18_drop_ceeaus_duplicates

March 21, 2017

```
In [4]: import pandas as pd
        import os

        path = '/research/ella/rivendell'

In [2]: ceeaus_drop = [
            u'chinese/ceecus_ptj_01.txt',
            u'chinese/ceecus_ptj_02.txt',
            u'chinese/ceecus_ptj_03.txt',
            u'chinese/ceecus_ptj_04.txt',
            u'chinese/ceecus_ptj_05.txt',
            u'chinese/ceecus_ptj_06.txt',
            u'chinese/ceecus_ptj_07.txt',
            u'chinese/ceecus_ptj_13.txt',
            u'chinese/ceecus_ptj_14.txt',
            u'chinese/ceecus_ptj_15.txt',
            u'chinese/ceecus_ptj_16.txt',
            u'chinese/ceecus_ptj_18.txt',
            u'chinese/ceecus_ptj_19.txt',
            u'chinese/ceecus_ptj_21.txt',
            u'chinese/ceecus_ptj_22.txt',
            u'chinese/ceecus_ptj_23.txt',
            u'chinese/ceecus_ptj_25.txt',
            u'chinese/ceecus_ptj_26.txt',
            u'chinese/ceecus_ptj_31.txt',
            u'chinese/ceecus_ptj_34.txt',
            u'chinese/ceecus_ptj_45.txt',
            u'chinese/ceecus_ptj_46.txt',
            u'chinese/ceecus_smk_01.txt',
            u'chinese/ceecus_smk_02.txt',
            u'chinese/ceecus_smk_03.txt',
            u'chinese/ceecus_smk_04.txt',
            u'chinese/ceecus_smk_05.txt',
            u'chinese/ceecus_smk_06.txt',
            u'chinese/ceecus_smk_07.txt',
            u'chinese/ceecus_smk_13.txt',
            u'chinese/ceecus_smk_14.txt',
```

```
u'chinese/ceecus_smk_15.txt',
u'chinese/ceecus_smk_16.txt',
u'chinese/ceecus_smk_18.txt',
u'chinese/ceecus_smk_19.txt',
u'chinese/ceecus_smk_21.txt',
u'chinese/ceecus_smk_22.txt',
u'chinese/ceecus_smk_23.txt',
u'chinese/ceecus_smk_25.txt',
u'chinese/ceecus_smk_26.txt',
u'chinese/ceecus_smk_29.txt',
u'chinese/ceecus_smk_32.txt',
u'chinese/ceecus_smk_41.txt',
u'chinese/ceecus_smk_43.txt',
u'chinese/ceecus_smk_44.txt',
u'english/ceenas_ptj_34.txt',
u'english/ceenas_ptj_35.txt',
u'english/ceenas_ptj_36.txt',
u'english/ceenas_ptj_37.txt',
u'english/ceenas_ptj_38.txt',
u'english/ceenas_ptj_39.txt',
u'english/ceenas_ptj_40.txt',
u'english/ceenas_ptj_41.txt',
u'english/ceenas_ptj_42.txt',
u'english/ceenas_ptj_43.txt',
u'english/ceenas_ptj_44.txt',
u'english/ceenas_ptj_45.txt',
u'english/ceenas_ptj_46.txt',
u'english/ceenas_ptj_47.txt',
u'english/ceenas_ptj_48.txt',
u'english/ceenas_ptj_49.txt',
u'english/ceenas_ptj_50.txt',
u'english/ceenas_ptj_51.txt',
u'english/ceenas_ptj_52.txt',
u'english/ceenas_ptj_54.txt',
u'english/ceenas_ptj_55.txt',
u'english/ceenas_ptj_56.txt',
u'english/ceenas_ptj_57.txt',
u'english/ceenas_ptj_58.txt',
u'english/ceenas_ptj_59.txt',
u'english/ceenas_ptj_60.txt',
u'english/ceenas_ptj_61.txt',
u'english/ceenas_ptj_62.txt',
u'english/ceenas_ptj_63.txt',
u'english/ceenas_ptj_64.txt',
u'english/ceenas_ptj_65.txt',
u'english/ceenas_ptj_66.txt',
u'english/ceenas_ptj_67.txt',
u'english/ceenas_ptj_68.txt',
```

```
