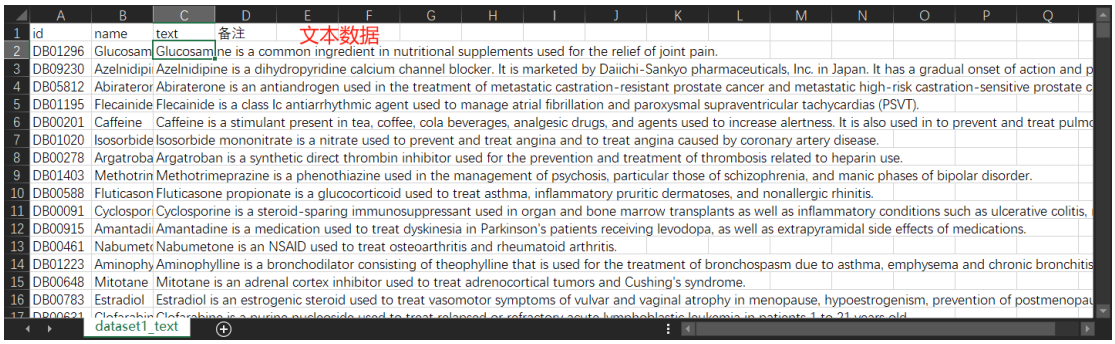
数据

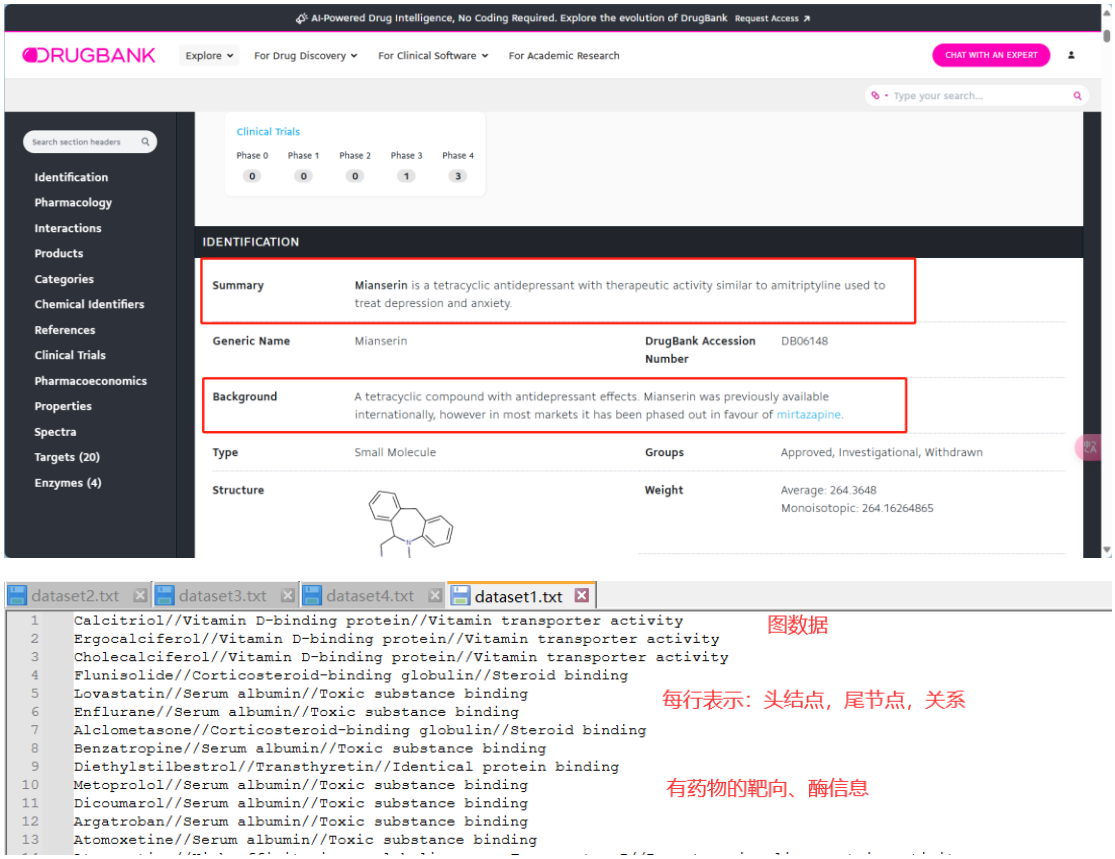
大致完成文本数据收集、对齐

进展

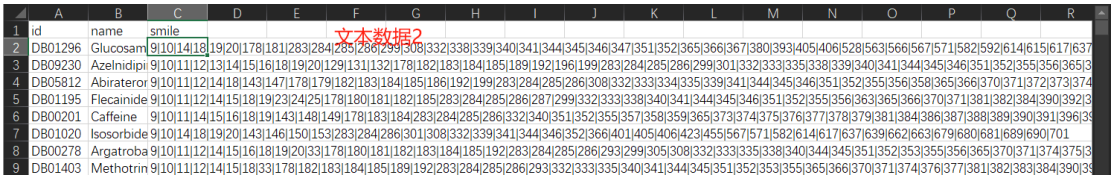
dataset1

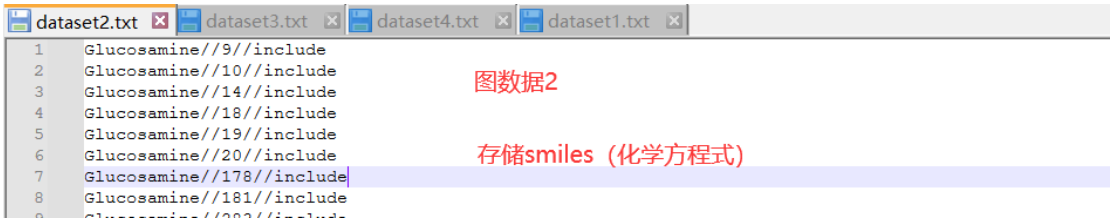


文本数据来自于 drugbank 数据库中对每个药物的 Summary 和 Background

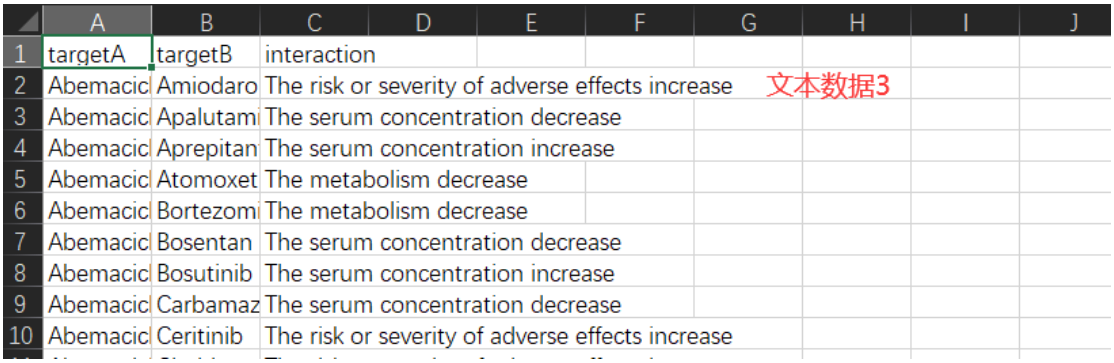


dataset2

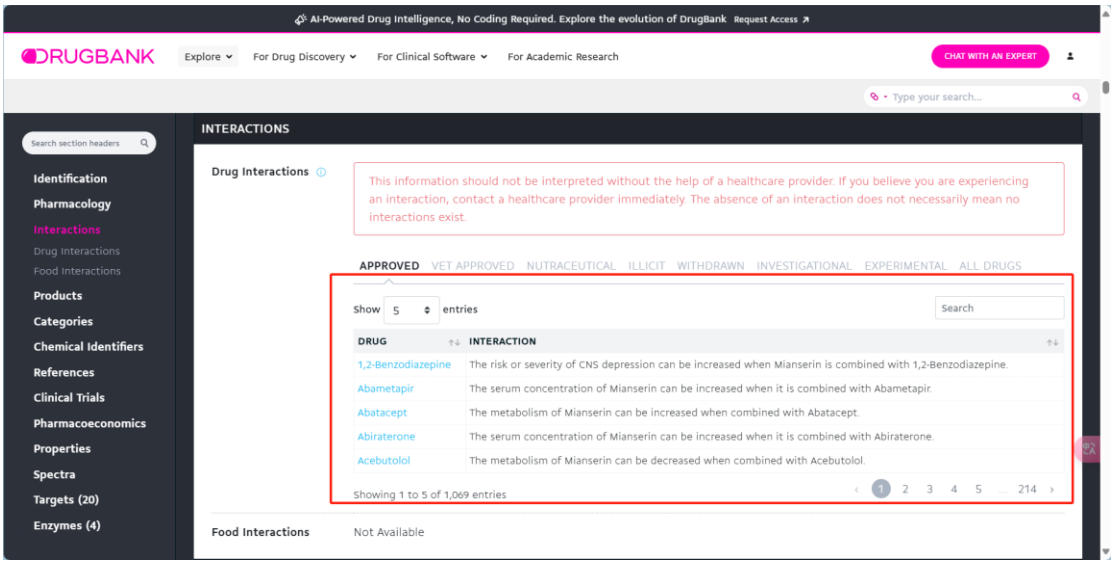


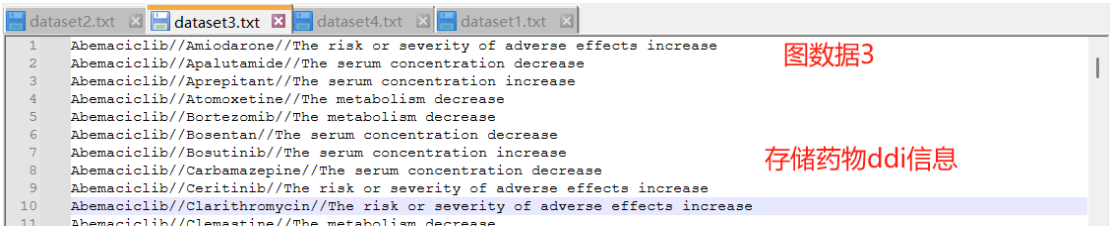


dataset3

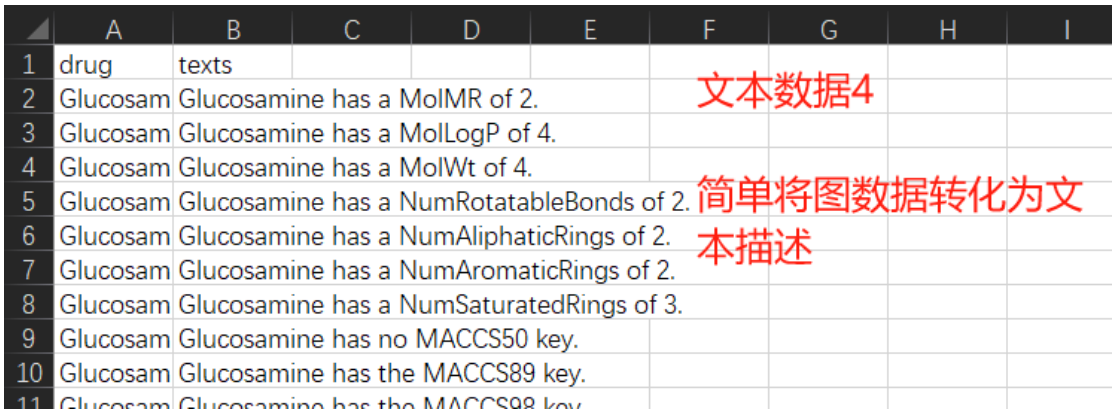


