**2022秋自然语言处理大作业**

刘轲迪 2019200076

* 任务描述

使用自然语言生成技术实现标题抽取(生成)功能，并通过UI界面实现展示和交互。

* 数据介绍

文本数据来源于新浪新闻，分别抽取新闻正文与标题，约68万条新闻数据。

* 实现方法

GPT2模型：模型由多层单向 Transformer 的解码器部分构成，本质上是自回归模型，自回归的意思是指，每次产生新单词后，将新单词加到原输入句后面，作为新的输入句。在具体实现中，主要对transformers包中GPT2LMHeadModel类进行重写，修改计算loss部分，只计算预测title部分的loss。

解码策略：仅保留top\_k个以及累积概率到达top\_p的标记，其他标记设为filter\_value，后续在选取标记的过程中会取不到值设为无穷小。

UI界面：通过streamlit插件实现交互界面

* 运行代码

streamlit run app.py