**主题模型**

我们通过两个步骤对Quora社交问答平台上特定关键词的回答内容进行主题分类。

首先，使用Python中的自然语言处理工具包—NLTK（Natural Language Tool Kit）对回答内容进行预处理。针对每个回答，使用NLTK工具包中的sent\_tokenize方法将其拆分为句子。然后，对每个句子进行单词拆分，并使用NLTK工具包中的pos\_tag方法为这些单词进行词性标注。保留名词形式（NN-名词单数、NNS-名词复数）的单词，并使用WordNetLemmatizer词形还原方法将这些名词还原为词根形式（时态、语态、单复数的转换）。最后，采用康奈尔大学提供的停用词列表，删除没有分析价值的单词。这样处理后的文本内容将作为下一步LDA模型训练的语料库。

接下来是模型的训练和主题分析。使用Python中的Gensim工具包实现LDA主题模型分析。通过计算主题的困惑度指标，确定最佳的主题数量。最后，为了获取每个主题内部的关键词及其概率权重，再次训练LDA模型，并设置最佳的主题数量来完成主题分析。

困惑度：在自然语言处理中，困惑度是被广泛应用于衡量模型的拟合效果。通过计算LDA困惑度，可以评估模型对于未见过的新数据的预测能力，同时也可以用于衡量主题模型有效性，将已训练好的模型用于特定领域或任务的数据集上。