Lab 5
Sentence Query

[Objective]

1. Learn using Gensim to implement a sentence query system.

In this lab, we will follow the gensim tutorial to learn the basic of a query system. The source code of gensim is located at
https://github.com/RaRe-Technologies/gensim

First of all, install gensim:

pip install gensim

and Levenshtein:

pip install python-Levenshtein

Then please go through the tutorial located at
https://radimrehurek.com/gensim/auto_examples/core/run_core_concepts.html

At the end of the tutorial, you can make a query with a sentence and the model will return the similarity scores of every documents in the corpus.

In this lab, you are going to convert the English query system into a Chinese one. You can change the English corpus into the followings:

```
text_corpus = [
    "此前拜登政府宣布将投入几千亿美元推动美国电动汽车行业发展",
    "解放军东部战区新闻发言人表示台湾及其附属岛屿是中国领土的一部分",
    "受利好消息刺激影响海南概念板块直线上行券商股拉升",
    "要做好疫情防控毫不放松抓实抓细疫情防控各项措施",
    "上海市出台高中招考新政实验性示范性高中将出其招生计划",
    "游客出游热情高涨红色旅游持续升温踏青游近郊游乡村游自驾游等需求加速释放",
    "忙为新剧拍摄表示完成拍摄后返来电视城为劲歌金曲担任嘉宾亦很开心见到偶像",
    "印度菜对于香港人并不陌生咖喱薄饼串烧等印度美食更深得港人欢心",
    "继上一圈斗巴塞隆拿两回合狂轰四球安巴比这次面对盟主拜仁慕尼黑又有佳作",
]
```


zhuym@sustc.edu.cn

You might need to change the text preprocessing part as we are now dealing with Chinese sentences. Remember to use Jieba to do the word segmentation.
At the end, you can make a query with a Chinese sentence and see which sentence in the text corpus has the highest similarity. Remember that you also need to segment the words in your query.

There is a lot more Gensim can do, please feel free to explore it.