来源于 drugbank

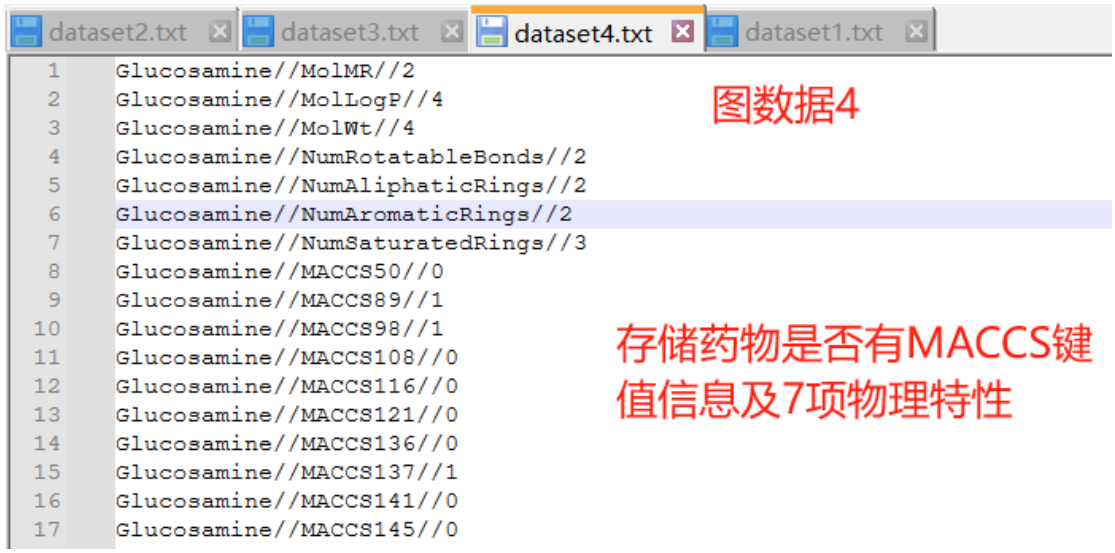




dataset4



	A	B	C	D	E	F	G	H	I
1	drug	texts							
2	Glucosam	Glucosamine has a MolMR of 2.							
3	Glucosam	Glucosamine has a MolLogP of 4.							
4	Glucosam	Glucosamine has a MolWt of 4.							
5	Glucosam	Glucosamine has a NumRotatableBonds of 2.							
6	Glucosam	Glucosamine has a NumAliphaticRings of 2.							
7	Glucosam	Glucosamine has a NumAromaticRings of 2.							
8	Glucosam	Glucosamine has a NumSaturatedRings of 3.							
9	Glucosam	Glucosamine has no MACCS50 key.							
10	Glucosam	Glucosamine has the MACCS89 key.							
11	Glucosam	Glucosamine has the MACCS98 key.							



dataset2.txt | dataset3.txt | dataset4.txt | dataset1.txt

```
1   Glucosamine//MolMR//2
2   Glucosamine//MolLogP//4
3   Glucosamine//MolWt//4
4   Glucosamine//NumRotatableBonds//2
5   Glucosamine//NumAliphaticRings//2
6   Glucosamine//NumAromaticRings//2
7   Glucosamine//NumSaturatedRings//3
8   Glucosamine//MACCS50//0
9   Glucosamine//MACCS89//1
10  Glucosamine//MACCS98//1
11  Glucosamine//MACCS108//0
12  Glucosamine//MACCS116//0
13  Glucosamine//MACCS121//0
14  Glucosamine//MACCS136//0
15  Glucosamine//MACCS137//1
16  Glucosamine//MACCS141//0
17  Glucosamine//MACCS145//0
```