u'english/ceenas_ptj_69.txt',
u'english/ceenas_ptj_71.txt',
u'english/ceenas_ptj_72.txt',
u'english/ceenas_smk_34.txt',
u'english/ceenas_smk_35.txt',
u'english/ceenas_smk_36.txt',
u'english/ceenas_smk_37.txt',
u'english/ceenas_smk_38.txt',
u'english/ceenas_smk_39.txt',
u'english/ceenas_smk_40.txt',
u'english/ceenas_smk_41.txt',
u'english/ceenas_smk_42.txt',
u'english/ceenas_smk_43.txt',
u'english/ceenas_smk_44.txt',
u'english/ceenas_smk_45.txt',
u'english/ceenas_smk_46.txt',
u'english/ceenas_smk_47.txt',
u'english/ceenas_smk_48.txt',
u'english/ceenas_smk_49.txt',
u'english/ceenas_smk_50.txt',
u'english/ceenas_smk_51.txt',
u'english/ceenas_smk_52.txt',
u'english/ceenas_smk_54.txt',
u'english/ceenas_smk_55.txt',
u'english/ceenas_smk_56.txt',
u'english/ceenas_smk_57.txt',
u'english/ceenas_smk_58.txt',
u'english/ceenas_smk_59.txt',
u'english/ceenas_smk_60.txt',
u'english/ceenas_smk_61.txt',
u'english/ceenas_smk_62.txt',
u'english/ceenas_smk_63.txt',
u'english/ceenas_smk_64.txt',
u'english/ceenas_smk_65.txt',
u'english/ceenas_smk_66.txt',
u'english/ceenas_smk_67.txt',
u'english/ceenas_smk_68.txt',
u'english/ceenas_smk_69.txt',
u'english/ceenas_smk_70.txt',
u'english/ceenas_smk_71.txt',
u'english/ceenas_smk_72.txt',
u'japanese/ceejus_L_ptj_01.txt',
u'japanese/ceejus_L_ptj_02.txt',
u'japanese/ceejus_L_ptj_03.txt',
u'japanese/ceejus_L_ptj_05.txt',
u'japanese/ceejus_L_ptj_17.txt',
u'japanese/ceejus_L_ptj_18.txt',
u'japanese/ceejus_L_ptj_19.txt',
```

```
u'japanese/ceejus_L_ptj_20.txt',
u'japanese/ceejus_L_ptj_21.txt',
u'japanese/ceejus_L_ptj_22.txt',
u'japanese/ceejus_L_ptj_23.txt',
u'japanese/ceejus_L_ptj_24.txt',
u'japanese/ceejus_L_ptj_26.txt',
u'japanese/ceejus_L_ptj_27.txt',
u'japanese/ceejus_L_ptj_28.txt',
u'japanese/ceejus_L_ptj_29.txt',
u'japanese/ceejus_L_ptj_30.txt',
u'japanese/ceejus_L_ptj_31.txt',
u'japanese/ceejus_L_ptj_32.txt',
u'japanese/ceejus_L_ptj_34.txt',
u'japanese/ceejus_L_ptj_35.txt',
u'japanese/ceejus_L_ptj_36.txt',
u'japanese/ceejus_L_ptj_37.txt',
u'japanese/ceejus_L_ptj_38.txt',
u'japanese/ceejus_L_ptj_39.txt',
u'japanese/ceejus_L_ptj_40.txt',
u'japanese/ceejus_L_smk_01.txt',
u'japanese/ceejus_L_smk_02.txt',
u'japanese/ceejus_L_smk_03.txt',
u'japanese/ceejus_L_smk_04.txt',
u'japanese/ceejus_L_smk_05.txt',
u'japanese/ceejus_L_smk_06.txt',
u'japanese/ceejus_L_smk_07.txt',
u'japanese/ceejus_L_smk_08.txt',
u'japanese/ceejus_L_smk_12.txt',
u'japanese/ceejus_L_smk_13.txt',
u'japanese/ceejus_L_smk_16.txt',
u'japanese/ceejus_L_smk_17.txt',
u'japanese/ceejus_L_smk_18.txt',
u'japanese/ceejus_L_smk_19.txt',
u'japanese/ceejus_L_smk_20.txt',
u'japanese/ceejus_L_smk_21.txt',
u'japanese/ceejus_L_smk_22.txt',
u'japanese/ceejus_L_smk_23.txt',
u'japanese/ceejus_L_smk_25.txt',
u'japanese/ceejus_L_smk_26.txt',
u'japanese/ceejus_L_smk_28.txt',
u'japanese/ceejus_L_smk_29.txt',
u'japanese/ceejus_L_smk_30.txt',
u'japanese/ceejus_L_smk_31.txt',
u'japanese/ceejus_L_smk_32.txt',
