

**实验报告**

****

课程名称： 自然语言处理I

人工智能 学院 人工智能 专业2020年级 1 班

姓名李建洲学号202083460016指导老师 王丽娜

人工智能学院

日期 2022 年 12 月

词向量

# **实验目的**

本章实验的主要目的是掌握自然语言处理中文本数据转换为数值型数据常用的模型方法。通过Python3.6和sklearn.preprocessing中的OneHotEncoder进行简单的One-hot编码操作演练，让学员进一步掌握One-hot编码的原理；通过TF-IDF算法的调用，演示关键词提取的过程，让学员进一步掌握TF-IDF算法原理；最后学员利用Python和gensim框架实现Word2vec中两种模型CBOW和skip-gram，并应用于文本数据集的处理。

# **实验环境**

Jupyter Notebook

# **实验原理**

1.One-hot编码也称之为独热编码，是NLP里的经常在预处理数据的时候使用的技术。通常是需要将离散型特征或者标签用One-hot编码。比如房价预测中，出现房屋朝向的特征，一共有南、东南、西南、东、东北、北、西北和西8个方向，那么这样的离散特征是需要One-hot编码的，比如“南”的one-hot编码为[1,0,0,0,0,0,0,0]，“东南”的One-hot编码为[0,1,0,0,0,0,0,0]…依此类推。

2.关键词是指能反映文本主题或者主要内容的词语。关键词提取是NLP领域中一个重要的子任务。在信息检索中，准确的关键词提取可以大幅提升效率；在对话系统中，机器可以通过关键词来理解用户意图；在文本摘要中，关键词提取有利于形成摘要；以及文本分类等相关应用场景中都有着重要作用。

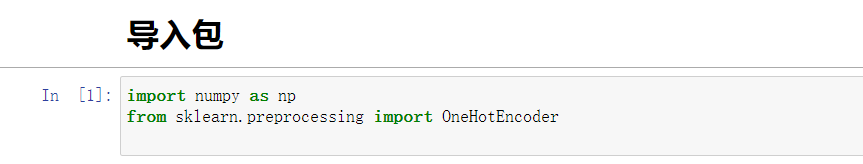
TF-IDF指的是词频-逆文档频率，TF指的是词频（Term Frequency），IDF指的是逆文档频率(Inverse Document Frequency)。其核心思想是：如果某个单词在一篇文章中出现的频率TF高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，即为关键词。

3.Word2vec是Google在2013年开源的一款用于词向量计算的工具，一经发布就引起了工业界和学术界的关注。首先，Word2vec可以在百万数量级的词典和上亿的数据集上进行高效地训练；其次，该工具训练得到的词向量（word embedding），可以很好地度量词与词之间的相似性。Word2vec不是一种深度学习算法，其后面只是一个浅层神经网络，包含两种模型：CBOW模型和Skip-gram模型。

