SimplifyToTradidtional

2018 年 7 月 5 日

1 簡繁轉換

1.1 介紹

我們在處理語言處理的時候，常常會遇到一個問題，

1.2 需要函式庫

1.2.1 OpenCC-Python

可以幫我們簡繁轉換
https://github.com/yichen0831/opencc-python

1.2.2 chardet(選用)

編碼猜測器, 有時候我們真的猜不出來到底是 GBK 還是 windows 編碼的時候, 可以選擇使用猜測器猜測一下
https://github.com/chardet/chardet

```
In [ ]: import os
        import jieba
        from opencc import OpenCC
        # import chardet

        # 我們要轉換的資料夾
        base_dir = "chinese_news"
        # 我們轉換完輸出的資料夾
        trans_dir = base_dir + "_trans"
        # 如果不存在, 我們把資料夾做出來
        if not os.path.exists(trans_dir):
```

```
        os.makedirs(trans_dir)

    # os.walk 會回出一個 tuple, 而且會輪迴的走過 base_dir 的所有資料夾, 直到他找到檔案為止
    for dirPath, dirNames, fileNames in os.walk(base_dir):
        # print(dirPath, dirNames, fileNames)
        # 準備一個簡繁轉換翻譯器
        openCC = OpenCC('s2tw')

        # 針對每一個檔案做出轉換
        for single_name in fileNames:
            if not single_name.startswith("."):
                # 先不解碼 (rb), 後面再用簡體解碼
                # 當然你也可以在這裡直接使用 encoding="GBK"
                f = open(os.path.join(dirPath, single_name), "rb")
                content = f.read()
                try:
                    content = content.decode("GBK")
                    f.close()
                    # 轉換!!
                    new_content = openCC.convert(content)
                    # 把處理的文章路徑印出來
                    # print(os.path.join(new_dirPath, single_name))
                    # 把轉換好的繁體用 utf-8 編碼寫入文件中
                    f = open(os.path.join(new_dirPath, single_name),
                             "w",
                             encoding="utf-8")
                    f.write(new_content)
                    f.close()
                except:
                    # 有時候就是無法正確的轉換, 我們可以選擇直接放棄那一點點無法轉換的文章
                    print("[Decode Error] 放棄此篇文章")
```