

02-analyze-juris-docs

2024-09-01

```
[ ]: %run "./common.py"
```

```
[ ]: import os, regex
import pandas as pd
```

```
[ ]: def read_doc(doc, path_to_docs):
    with open(f'{path_to_docs}/{doc}') as f:
        text = f.read()
    return text

regexes = {
    'mitverschulden': '.{0,20}mitverschuld.{0,20}',
    'p254': '.{0,20}$\s*254.{0,20}',
    'p341': '.{0,20}$\s*341.{0,20}',
}

def find_matches(text, regexes:dict):
    matches = {regname:regex.findall(reg, text) for regname, reg in regexes.
    ↪items()}
    match_lengths = {f'{regname}_n':len(matches[regname]) for regname in
    ↪matches}
    return {**matches, **match_lengths}

def get_matchdicts(docs, path_to_docs):
    matchdicts = []
    for doc in docs:
        text = read_doc(doc, path_to_docs).lower()
        matches = find_matches(text, regexes)
        matchdict = {**matches, 'dateiname':doc}
        matchdicts.append(matchdict)
    return matchdicts
```

```
[ ]: path_to_helpers = ensure_exists('../data/BGH-XI-helpers')
path_to_docs = ensure_exists('../data/BGH-XI-cleaned')
docs = [x for x in os.listdir(path_to_docs) if x.endswith('.txt')]
len(docs)
```

```
[ ]: matchdicts = get_matchdicts(docs, path_to_docs)
```

```

[ ]: df = pd.DataFrame(matchdicts).set_index('dateiname')
[ ]: df_filtered = df[(df.mitverschulden_n > 0) | (df.p254_n > 0) | (df.p341_n > 0)]
[ ]: len(df_filtered)
[ ]: df_filtered.to_csv(f'{path_to_helpers}/mitverschulden-matches.csv')
[ ]: len(df[df.mitverschulden_n > 0]), len(df[df.p254_n > 0]), len(df[df.p341_n > 0])
[ ]: df[(df.mitverschulden_n > 0) & (df.p254_n > 0) & (df.p341_n > 0)]
[ ]: df[(df.p254_n > 0) & (df.p341_n > 0)]
[ ]: df[(df.mitverschulden_n > 0) & (df.p341_n > 0)]
[ ]: len(df[(df.mitverschulden_n > 0) & (df.p254_n > 0)])
[ ]: df[(df.mitverschulden_n > 0) & (df.p254_n > 0)]
[ ]:

```