u'japanese/ceejus_L_smk_33.txt',
u'japanese/ceejus_L_smk_34.txt',
u'japanese/ceejus_L_smk_35.txt',
u'japanese/ceejus_L_smk_36.txt',
```

```
u'japanese/ceejus_L_smk_37.txt',
u'japanese/ceejus_L_smk_39.txt',
u'japanese/ceejus_L_smk_40.txt',
u'japanese/ceejus_M_ptj_001.txt',
u'japanese/ceejus_M_ptj_002.txt',
u'japanese/ceejus_M_ptj_003.txt',
u'japanese/ceejus_M_ptj_004.txt',
u'japanese/ceejus_M_ptj_005.txt',
u'japanese/ceejus_M_ptj_006.txt',
u'japanese/ceejus_M_ptj_007.txt',
u'japanese/ceejus_M_ptj_008.txt',
u'japanese/ceejus_M_ptj_009.txt',
u'japanese/ceejus_M_ptj_010.txt',
u'japanese/ceejus_M_ptj_011.txt',
u'japanese/ceejus_M_ptj_012.txt',
u'japanese/ceejus_M_ptj_013.txt',
u'japanese/ceejus_M_ptj_014.txt',
u'japanese/ceejus_M_ptj_015.txt',
u'japanese/ceejus_M_ptj_016.txt',
u'japanese/ceejus_M_ptj_018.txt',
u'japanese/ceejus_M_ptj_021.txt',
u'japanese/ceejus_M_ptj_022.txt',
u'japanese/ceejus_M_ptj_024.txt',
u'japanese/ceejus_M_ptj_025.txt',
u'japanese/ceejus_M_ptj_028.txt',
u'japanese/ceejus_M_ptj_029.txt',
u'japanese/ceejus_M_ptj_030.txt',
u'japanese/ceejus_M_ptj_032.txt',
u'japanese/ceejus_M_ptj_033.txt',
u'japanese/ceejus_M_ptj_034.txt',
u'japanese/ceejus_M_ptj_035.txt',
u'japanese/ceejus_M_ptj_037.txt',
u'japanese/ceejus_M_ptj_042.txt',
u'japanese/ceejus_M_ptj_043.txt',
u'japanese/ceejus_M_ptj_044.txt',
u'japanese/ceejus_M_ptj_045.txt',
u'japanese/ceejus_M_ptj_046.txt',
u'japanese/ceejus_M_ptj_047.txt',
u'japanese/ceejus_M_ptj_049.txt',
u'japanese/ceejus_M_ptj_050.txt',
u'japanese/ceejus_M_ptj_051.txt',
u'japanese/ceejus_M_ptj_053.txt',
u'japanese/ceejus_M_ptj_054.txt',
u'japanese/ceejus_M_ptj_056.txt',
u'japanese/ceejus_M_ptj_058.txt',
u'japanese/ceejus_M_ptj_081.txt',
u'japanese/ceejus_M_ptj_082.txt',
u'japanese/ceejus_M_ptj_083.txt',
```

```
u'japanese/ceejus_M_ptj_084.txt',
u'japanese/ceejus_M_ptj_085.txt',
u'japanese/ceejus_M_ptj_086.txt',
u'japanese/ceejus_M_ptj_087.txt',
u'japanese/ceejus_M_ptj_088.txt',
u'japanese/ceejus_M_ptj_090.txt',
u'japanese/ceejus_M_ptj_091.txt',
u'japanese/ceejus_M_ptj_092.txt',
u'japanese/ceejus_M_ptj_093.txt',
u'japanese/ceejus_M_ptj_094.txt',
u'japanese/ceejus_M_ptj_095.txt',
u'japanese/ceejus_M_ptj_096.txt',
u'japanese/ceejus_M_ptj_097.txt',
u'japanese/ceejus_M_ptj_098.txt',
u'japanese/ceejus_M_ptj_099.txt',
u'japanese/ceejus_M_ptj_102.txt',
u'japanese/ceejus_M_ptj_103.txt',
u'japanese/ceejus_M_ptj_104.txt',
u'japanese/ceejus_M_ptj_105.txt',
u'japanese/ceejus_M_ptj_106.txt',
u'japanese/ceejus_M_ptj_107.txt',
u'japanese/ceejus_M_ptj_108.txt',
u'japanese/ceejus_M_ptj_110.txt',
u'japanese/ceejus_M_ptj_112.txt',
u'japanese/ceejus_M_ptj_113.txt',
u'japanese/ceejus_M_ptj_114.txt',
u'japanese/ceejus_M_ptj_115.txt',
u'japanese/ceejus_M_ptj_116.txt',
u'japanese/ceejus_M_ptj_119.txt',
u'japanese/ceejus_M_ptj_121.txt',
u'japanese/ceejus_M_ptj_122.txt',
u'japanese/ceejus_M_ptj_123.txt',
u'japanese/ceejus_M_ptj_125.txt',
u'japanese/ceejus_M_ptj_126.txt',
