NLP 自然语义处理文档

1. 什么是自然语义处理？

自然语义处理（NLP）是计算机科学领域和人工智能领域中的一个分支，它与计算机和人类之间使用自然语言进行互动密切相关，主要致力于让计算机能够理解和生成人类语言。语言的两大要素，句法和语义。“句法”是指文本的语法结构，“语义”指的是由它表达的含义。所谓语义处理就是实现对文本，或是音频，或是图像的处理，挖掘其中的含义。

1. 为什么要进行自然语义处理？

自然语义处理涉及的范围极广，包含词语预测，语言识别，文本纠错，英汉互译，情感分析等等非常多，所以其带来的商业应用也非常有用。例如在该项目中会涉及到的舆情分析，涉诉分析。通过舆情与涉诉分析可以得到大众对于某行业的某些公司或地区独有的态度或某些特点。

1. 什么是舆情分析？

舆情可以理解为是在各大平台，论坛，新闻，微博等出现的大大小小的事件，评论，其往往带有个人的情感导向。通过收集关于特定地区，行业，公司的舆情可以分析总结出关于地区，行业，公司的态度与可能出现的风险。

1. 怎么进行舆情分析？

以此煤炭项目中的数据为例，该项目中收集的数据为同花顺数据库中的新闻舆情数据，数据量 18450x12。在做类似的舆情分析的时候，我们的目标主要是从时间，公司，地区，风险等多个维度找到各个维度上的特点，例如：在山东的公司普遍较多发生信用预警的风险。首先，由于技术成熟的程度，我们需要先剔除数据量少的年份。对于公司-风险为度稍作处理分析后，就可以对新闻舆情的内容进行处理。在对新闻舆情的内容处理前，我们需要先将其内容中的标点符号全部去掉。之后，这里由于中文的词法，语法，句法与英文完全不同，所以需要调用一个特别针对中文处理的包，jieba（其具体的各种function都可以在其网站查到，如不同的分词方法）利用jieba，我们可以把舆情新闻的内容拆分成有实际含义的词或句。拆分好之后我们就可以根据词与句的出现次数的多少，大致的总结出地区，时间，风险，公司等维度之间的关系。

1. 代码

<https://github.com/JackieSun66/Glp_Job/blob/main/%E7%85%A4%E7%82%AD%E6%96%B0%E9%97%BB.ipynb>

六. 做图

做图的话，可以选择tableau进行可视化，preview 如下：

Chart, scatter chart

Description automatically generated

除了tableau之外，也可以选择高的开发者平台的可视化工具进行可视化。效果如下：

https://maplab.amap.com/share/mapv/90555a6a8c71ccfdbd884e551611a755