本章实验主要是基于Python和gensim框架实现Word2vec在Wikipedia语料集上面的应用，并且获取词的词向量以及寻找相近词。

# **实验步骤**

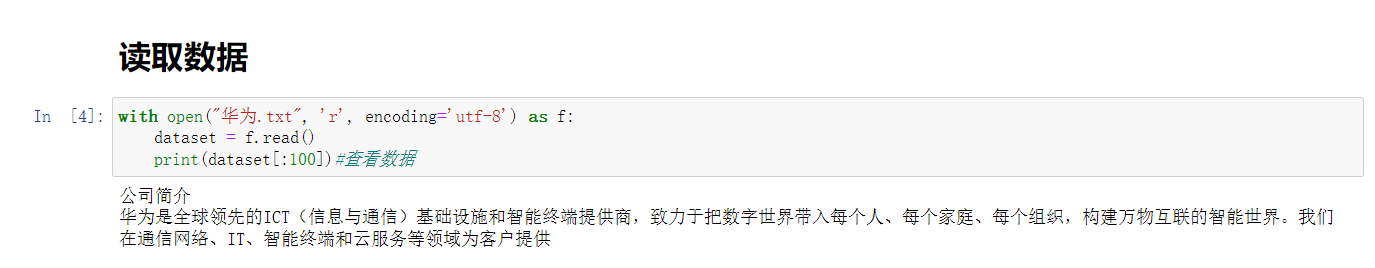
1. One-hot

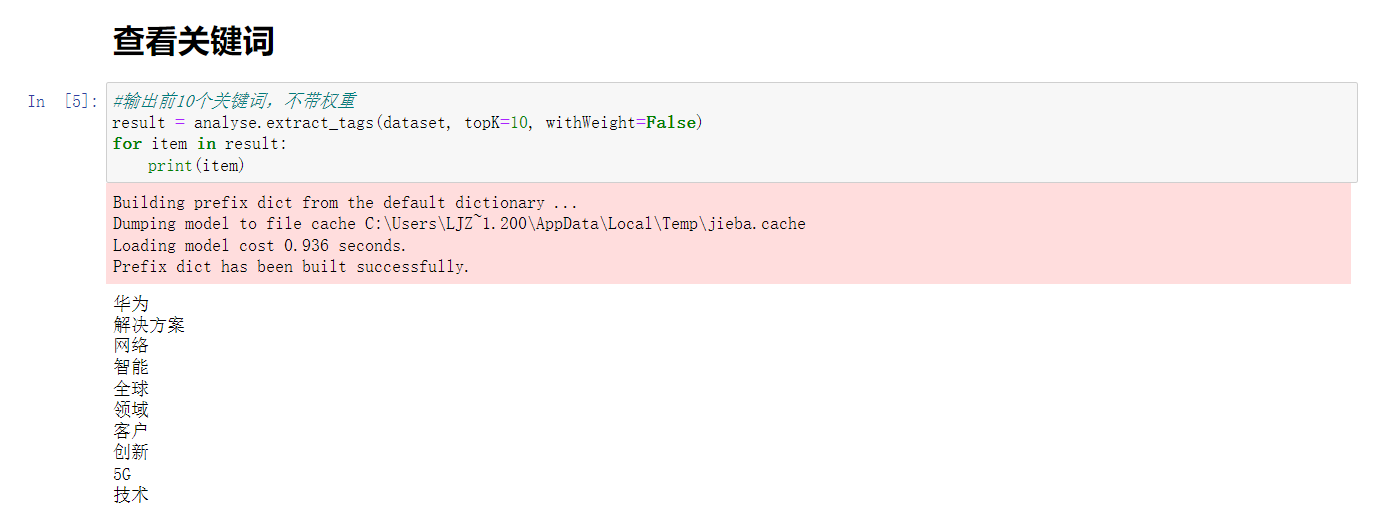