u'japanese/ceejus_M_ptj_127.txt',
u'japanese/ceejus_M_ptj_128.txt',
u'japanese/ceejus_M_ptj_129.txt',
u'japanese/ceejus_M_ptj_130.txt',
u'japanese/ceejus_M_ptj_131.txt',
u'japanese/ceejus_M_ptj_132.txt',
u'japanese/ceejus_M_ptj_133.txt',
u'japanese/ceejus_M_ptj_134.txt',
u'japanese/ceejus_M_ptj_139.txt',
u'japanese/ceejus_M_ptj_140.txt',
u'japanese/ceejus_M_ptj_142.txt',
u'japanese/ceejus_M_ptj_143.txt',
u'japanese/ceejus_M_ptj_144.txt',
u'japanese/ceejus_M_ptj_145.txt',
```

```
u'japanese/ceejus_M_ptj_146.txt',
u'japanese/ceejus_M_ptj_147.txt',
u'japanese/ceejus_M_ptj_148.txt',
u'japanese/ceejus_M_ptj_149.txt',
u'japanese/ceejus_M_ptj_152.txt',
u'japanese/ceejus_M_ptj_153.txt',
u'japanese/ceejus_M_ptj_154.txt',
u'japanese/ceejus_M_ptj_155.txt',
u'japanese/ceejus_M_ptj_156.txt',
u'japanese/ceejus_M_ptj_157.txt',
u'japanese/ceejus_M_ptj_158.txt',
u'japanese/ceejus_M_ptj_159.txt',
u'japanese/ceejus_M_ptj_160.txt',
u'japanese/ceejus_M_ptj_161.txt',
u'japanese/ceejus_M_ptj_162.txt',
u'japanese/ceejus_M_ptj_163.txt',
u'japanese/ceejus_M_ptj_164.txt',
u'japanese/ceejus_M_ptj_165.txt',
u'japanese/ceejus_M_ptj_166.txt',
u'japanese/ceejus_M_ptj_167.txt',
u'japanese/ceejus_M_ptj_168.txt',
u'japanese/ceejus_M_ptj_169.txt',
u'japanese/ceejus_M_ptj_170.txt',
u'japanese/ceejus_M_ptj_171.txt',
u'japanese/ceejus_M_ptj_172.txt',
u'japanese/ceejus_M_ptj_173.txt',
u'japanese/ceejus_M_ptj_174.txt',
u'japanese/ceejus_M_ptj_175.txt',
u'japanese/ceejus_M_ptj_176.txt',
u'japanese/ceejus_M_ptj_177.txt',
u'japanese/ceejus_M_ptj_178.txt',
u'japanese/ceejus_M_ptj_179.txt',
u'japanese/ceejus_M_ptj_180.txt',
u'japanese/ceejus_M_ptj_181.txt',
u'japanese/ceejus_M_ptj_183.txt',
u'japanese/ceejus_M_ptj_185.txt',
u'japanese/ceejus_M_ptj_186.txt',
u'japanese/ceejus_M_ptj_187.txt',
u'japanese/ceejus_M_ptj_188.txt',
u'japanese/ceejus_M_ptj_189.txt',
u'japanese/ceejus_M_ptj_190.txt',
u'japanese/ceejus_M_ptj_191.txt',
u'japanese/ceejus_M_ptj_192.txt',
u'japanese/ceejus_M_ptj_193.txt',
u'japanese/ceejus_M_ptj_194.txt',
u'japanese/ceejus_M_ptj_195.txt',
u'japanese/ceejus_M_smk_001.txt',
u'japanese/ceejus_M_smk_002.txt',
```

```
u'japanese/ceejus_M_smk_003.txt',
u'japanese/ceejus_M_smk_004.txt',
u'japanese/ceejus_M_smk_005.txt',
u'japanese/ceejus_M_smk_006.txt',
u'japanese/ceejus_M_smk_007.txt',
u'japanese/ceejus_M_smk_008.txt',
u'japanese/ceejus_M_smk_009.txt',
u'japanese/ceejus_M_smk_010.txt',
u'japanese/ceejus_M_smk_011.txt',
u'japanese/ceejus_M_smk_012.txt',
u'japanese/ceejus_M_smk_013.txt',
u'japanese/ceejus_M_smk_014.txt',
u'japanese/ceejus_M_smk_015.txt',
u'japanese/ceejus_M_smk_016.txt',
u'japanese/ceejus_M_smk_017.txt',
u'japanese/ceejus_M_smk_019.txt',
u'japanese/ceejus_M_smk_020.txt',
u'japanese/ceejus_M_smk_022.txt',
u'japanese/ceejus_M_smk_023.txt',
u'japanese/ceejus_M_smk_024.txt',
u'japanese/ceejus_M_smk_025.txt',
u'japanese/ceejus_M_smk_026.txt',
u'japanese/ceejus_M_smk_027.txt',
u'japanese/ceejus_M_smk_028.txt',
