实验1：训练歌词词向量

1. 题目

题目：请基于给定歌词语料“lyrics_10k.txt”，训练歌词词向量，用gensim、jieba库训练中文歌词词向量，结合自己喜好任选10个词，输出每个词的三个同义词，打印输出结果。可选择去除停用词后训练词向量模型。

2. 示例代码

下载包：

```
pip install gensim jieba
```

代码：

```
import jieba
import gensim
from gensim.models import Word2Vec

# 假设你的歌词语料是一个列表，每项是每首歌词的文本字符串
# 例如: ['我爱你', '歌曲很美', ...]

# 示例中文歌词语料
lyrics_corpus = [
    '我爱你',
    '歌曲很美',
    '心情愉快',
    '音乐让我开心',
    '爱情很美好',
    '每一天都是新的开始',
    '梦想在远方',
    '生活中充满希望',
    '音乐是我最好的朋友',
    '每一个笑容都能温暖心灵',
    # 更多歌词...
]

# 使用jieba分词将歌词语料转换为词的列表
tokenized_corpus = [list(jieba.cut(song)) for song in lyrics_corpus]

# 训练Word2Vec模型
model = Word2Vec(sentences=tokenized_corpus, vector_size=100, window=5,
min_count=1, workers=4)

# 保存训练好的词向量模型（可选）
model.save("lyrics_word2vec.model")

# 加载模型（如果从文件加载）
# model = gensim.models.Word2Vec.load("lyrics_word2vec.model")

# 选择10个词来测试（确保这些词出现在词汇表中）
sample_words = ['我', '爱', '歌曲', '很', '美', '心情', '愉快', '音乐', '每', '生活']
```

```
# 输出这些词的同义词
for word in sample_words:
    try:
        # 获取同义词（最相似的词）
        similar_words = model.wv.most_similar(word, topn=3)
        print(f"同义词 - {word}:")
        for similar_word, similarity in similar_words:
            print(f"  {similar_word} (相似度: {similarity:.4f})")
    except KeyError:
        print(f"词 '{word}' 不在词汇表中，跳过。")
print()
```

- **分词**：使用 jieba.cut() 对中文歌词进行分词，将每首歌词转换成词的列表。tokenized_corpus 是一个包含每首歌词分词后列表的列表。
- **训练Word2Vec模型**：使用 Word2Vec 训练中文歌词的词向量，vector_size=100 设置词向量的维度，window=5 设置上下文窗口大小，min_count=1 设置最小词频。
- **输出同义词**：通过 model.wv.most_similar 获取与指定词最相似的词（即同义词）。

3. 效果

```
同义词 - 我:
  爱 (相似度: 0.8571)
  每 (相似度: 0.7223)
  心情 (相似度: 0.6715)

同义词 - 爱:
  我 (相似度: 0.8571)
  每 (相似度: 0.7223)
  音乐 (相似度: 0.6014)

...
```