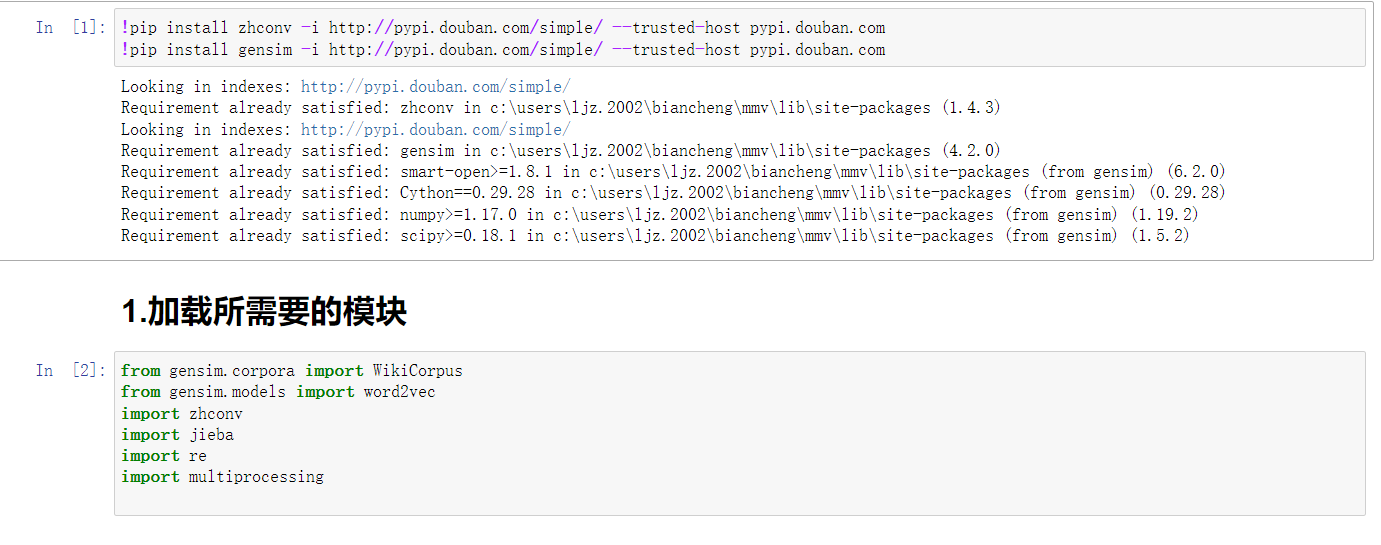
2. TF-IDF



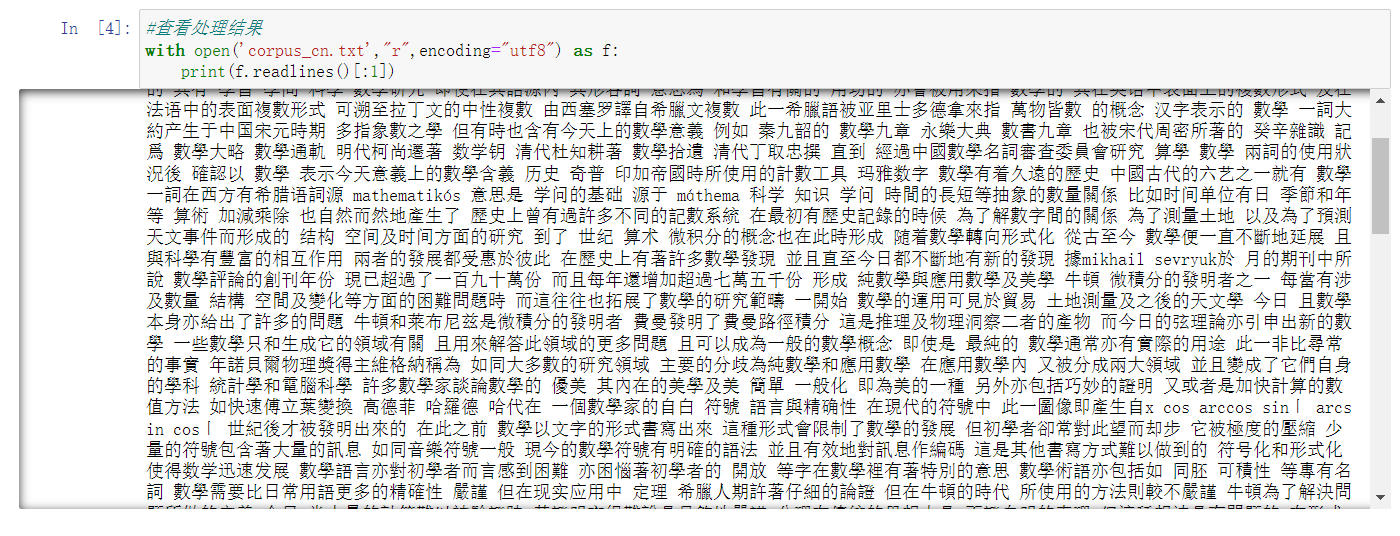




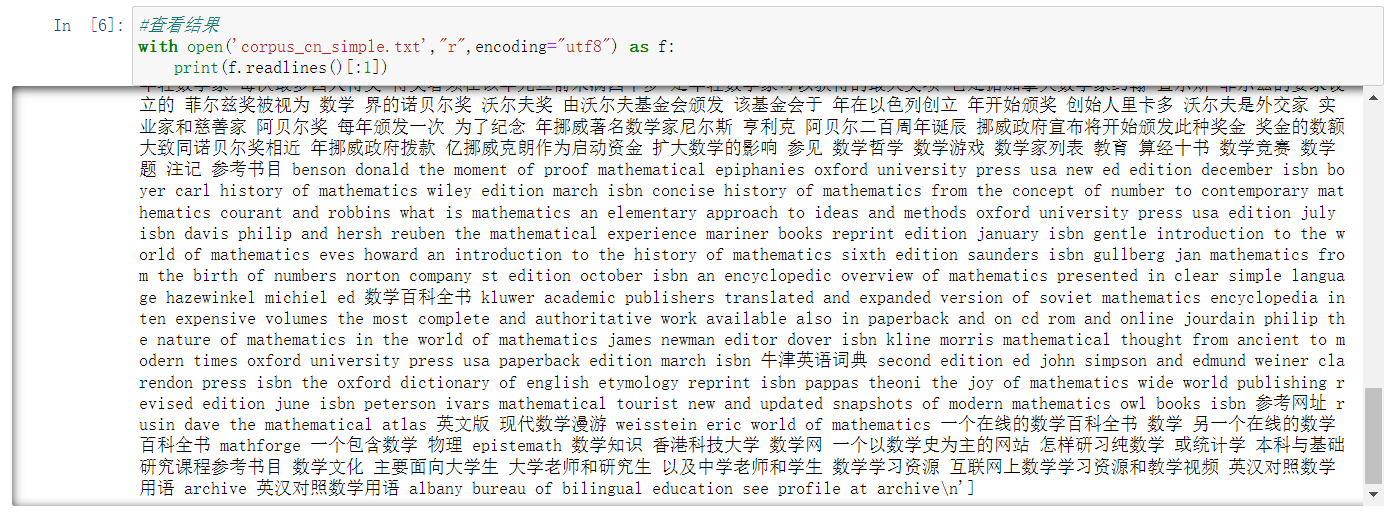
3. Word2vec







