Word Segmentation
Lab 4

[Getting started]

Word segmentation is a necessary process in creating language model, information extraction and search engines.

Jieba and HanLP are two very popular Chinese word segmentation tools. Today we are going to learn Jieba. For a full description of Jieba, please visit
https://github.com/fxsjy/jieba

In this lab, we will try the most commonly used functions in Jieba.

First, install Jieba with

```
pip install jieba
```

or

```
pip3 install jieba
```

Run “import jieba” in python to see if Jieba is installed successfully.

[Functions]

Let us examine the functions one by one.

First we try the cut function.

There are three major modes: the accurate mode, full mode and search engine mode.

Code example: segmentation

```
#encoding=utf-8
import jieba

seg_list = jieba.cut("我来到北京清华大学", cut_all=False)  # 默认模式
print("Accurate Mode: " + "/ ".join(seg_list))
seg_list = jieba.cut("我来到北京清华大学", cut_all=True)   # 全模式
print("Full Mode: " + "/ ".join(seg_list))
seg_list = jieba.cut_for_search("我来到北京清华大学")  # 搜索引擎模式
```


zhuym@sustc.edu.cn

```
print("Search Engine Mode: " + "/ ".join(seg_list))
seg_list = jieba.cut("他来到了网易杭研大厦")
print("Unknown Words Recognize: " + ", ".join(seg_list))
```

Output:

```
[Accurate Mode]: 我/ 来到/ 北京/ 清华大学
[Full Mode]: 我/ 来到/ 北京/ 清华/ 清华大学/ 华大/ 大学
[Search Engine Mode]：我/ 来到/ 北京/ 清华/ 华大/ 大学/ 清华大学
[Unknown Words Recognize] 他, 来到, 了, 网易, 杭研, 大厦    (In this case, "杭研" is not in the dictionary, but is identified by the Viterbi algorithm)
```

- The `jieba.cut` function accepts three input parameters: the first parameter is the string to be cut; the second parameter is `cut_all`, controlling the cut mode; the third parameter is to control whether to use the Hidden Markov Model.
- `jieba.cut_for_search` accepts two parameter: the string to be cut; whether to use the Hidden Markov Model. This will cut the sentence into short words suitable for search engines.
- The input string can be an unicode/str object, or a str/bytes object which is encoded in UTF-8 or GBK. Note that using GBK encoding is not recommended because it may be unexpectly decoded as UTF-8.
- `jieba.cut` and `jieba.cut_for_search` returns an generator, from which you can use a `for` loop to get the segmentation result (in unicode).
- `jieba.lcut` and `jieba.lcut_for_search` returns a list.

Next we want to modify the dictionary to improve the segmentation result.

Example:

```
>>>print('/'.join(jieba.cut('李小福是创新办主任也是云计算方面的专家')))
李小福 / 是 / 创新 / 办 / 主任 / 也 / 是 / 云 / 计算 / 方面 / 的 / 专家 /
>>> jieba.add_word('创新办')
>>> jieba.add_word('云计算')
李小福 / 是 / 创新办 / 主任 / 也 / 是 / 云计算 / 方面 / 的 / 专家 /
```

```
>>> print('/'.join(jieba.cut('如果放到 post 中将出错。', HMM=False)))
如果/放到/post/中将/出错/。
>>> jieba.suggest_freq(('中', '将'), True)
494
>>> print('/'.join(jieba.cut('如果放到 post 中将出错。', HMM=False)))
如果/放到/post/中/将/出错/。
>>> print('/'.join(jieba.cut('「台中」正确应该不会被切开', HMM=False)))
「/台/中/」/正确/应该/不会/被/切开
>>> jieba.suggest_freq('台中', True)
69
>>> print('/'.join(jieba.cut('「台中」正确应该不会被切开', HMM=False)))
「/台中/」/正确/应该/不会/被/切开
```

- Use `add_word(word, freq=None, tag=None)` and `del_word(word)` to modify the dictionary dynamically in programs.
- Use `suggest_freq(segment, tune=True)` to adjust the frequency of a single word so that it can (or cannot) be segmented. (Note that this function will increase the frequency by 1. If you want to increase the frequency by x, call this function x times.)
- Note that HMM may affect the final result.

[Exercises]

1. Use Jieba to segment the following sentence:

    江州市长江大桥参加了长江大桥的通车仪式

    我們在野生動物園玩

    ETF 新兵登场元大宝来推 ETF 伞型证券投资信托基金

    Try to use the Jieba functions to segment the sentences in a correct way.

2. Can you modify the Jieba setting so that it will separate every single Chinese character? Can Jieba help you in your assignment 1?

Please also refer to https://github.com/fxsjy/jieba and following the “Keyword Extraction”, “Part of Speech Tagging” and “Tokenize: return words with position” sections on your own.