问题

1、dataset2 中的文本描述可以用 CN1CCN2C(C1)C1=CC=CC=C1CC1=CC=CC=C21 该种方式表示，看后续实验效果是否修改

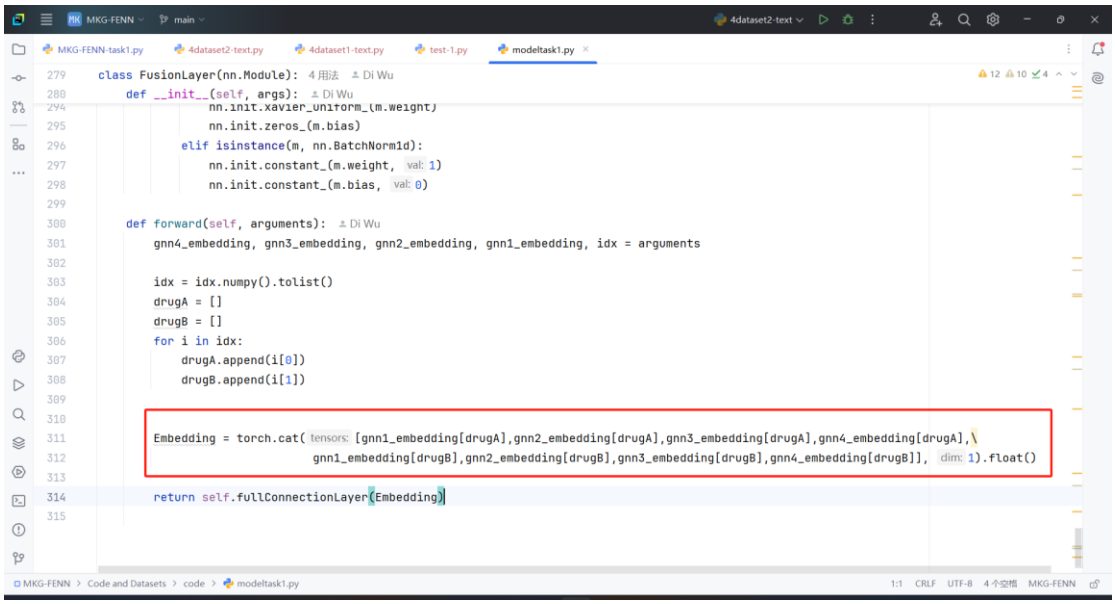
2、dataset4 中还没找到关于药物 MACCS 键值的描述，所以暂时只是简单将图数据转化成文本数据

模型、代码

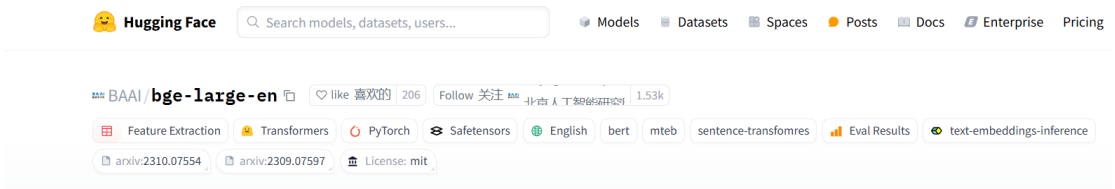
进展

代码方面还没做出修改

初步打算在代码的 Embeding 那边拼接文本数据的向量，然后在 Embeding 加一个 LLM 充分融合两种向量。



准备使用 bge-large-en 作为 LLM



问题

1、文本数据通过 LLM 输出的向量是否需要动态？

目前打算先用静态的试一试，后续试试动态。

流程

整个项目实现的流程是清楚了

下一步计划

完成文本数据向量和图数据向量的拼接、融合