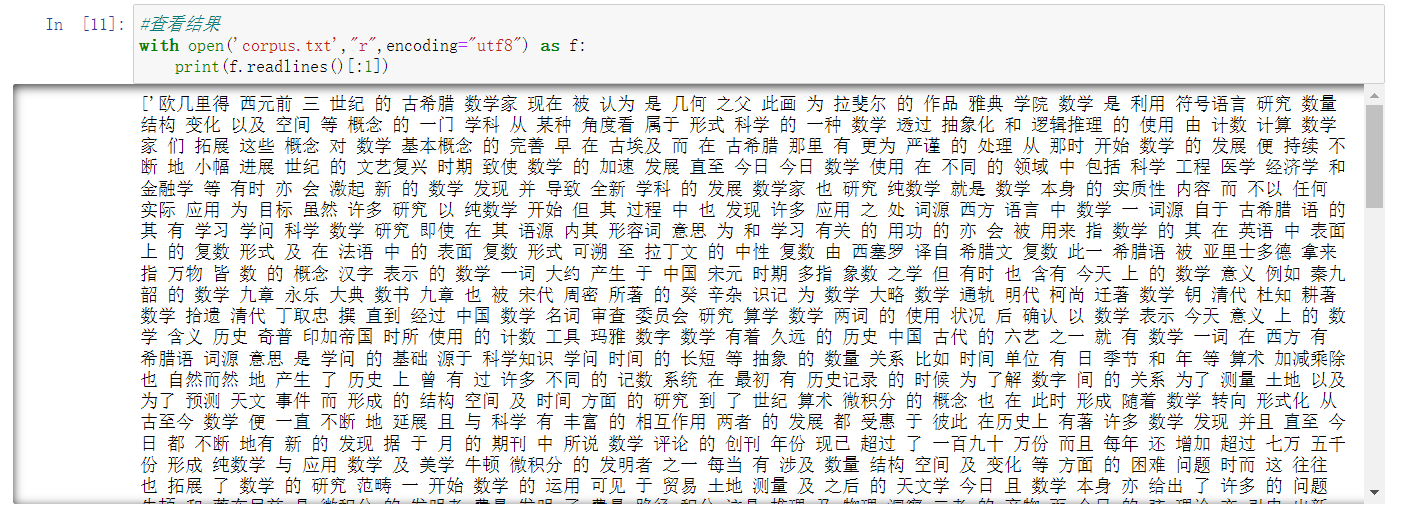
















# **实验验证分析**

1. One-hot：

运行代码，生成所给集合x的独热编码。

2. TF-IDF

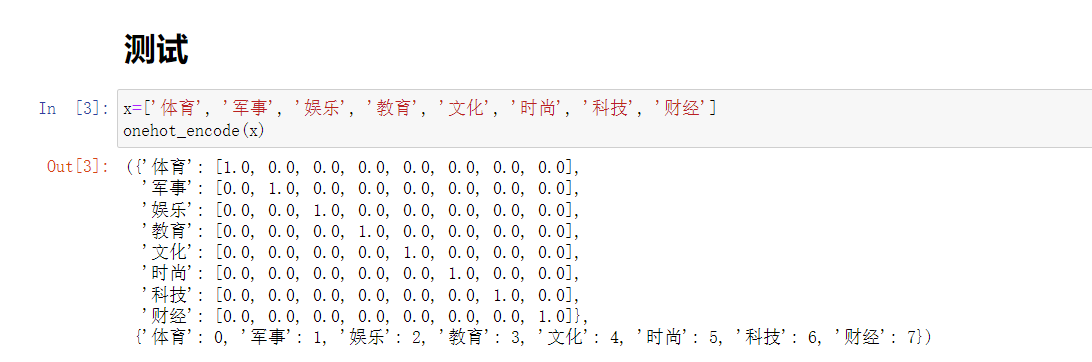
运行TF-IDF代码，分别生成带权重和不带权重的10个关键词

3. Word2vec

