# -\*- coding: utf-8 -\*-  
*"""  
 @Time : 2023/3/19 10:43  
 @Author : liao.sc  
 @File : demo02  
 @Contact : 446773160@qq.com  
"""*import jieba  
import thulac  
  
# 分词的自我介绍  
text = """您好，我是廖山川，来自安徽，目前就读于上海师范大学。我是一名喜欢编程、音乐和吉他的学生，同时也有着坚持锻炼的习惯。在编程方面,我的主要工作语言是Java。我喜欢使用Java开发各种应用程序，包括Web应用、移动应用等等。我深入研究Java语言的特性和应用场景，并不断学习新技术以提高自己的技能和编程水平。在科研方面,我主要使用Python语言进行数据分析和建模。我熟练掌握Python语言和各种相关的数据分析工具，能够对大量数据进行处理和分析，并通过数据建模来预测未来趋势。此外，我也是一位音乐爱好者和吉他手。我喜欢弹奏吉他，演奏不同风格的音乐，通过音乐来放松和享受生活。总的来说，我是一位热爱学习和不断提高自己的人。我希望能够在未来的职业生涯中继续探索和发展我的技能，为社会做出贡献。"""  
  
# 使用jieba分词  
jieba\_tokens = jieba.lcut(text)  
  
# 使用THULAC分词  
thu = thulac.thulac()  
thulac\_tokens = thu.cut(text, text=True)  
  
# 比较分词结果  
def compare\_tokens(ref\_tokens, test\_tokens):  
 ref\_set = set(ref\_tokens)  
 test\_set = set(test\_tokens)  
 common\_set = ref\_set & test\_set  
 iv\_recall = len(common\_set) / len(ref\_set)  
 oov\_recall = len(common\_set) / len(test\_set - ref\_set)  
 precision = len(common\_set) / len(test\_set)  
 recall = len(common\_set) / len(ref\_set)  
 return {'IV Recall': iv\_recall, 'OOV Recall': oov\_recall, 'Precision': precision, 'Recall': recall}  
  
result\_jieba = compare\_tokens(jieba\_tokens, thulac\_tokens)  
result\_thulac = compare\_tokens(thulac\_tokens, jieba\_tokens)  
  
print('jieba分词结果:', jieba\_tokens)  
print('THULAC分词结果:', thulac\_tokens)  
print('jieba分词结果与THULAC分词结果的比较:')  
print('Precision:', result\_jieba['Precision'])  
print('Recall:', result\_jieba['Recall'])  
print('IV Recall:', result\_jieba['IV Recall'])  
print('OOV Recall:', result\_jieba['OOV Recall'])  
print('THULAC分词结果与jieba分词结果的比较:')  
print('Precision:', result\_thulac['Precision'])  
print('Recall:', result\_thulac['Recall'])  
print('IV Recall:', result\_thulac['IV Recall'])  
print('OOV Recall:', result\_thulac['OOV Recall'])