u'japanese/ceejus_M_smk_029.txt',
u'japanese/ceejus_M_smk_030.txt',
u'japanese/ceejus_M_smk_032.txt',
u'japanese/ceejus_M_smk_033.txt',
u'japanese/ceejus_M_smk_034.txt',
u'japanese/ceejus_M_smk_035.txt',
u'japanese/ceejus_M_smk_036.txt',
u'japanese/ceejus_M_smk_037.txt',
u'japanese/ceejus_M_smk_038.txt',
u'japanese/ceejus_M_smk_039.txt',
u'japanese/ceejus_M_smk_040.txt',
u'japanese/ceejus_M_smk_041.txt',
u'japanese/ceejus_M_smk_043.txt',
u'japanese/ceejus_M_smk_044.txt',
u'japanese/ceejus_M_smk_046.txt',
u'japanese/ceejus_M_smk_047.txt',
u'japanese/ceejus_M_smk_053.txt',
u'japanese/ceejus_M_smk_054.txt',
u'japanese/ceejus_M_smk_055.txt',
u'japanese/ceejus_M_smk_056.txt',
u'japanese/ceejus_M_smk_057.txt',
u'japanese/ceejus_M_smk_058.txt',
u'japanese/ceejus_M_smk_059.txt',
u'japanese/ceejus_M_smk_060.txt',
```

```
u'japanese/ceejus_M_smk_062.txt',
u'japanese/ceejus_M_smk_063.txt',
u'japanese/ceejus_M_smk_064.txt',
u'japanese/ceejus_M_smk_066.txt',
u'japanese/ceejus_M_smk_067.txt',
u'japanese/ceejus_M_smk_069.txt',
u'japanese/ceejus_M_smk_070.txt',
u'japanese/ceejus_M_smk_071.txt',
u'japanese/ceejus_M_smk_072.txt',
u'japanese/ceejus_M_smk_073.txt',
u'japanese/ceejus_M_smk_075.txt',
u'japanese/ceejus_M_smk_076.txt',
u'japanese/ceejus_M_smk_077.txt',
u'japanese/ceejus_M_smk_078.txt',
u'japanese/ceejus_M_smk_079.txt',
u'japanese/ceejus_M_smk_080.txt',
u'japanese/ceejus_M_smk_081.txt',
u'japanese/ceejus_M_smk_083.txt',
u'japanese/ceejus_M_smk_085.txt',
u'japanese/ceejus_M_smk_086.txt',
u'japanese/ceejus_M_smk_087.txt',
u'japanese/ceejus_M_smk_088.txt',
u'japanese/ceejus_M_smk_089.txt',
u'japanese/ceejus_M_smk_092.txt',
u'japanese/ceejus_M_smk_094.txt',
u'japanese/ceejus_M_smk_095.txt',
u'japanese/ceejus_M_smk_096.txt',
u'japanese/ceejus_M_smk_098.txt',
u'japanese/ceejus_M_smk_099.txt',
u'japanese/ceejus_M_smk_100.txt',
u'japanese/ceejus_M_smk_101.txt',
u'japanese/ceejus_M_smk_102.txt',
u'japanese/ceejus_M_smk_103.txt',
u'japanese/ceejus_M_smk_104.txt',
u'japanese/ceejus_M_smk_105.txt',
u'japanese/ceejus_M_smk_106.txt',
u'japanese/ceejus_M_smk_107.txt',
u'japanese/ceejus_M_smk_111.txt',
u'japanese/ceejus_M_smk_112.txt',
u'japanese/ceejus_M_smk_114.txt',
u'japanese/ceejus_M_smk_115.txt',
u'japanese/ceejus_M_smk_116.txt',
u'japanese/ceejus_M_smk_117.txt',
u'japanese/ceejus_M_smk_118.txt',
u'japanese/ceejus_M_smk_119.txt',
u'japanese/ceejus_M_smk_120.txt',
u'japanese/ceejus_M_smk_121.txt',
u'japanese/ceejus_M_smk_122.txt',
```

```
u'japanese/ceejus_M_smk_123.txt',
u'japanese/ceejus_M_smk_126.txt',
u'japanese/ceejus_M_smk_127.txt',
u'japanese/ceejus_M_smk_128.txt',
u'japanese/ceejus_M_smk_129.txt',
u'japanese/ceejus_M_smk_130.txt',
u'japanese/ceejus_M_smk_131.txt',
u'japanese/ceejus_M_smk_133.txt',
u'japanese/ceejus_M_smk_134.txt',
u'japanese/ceejus_M_smk_135.txt',
u'japanese/ceejus_M_smk_149.txt',
u'japanese/ceejus_M_smk_150.txt',
u'japanese/ceejus_M_smk_151.txt',
