**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 sent\_tokenize方法将其拆分为句子。然后，利用word\_tokenize方法对每个句子进行单词拆分，并使用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，同时使用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。困惑度是一个广泛用于评估自然语言处理模型性能的标准，特别是模型拟合数据的程度。因此，利用已经训练好的LDA模型，通过将其应用到特定关键词的回答文本数据集上，基于计算得到的困惑度可以确定最佳的主题数量。最后，一旦确定了最佳的主题数量，我们再次训练LDA模型，并设置确定后的最佳主题数量来得到每个主题中的核心关键词及其在主题中的权重。

**困惑度计算的过程及其具体案例**

**计算过程：**

1. 首先，将给定的文本（例如，关于碳中和的数据集中的一条回答）输入到已经训练好的LDA模型中。

2. 然后，在这条回答的每个单词上遍历。

3. 对于每个单词，我们观察其在每个主题对应关键词中的权重（prob\_topic\_word）。这可以被看作是该单词属于特定主题的概率。

4. 接着，将主题对应关键词中的权重（prob\_topic\_word）与这条回答对应的主题权重（prob\_topic）相乘，得到单词的条件概率（prob\_word）。这可以被看作是给定该回答的前提下，单词出现的条件概率。

5. 之后，对这一条回答的所有单词重复3-4这个过程，并累加这些单词条件概率（prob\_word），就得到了这条回答的总概率分布（prob\_doc）。

6. 然后，将数据库中碳中和相关的所有回答重复1-5过程，并将这些回答的总概率分布进行累加，得到碳中和相关文本的总概率分布（prob\_doc\_sum）。

7. 最后，我们通过如下公式对模型的困惑度进行计算：

其中是计算得到的困惑度，是过程6计算得到的概率分布累加和，是数据库中碳中和相关所有回答单词的数量。

**具体案例：**

假设以“carbon neutral ….”这条回答为例。

1. 将这条输入到训练好主题数量的LDA模型中，假设得到[(0, 0.35), (1, 0.2), (2, 0.45)]，表示这条文本对应主题1的概率为0.35，主题2的概率为0.2，主题3的概率为0.45。

2. 假设该LDA模型的3个主题及其关键词和权重为：

【0， carbon\*0.08 neutral\*0.05, climate\*0.01】

【1， co2\*0.05 carbon\*0.03 neutral\*0.01】

【2， neutral\*0.05, carbon\*0.02, peak\*0.01】

3. “carbon neutral ….”包括carbon、neutral两个单词，carbon在LDA模型的3个主题中关键词权重（prob\_topic\_word）分别为0.08、0.03、0.02。

4. 由于“carbon neutral ….”在第1个主题中的权重（prob\_topic）为0.35；在第2个主题中的权重（prob\_topic）为0.2，在第3个主题中的权重（prob\_topic）为0.02，因此carbon单词的条件概率（prob\_word）为：0.08\*0.35+0.03\*0.2+0.02\*0.45=0.043

5. 之后对“carbon neutral ….”中的每个单词都采用上述4过程的计算方式并累加得到这条回答的总概率分布（prob\_doc）

6. 之后对碳中和相关4359条回答都依次计算prob\_doc并累加得到碳中和相关文本的总概率分布（prob\_doc\_sum）

7. 最后，我们通过如下公式对模型的困惑度进行计算：

其中是计算得到的困惑度，是过程6计算得到的概率分布累加和，是数据库中碳中和相关所有回答单词的数量。