文本处理的第一步骤： Tokenization

基础环境：

Operation system：Centos7
Python version：Python-3.6.1
Pip version: pip 3

环境配置：

pip3 install foolnltk
pip3 install tensorflow

pip3 install jieba

Reference：
https://www.oschina.net/p/foolnltk
https://www.oschina.net/p/jieba
http://blog.csdn.net/eastmount/article/details/50256163
http://blog.csdn.net/eastmount/article/details/50256163

切词基础资源

PorterStemmer Python code：
http://ccl.pku.edu.cn/doubtfire/NLP/Lexical_Analysis/Word_Lemmatization/Porter/Porter%20Stemming%20Algorithm.htm
https://tartarus.org/martin/PorterStemmer/
https://pypi.python.org/pypi/PorterStemmer/
http://blog.csdn.net/noobzc1/article/details/8902881

英文 stop words
http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words

Python NLTK（python 自然语言处理包）：
https://www.cnblogs.com/huiyang865/p/5571421.html

英文关键词提取

```
#Porter’ s algorithm
# 第一步，处理复数，以及 ed 和 ing 结束的单词。
# 第二步，如果单词中包含元音，并且以 y 结尾，将 y 改为 i。
# 第三步，将双后缀的单词映射为单后缀。
# 第四步，处理-ic-，-full，-ness 等等后缀。
# 第五步，在<c>vcvc<v>情形下，去除-ant，-ence 等后缀。
# 第六步，也就是最后一步，在 m()>1 的情况下，移除末尾的“e”。
```

```python
#0 get stop words list
stop_words = get_stop_words()

# 1 get documents name from the specific folder
file_names = get_file_names()
# 2 get documents text
for file_key, file_value in file_names.items():
    raw_text = get_file_content(file_value)
    #确认效果
    #print(raw_text)

    #3 Text Normalization
    normalized_text = text_normalization(raw_text)
    # 确认效果
    #print(normalized_text)

    # 4 Text Lowercasing
    lowercasing_text = text_lowercasing(normalized_text)
    # 确认效果
    #print(lowercasing_text)

    #5 Text Tokenization Text -> terms
    tokenized_list = text_tokenization(lowercasing_text)
    # 确认效果
    #print(tokenized_list)

    #6 Stop word dropping
    highlights_list = text_stopword_dropping(tokenized_list, stop_words)
    #print(highlights_text)

    #7 Stemming and lemmatization linguistic preprocessing of tokens（词干提取和词性还原）
    stm_len_list = text_stemming_lemmatization(highlights_list)
    #print(stm_len_list)

    #8 Stop word dropping
    final_list = text_stopword_dropping(stm_len_list, stop_words)
    #print(final_list)

    #9 保存成文本
    save_document_terms(file_name=file_key, document_terms=final_list, file_folder=’temp’)
```

中文关键词提取

```
#documents_terms ={}
#0 get stop words list
stop_words = get_cn_stop_words(lang_code = 'zh-utf8')

# 1 get documents name from the specific folder
file_names = get_file_names()
# 2 get documents text
for file_key, file_value in file_names.items():
    raw_text = get_cn_file_content(file_value)
    #确认效果
    #print(raw_text)

    # 3Text Tokenization Text -> terms
    # 导入自定义词典
    #jieba.load_userdict("dict.txt")
    # 全模式
    #text = "故宫的著名景点包括乾清宫、太和殿和黄琉璃瓦等"
    #seg_list = jieba.cut(text, cut_all=True)
    #print u"[全模式]: ", "/ ".join(seg_list)
    # 精确模式
    #seg_list = jieba.cut(text, cut_all=False)
    #print u"[精确模式]: ", "/ ".join(seg_list)
    # 搜索引擎模式
    #seg_list = jieba.cut_for_search(text)
    #print u"[搜索引擎模式]: ", "/ ".join(seg_list)

    if cut_type == 'jieba':
        temp = jieba.cut(raw_text, cut_all=False)
        temp = '.'.join(temp)
        tokenized_list = temp.split('.')
    else:
        tokenized_list = fool.cut(raw_text)
        # 确认效果
    # print(tokenized_list)

    #4 Text Normalization
    normalized_list = cn_text_normalization(tokenized_list)
    # 确认效果
    #print(normalized_text)

    #5 Stop word dropping
    final_list = text_stopword_dropping(normalized_list, stop_words)
    #print(final_list)
```

```
#6 保存成文本
save_document_terms(file_name=file_key, document_terms=final_list, file_folder=’temp’)
```