u'japanese/ceejus_M_smk_152.txt',
u'japanese/ceejus_M_smk_153.txt',
u'japanese/ceejus_M_smk_154.txt',
u'japanese/ceejus_M_smk_155.txt',
u'japanese/ceejus_M_smk_156.txt',
u'japanese/ceejus_M_smk_157.txt',
u'japanese/ceejus_M_smk_158.txt',
u'japanese/ceejus_M_smk_159.txt',
u'japanese/ceejus_M_smk_160.txt',
u'japanese/ceejus_M_smk_161.txt',
u'japanese/ceejus_M_smk_162.txt',
u'japanese/ceejus_M_smk_163.txt',
u'japanese/ceejus_M_smk_164.txt',
u'japanese/ceejus_M_smk_165.txt',
u'japanese/ceejus_M_smk_166.txt',
u'japanese/ceejus_M_smk_167.txt',
u'japanese/ceejus_M_smk_168.txt',
u'japanese/ceejus_M_smk_169.txt',
u'japanese/ceejus_M_smk_170.txt',
u'japanese/ceejus_M_smk_171.txt',
u'japanese/ceejus_M_smk_172.txt',
u'japanese/ceejus_M_smk_173.txt',
u'japanese/ceejus_M_smk_174.txt',
u'japanese/ceejus_M_smk_175.txt',
u'japanese/ceejus_M_smk_176.txt',
u'japanese/ceejus_M_smk_177.txt',
u'japanese/ceejus_M_smk_178.txt',
u'japanese/ceejus_M_smk_180.txt',
u'japanese/ceejus_M_smk_182.txt',
u'japanese/ceejus_M_smk_183.txt',
u'japanese/ceejus_M_smk_185.txt',
u'japanese/ceejus_M_smk_187.txt',
u'japanese/ceejus_M_smk_188.txt',
u'japanese/ceejus_M_smk_189.txt',
u'japanese/ceejus_M_smk_190.txt',
```

```
u'japanese/ceejus_M_smk_191.txt',
u'japanese/ceejus_M_smk_192.txt',
u'japanese/ceejus_M_smk_193.txt',
u'japanese/ceejus_M_smk_194.txt',
u'japanese/ceejus_S_ptj_001.txt',
u'japanese/ceejus_S_ptj_043.txt',
u'japanese/ceejus_S_ptj_044.txt',
u'japanese/ceejus_S_ptj_045.txt',
u'japanese/ceejus_S_ptj_046.txt',
u'japanese/ceejus_S_ptj_047.txt',
u'japanese/ceejus_S_ptj_048.txt',
u'japanese/ceejus_S_ptj_049.txt',
u'japanese/ceejus_S_ptj_050.txt',
u'japanese/ceejus_S_ptj_052.txt',
u'japanese/ceejus_S_ptj_053.txt',
u'japanese/ceejus_S_ptj_054.txt',
u'japanese/ceejus_S_ptj_056.txt',
u'japanese/ceejus_S_ptj_057.txt',
u'japanese/ceejus_S_ptj_058.txt',
u'japanese/ceejus_S_ptj_059.txt',
u'japanese/ceejus_S_ptj_060.txt',
u'japanese/ceejus_S_ptj_061.txt',
u'japanese/ceejus_S_ptj_062.txt',
u'japanese/ceejus_S_ptj_063.txt',
u'japanese/ceejus_S_ptj_064.txt',
u'japanese/ceejus_S_ptj_065.txt',
u'japanese/ceejus_S_ptj_066.txt',
u'japanese/ceejus_S_ptj_067.txt',
u'japanese/ceejus_S_ptj_068.txt',
u'japanese/ceejus_S_ptj_069.txt',
u'japanese/ceejus_S_ptj_070.txt',
u'japanese/ceejus_S_ptj_071.txt',
u'japanese/ceejus_S_ptj_072.txt',
u'japanese/ceejus_S_ptj_073.txt',
u'japanese/ceejus_S_ptj_074.txt',
u'japanese/ceejus_S_ptj_075.txt',
u'japanese/ceejus_S_ptj_076.txt',
u'japanese/ceejus_S_ptj_077.txt',
u'japanese/ceejus_S_ptj_078.txt',
u'japanese/ceejus_S_ptj_079.txt',
u'japanese/ceejus_S_ptj_080.txt',
u'japanese/ceejus_S_ptj_081.txt',
u'japanese/ceejus_S_ptj_082.txt',
u'japanese/ceejus_S_ptj_083.txt',
u'japanese/ceejus_S_ptj_084.txt',
u'japanese/ceejus_S_ptj_085.txt',
u'japanese/ceejus_S_ptj_086.txt',
u'japanese/ceejus_S_ptj_087.txt',
```

```
u'japanese/ceejus_S_ptj_088.txt',
u'japanese/ceejus_S_ptj_089.txt',
u'japanese/ceejus_S_ptj_091.txt',