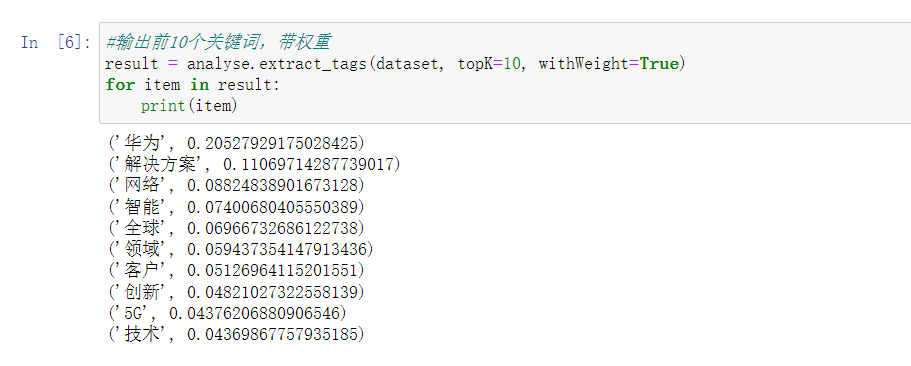
运行Word2vec代码，分别生成与“华为”和“自然语言”最接近的词语。

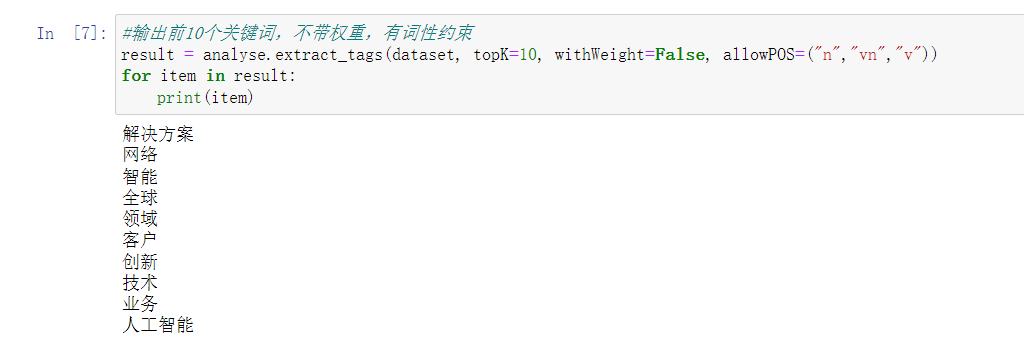
# **实验结论与心得**

1. One-hot



2. TF-IDF





3. Word2vec

