TXT-Sammeldateien in einzelne Entscheidungen trennen und benennen

Sowohl die zur Trennung eingesetzte Zeichenkette als auch die Orte, an denen man Entscheidungsdatum und Aktenzeichen findet, unterscheiden sich, je nachdem, welche Datenquelle man verwendet. Das Skript geht davon aus, dass Batch-Dateien vorliegen, die mehrere Entscheidungen enthalten, deren Anzahl sich aus dem Dateinamen (z.B. '0000-0400.txt') ermitteln lässt.

In [ ]:

```
sourcepath = # Pfad zum Ordner, in dem die TXT-Sammeldateien liegen
targetpath = # Pfad zum Ordner, in den die einzelnen Entscheidungen gelegt werden so
```

In [ ]:

```
import re, string, os
```

In [ ]:

```
def txtsplit(collectionfile, number_of_decisions, output_path):
    """Receives a collectionfile with number_of_decisions different decisions,
       splits it into individual decisions, and writes the files to output_path."""

    splitter = # Zeichenkette, an welcher der Übergang zwischen Entscheidungen erken

    with open(collectionfile, "r") as f:
        text = f.read()
    files = text.split(f"{splitter}")[:-1]
    # [:-1] ggf. entfernen, falls Splitter nicht Symbol am Ende jeder Entscheidung i

    assert len(files) == number_of_decisions

    filenames = {f[:-4] for f in os.listdir(output_path) if f.endswith('.txt')}
    letters = string.ascii_lowercase
    new_filenames = set()

    # Suchmethoden für Datum und Aktenzeichen ggf. anpassen
    for text in files:
        raw_date = re.search("(?<=Entscheidungsdatum:.).*?\d{4}(?=\s)", text, re.DOT
        file_date = f"{raw_date[-1]}-{raw_date[-2]}-{raw_date[0][-2:]}"
        raw_az = re.search("(?<=Aktenzeichen:).*?(XI.*?/\d{2})", text, re.DOTALL).gr
        file_az = raw_az.replace("/", "-")
        original_filename = f"{file_date} {file_az}"

        # Eindeutige Dateinamen sicherstellen
        filename = original_filename
        idx = 0
        while filename in filenames:
            filename = f"{original_filename}_{letters[idx]}"
            idx += 1
        filenames.add(filename)
        new_filenames.add(filename)

        with open(f"{output_path}/{filename}.txt", "w") as f:
            f.write(text)

    assert len(new_filenames) == len(files)
```

In [ ]:

```
for f in os.listdir(sourcepath):
    number_of_decisions = int(f[:-4][-4:])-int(f[:4])+1
    txtsplit(f'{sourcepath}/{f}', number_of_decisions, targetpath)
```

Ende.