u'japanese/ceejus_S_ptj_092.txt',
u'japanese/ceejus_S_ptj_093.txt',
u'japanese/ceejus_S_ptj_094.txt',
u'japanese/ceejus_S_ptj_095.txt',
u'japanese/ceejus_S_ptj_096.txt',
u'japanese/ceejus_S_ptj_097.txt',
u'japanese/ceejus_S_ptj_098.txt',
u'japanese/ceejus_S_ptj_099.txt',
u'japanese/ceejus_S_ptj_100.txt',
u'japanese/ceejus_S_ptj_101.txt',
u'japanese/ceejus_S_ptj_102.txt',
u'japanese/ceejus_S_ptj_103.txt',
u'japanese/ceejus_S_ptj_104.txt',
u'japanese/ceejus_S_ptj_105.txt',
u'japanese/ceejus_S_ptj_106.txt',
u'japanese/ceejus_S_ptj_107.txt',
u'japanese/ceejus_S_ptj_108.txt',
u'japanese/ceejus_S_ptj_109.txt',
u'japanese/ceejus_S_ptj_110.txt',
u'japanese/ceejus_S_ptj_111.txt',
u'japanese/ceejus_S_ptj_112.txt',
u'japanese/ceejus_S_ptj_113.txt',
u'japanese/ceejus_S_ptj_114.txt',
u'japanese/ceejus_S_ptj_116.txt',
u'japanese/ceejus_S_ptj_117.txt',
u'japanese/ceejus_S_ptj_118.txt',
u'japanese/ceejus_S_ptj_119.txt',
u'japanese/ceejus_S_ptj_120.txt',
u'japanese/ceejus_S_ptj_122.txt',
u'japanese/ceejus_S_ptj_123.txt',
u'japanese/ceejus_S_ptj_125.txt',
u'japanese/ceejus_S_ptj_126.txt',
u'japanese/ceejus_S_ptj_128.txt',
u'japanese/ceejus_S_smk_001.txt',
u'japanese/ceejus_S_smk_002.txt',
u'japanese/ceejus_S_smk_004.txt',
u'japanese/ceejus_S_smk_005.txt',
u'japanese/ceejus_S_smk_006.txt',
u'japanese/ceejus_S_smk_008.txt',
u'japanese/ceejus_S_smk_009.txt',
u'japanese/ceejus_S_smk_011.txt',
u'japanese/ceejus_S_smk_014.txt',
u'japanese/ceejus_S_smk_015.txt',
u'japanese/ceejus_S_smk_016.txt',
u'japanese/ceejus_S_smk_018.txt',
```

```
u'japanese/ceejus_S_smk_019.txt',
u'japanese/ceejus_S_smk_020.txt',
u'japanese/ceejus_S_smk_021.txt',
u'japanese/ceejus_S_smk_022.txt',
u'japanese/ceejus_S_smk_024.txt',
u'japanese/ceejus_S_smk_025.txt',
u'japanese/ceejus_S_smk_026.txt',
u'japanese/ceejus_S_smk_028.txt',
u'japanese/ceejus_S_smk_029.txt',
u'japanese/ceejus_S_smk_031.txt',
u'japanese/ceejus_S_smk_032.txt',
u'japanese/ceejus_S_smk_033.txt',
u'japanese/ceejus_S_smk_035.txt',
u'japanese/ceejus_S_smk_036.txt',
u'japanese/ceejus_S_smk_037.txt',
u'japanese/ceejus_S_smk_038.txt',
u'japanese/ceejus_S_smk_039.txt',
u'japanese/ceejus_S_smk_040.txt',
u'japanese/ceejus_S_smk_041.txt',
u'japanese/ceejus_S_smk_042.txt',
u'japanese/ceejus_S_smk_043.txt',
u'japanese/ceejus_S_smk_044.txt',
u'japanese/ceejus_S_smk_045.txt',
u'japanese/ceejus_S_smk_046.txt',
u'japanese/ceejus_S_smk_047.txt',
u'japanese/ceejus_S_smk_048.txt',
u'japanese/ceejus_S_smk_049.txt',
u'japanese/ceejus_S_smk_050.txt',
u'japanese/ceejus_S_smk_051.txt',
u'japanese/ceejus_S_smk_052.txt',
u'japanese/ceejus_S_smk_053.txt',
u'japanese/ceejus_S_smk_054.txt',
u'japanese/ceejus_S_smk_055.txt',
u'japanese/ceejus_S_smk_056.txt',
u'japanese/ceejus_S_smk_057.txt',
u'japanese/ceejus_S_smk_058.txt',
u'japanese/ceejus_S_smk_059.txt',
u'japanese/ceejus_S_smk_060.txt',
u'japanese/ceejus_S_smk_061.txt',
u'japanese/ceejus_S_smk_062.txt',
