19_newline_to_space

March 21, 2017

02/25/17 - smiel ### Not all data sets use newlines, so to prevent it from becoming a confound, we'll force all data sets to use a single space instead.

In [1]:
```
import pandas as pd
import os

path = '/research/ella/rivendell'
```

In [2]:
```
# if original versions exist, use those
orig = '.orig_19'

# essays
essays = os.path.join(path, 'all_essays.csv')
if os.path.exists(essays + orig):
    df = pd.read_csv(essays + orig, encoding='utf8')
else:
    df = pd.read_csv(essays, encoding='utf8')

print('{} essays'.format(len(df)))

# make a copy
df.to_csv(essays + orig, encoding='utf8', index=False)

# replace all spaces with a single space
df.text = df.text.str.replace('\s+', ' ')

# save
df.to_csv(essays, encoding='utf8', index=False)

print('{} essays'.format(len(df)))

# sentences
sentences = os.path.join(path, 'all_sentences.csv')
if os.path.exists(sentences + orig):
    df = pd.read_csv(sentences + orig, encoding='utf8')
else:
    df = pd.read_csv(sentences, encoding='utf8')
```

```
        print('{} sentences'.format(len(df)))

        # make a copy
        df.to_csv(sentences + orig, encoding='utf8', index=False)

        # replace all spaces with a single space
        df.text = df.text.str.replace('\s+', ' ')
        df.leading_whitespace = df.leading_whitespace.str.replace('\s+', ' ')

        # save
        df.to_csv(sentences, encoding='utf8', index=False)

        print('{} sentences'.format(len(df)))
```

```
/home/smiel/.venvs/rivendell/local/lib/python2.7/site-packages/IPython/core/interactiveshell.py:
  interactivity=interactivity, compiler=compiler, result=result)


141520 essays
141520 essays


/home/smiel/.venvs/rivendell/local/lib/python2.7/site-packages/IPython/core/interactiveshell.py:
  interactivity=interactivity, compiler=compiler, result=result)


2382191 sentences
2382191 sentences
```