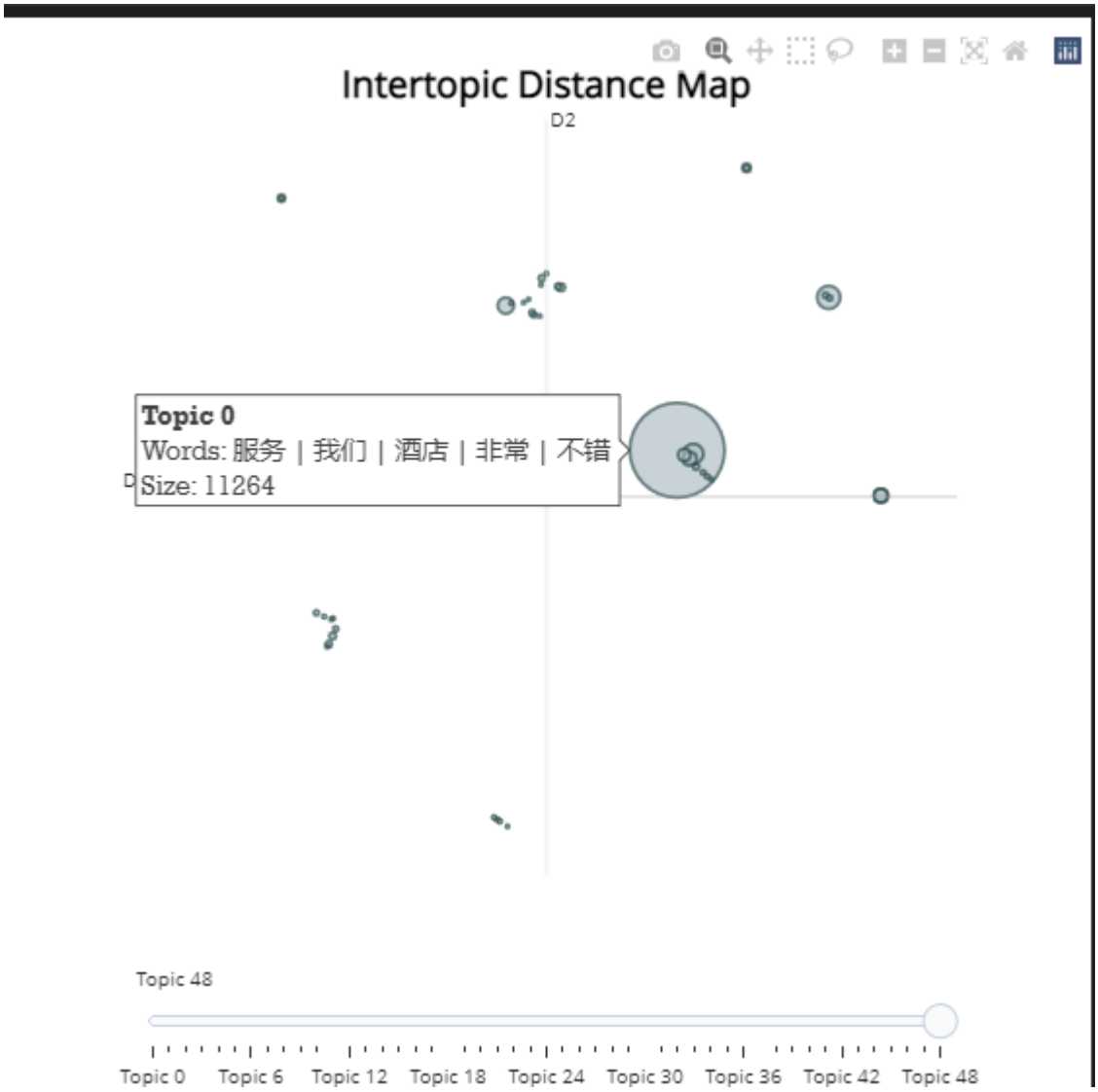
遇到了一些问题向老师反馈一下：

进展部分

1. bertopic能够跑通并运行，出图了。但是效果我觉得非常不好（一次训练大概是45min，非常费时间）

我是基于分词之后的90家酒店评论数据，大概**43178**条评论数据。训练的bertopic模型，没有调参之前跑出来了快两百个主题，我调参之后将主题限制到了50个，但是我大概看了一下词距离图感觉分类效果不好，很多选出来作为一个主题的所谓“关键词”，其实没有什么实际意义。比如一些语气词和人物姓名。




```
topic_model.visualize_barchart()
```

✓ 1.8s

疑惑的部分

1. 中文文本分词不太好搞，感觉bert在英文评论上的效果应该要好一点，因为中文的同义词太多了不太好处理。分词部分我是这样做的：

```
def clean_text(text):
    try:
        # 非中文先全去掉
        text=re.sub("[^\u4e00-\u9fa5]",'',text)
        # text=re.sub(r'[A-Za-z0-9_]','',text)
        cutor=jieba.lcut(text)
    except:
        return ''
    # 空格分词
    return ' '.join([i for i in cutor])
a="酒店离西安火车站一街之隔，但算是闹中取静；服务员非常热情，对待客人礼貌有加；房间内卫生良好，隔音情况良好；免费的无线网络连接比较顺畅。"
clean_text(a)
```

结果是：

```
'酒店 离 西安 火车站 一街 之 隔 但 算是 闹中取静 服务员 非常 热情 对待 客人 礼貌 有加 房间内 卫生 良好 隔音 情况 良好 免费 的 无线网络 连接 比较 顺畅'
```

结果还不是很好看。一些无关的词语还没有被去掉？

- 我的一个想法是，利用jieba库按照词性选一些词（比如我只要名词，形容词这些，然后把这些当作短语放入bertopic训练？还是说要把整个句子分词然后选关键词，把topxx的关键词放入bertopic训练？）

2. **样本不平衡的问题还没有解决**，我看了一些方法，都是针对于float、int类型的数据结构进行重采样的（SMOTE、ADASYN等oversample方法），对于文本型数据怎么进行重采样呢？
    1. 首先bertopic也是基于tfidf算法，词频就非常重要，如果盲目增加 `label=-1`的样本的数量，那么我觉得应该会对结果造成很大影响？
3. 如果得到了比较好的模型（主题比较明确，可解释性好），怎么利用模型预测label呢....感觉这个bertopic是一个无监督，只要扔数据进去就能输出话题。怎么和我的label对应上呢？