u'japanese/ceejus_S_smk_063.txt',
u'japanese/ceejus_S_smk_064.txt',
u'japanese/ceejus_S_smk_065.txt',
u'japanese/ceejus_S_smk_066.txt',
u'japanese/ceejus_S_smk_067.txt',
u'japanese/ceejus_S_smk_068.txt',
u'japanese/ceejus_S_smk_069.txt',
u'japanese/ceejus_S_smk_070.txt',
```

```
u'japanese/ceejus_S_smk_071.txt',
u'japanese/ceejus_S_smk_072.txt',
u'japanese/ceejus_S_smk_073.txt',
u'japanese/ceejus_S_smk_074.txt',
u'japanese/ceejus_S_smk_075.txt',
u'japanese/ceejus_S_smk_076.txt',
u'japanese/ceejus_S_smk_077.txt',
u'japanese/ceejus_S_smk_078.txt',
u'japanese/ceejus_S_smk_079.txt',
u'japanese/ceejus_S_smk_080.txt',
u'japanese/ceejus_S_smk_081.txt',
u'japanese/ceejus_S_smk_082.txt',
u'japanese/ceejus_S_smk_083.txt',
u'japanese/ceejus_S_smk_084.txt',
u'japanese/ceejus_S_smk_085.txt',
u'japanese/ceejus_S_smk_086.txt',
u'japanese/ceejus_S_smk_087.txt',
u'japanese/ceejus_S_smk_088.txt',
u'japanese/ceejus_S_smk_089.txt',
u'japanese/ceejus_S_smk_090.txt',
u'japanese/ceejus_S_smk_091.txt',
u'japanese/ceejus_S_smk_092.txt',
u'japanese/ceejus_S_smk_093.txt',
u'japanese/ceejus_S_smk_095.txt',
u'japanese/ceejus_S_smk_096.txt',
u'japanese/ceejus_S_smk_097.txt',
u'japanese/ceejus_S_smk_098.txt',
u'japanese/ceejus_S_smk_106.txt',
u'japanese/ceejus_S_smk_107.txt',
u'japanese/ceejus_S_smk_108.txt',
u'japanese/ceejus_S_smk_109.txt',
u'japanese/ceejus_S_smk_110.txt',
u'japanese/ceejus_S_smk_111.txt',
u'japanese/ceejus_S_smk_112.txt',
u'japanese/ceejus_S_smk_113.txt',
u'japanese/ceejus_S_smk_115.txt',
u'japanese/ceejus_S_smk_116.txt',
u'japanese/ceejus_S_smk_117.txt',
u'japanese/ceejus_S_smk_118.txt',
u'japanese/ceejus_S_smk_119.txt',
u'japanese/ceejus_S_smk_121.txt',
u'japanese/ceejus_S_smk_122.txt',
u'japanese/ceejus_S_smk_125.txt',
u'japanese/ceejus_S_smk_126.txt',
u'japanese/ceejus_S_smk_128.txt',
u'japanese/ceejus_U_ptj_01.txt',
u'japanese/ceejus_U_ptj_11.txt',
u'japanese/ceejus_U_ptj_13.txt',
```

```
        u'japanese/ceejus_U_smk_02.txt',
        u'japanese/ceejus_U_smk_03.txt',
        u'japanese/ceejus_U_smk_04.txt',
        u'japanese/ceejus_U_smk_05.txt',
        u'japanese/ceejus_U_smk_06.txt',
        u'japanese/ceejus_U_smk_07.txt',
        u'japanese/ceejus_U_smk_08.txt',
        u'japanese/ceejus_U_smk_09.txt',
        u'japanese/ceejus_U_smk_11.txt',
        u'japanese/ceejus_U_smk_17.txt',
        u'japanese/ceejus_U_smk_18.txt',
        u'japanese/ceejus_U_smk_19.txt'
    ]
```

In [5]:
```
# if original versions exist, use those
orig = '.orig_18'

# 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)

# drop essays
df = df[~(df.essay_id.isin(ceeaus_drop))]

# 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)
```

```
        # drop essays
        df = df[~(df.essay_id.isin(ceeaus_drop))]

        # 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)


142187 essays
141520 essays


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


2392918 sentences
2382191 sentences
```