14_remove_indentation

March 21, 2017

2/24/17 - smiel

1 In order to not let it be a confound, we are going to normalize whitespace by removing whitespace after newlines, and compressing repeated spaces to a single space.

```
In [7]: %matplotlib inline
        # run me when first starting this notebook
        import os

        import numpy as np
        import pandas as pd

        path = '/research/ella/rivendell'

        datasets = {
            'ceeaus': 'CEEAUS',
            'fce': 'FCE',
            'icnale': 'ICNALE',
            'nog': 'NOG',
            'ra': 'RA-2016',
            'teccl': 'TECCL',
            'asap': 'ASAP',
            'chungdahm': 'Chungdahm',
            'gachon': 'Gachon',
            'moecs': 'MOECS',
            'nucle': 'NUCLE',
            'sbac': 'SBAC-Field',
            'toefl': 'TOEFL-11'
        }

        for folder, dataset in datasets.items():
            print(dataset)
            data_path = os.path.join(path, folder)

            if os.path.exists(os.path.join(data_path, '{}_sentences_w_indentation.csv'.format(da
                df = pd.read_csv(os.path.join(data_path, '{}_sentences_w_indentation.csv'.format
            else:
                df = pd.read_csv(os.path.join(data_path, '{}_sentences.csv'.format(dataset)), en

            # save a backup copy
            df.to_csv(
                os.path.join(data_path, '{}_sentences_w_indentation.csv'.format(dataset)), encod
            )

            # rename trailing_whitespace
            df['leading_whitespace'] = df.trailing_whitespace
            df.drop('trailing_whitespace', axis=1, inplace=True)

            # remove paragraph indentation, multiple newlines, repeated spaces, etc.
            df.leading_whitespace = df.leading_whitespace.str.replace(r'\n\s*', '\n')
            df.leading_whitespace = df.leading_whitespace.str.replace(r'[ ]+', ' ')

            df.text = df.text.str.replace(r'\n\s*', '\n')
            df.text = df.text.str.replace(r'[ ]+', ' ')

            # save sentences
            df.to_csv(os.path.join(data_path, '{}_sentences.csv'.format(dataset)), encoding='utf
        print('Done!')
```

```
NUCLE
ASAP
NOG
CEEAUS
FCE
SBAC-Field
TECCL
ICNALE
TOEFL-11
Chungdahm
RA-2016
Gachon
MOECS
Done!
```