.

Opgaven

Udvid jeres Python program fra sidste øvelse med følgende funktionalitet:

Forberedelse

- Saml alle målepinde for Programmerings-uddannelsen i én liste
- Fjern den foranstillede nummerering
- Fjern "stopord" (dvs. 'og', 'i', 'jeg', 'det', 'at', 'en' osv.)
- Fjerne special tegn som punktum, komma osv.
- Saml alle ord i én liste af enkelte ord.

Mål

- Tegn en WordCloud viser de mest brugte ord i målepindene
- Tegn et diagram, som viser de 30 meste brugte ord i målepindene

Hints

Der er Python pakker, som kan hjælpe os:

Pakken *nlTK* indeholder lister med Stopord - også dansk:

```
from nltk.corpus import stopwords
nltk.download('stopwords')

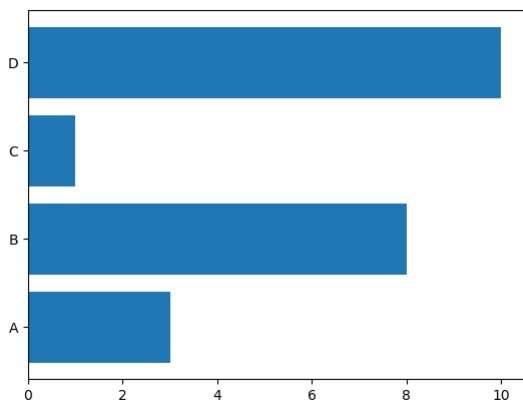
print(stopwords.words("danish"))
```

Pakken *Matplotlib* gør det nemt at tegne diagrammer:

```
import matplotlib.pyplot as plt
import numpy as np

x = np.array(["A", "B", "C", "D"])
y = np.array([3, 8, 1, 10])

plt.barh(x, y)
plt.show()
```



Pakken *WordCloud* kan tegne flotte billeder:

```
from wordcloud import WordCloud
import matplotlib.pyplot as plt

meta_text = '''
1. Lærlingen kan redegøre for, hvordan en maskine kan lære.
2. Lærlingen kan anvende maskinlæring til at løse praktiske opgaver.
3. Lærlingen kan samle og forberede data til maskinlæring.
'''

wc = WordCloud(background_color='black', width=800, height=500).generate(meta_text)
plt.axis("off")
```

```
plt.imshow(wc)
plt.show()
```



<https://medium.com/mlearning-ai/wordclouds-with-python-c287887acc8b>

Kontrol spørgsmål:

- Hvor mange målepinde er der i Programmerings-uddannelsen?
- Hvor mange "ikke stopord" er der?
- Hvad er 30 mest brugte ord?