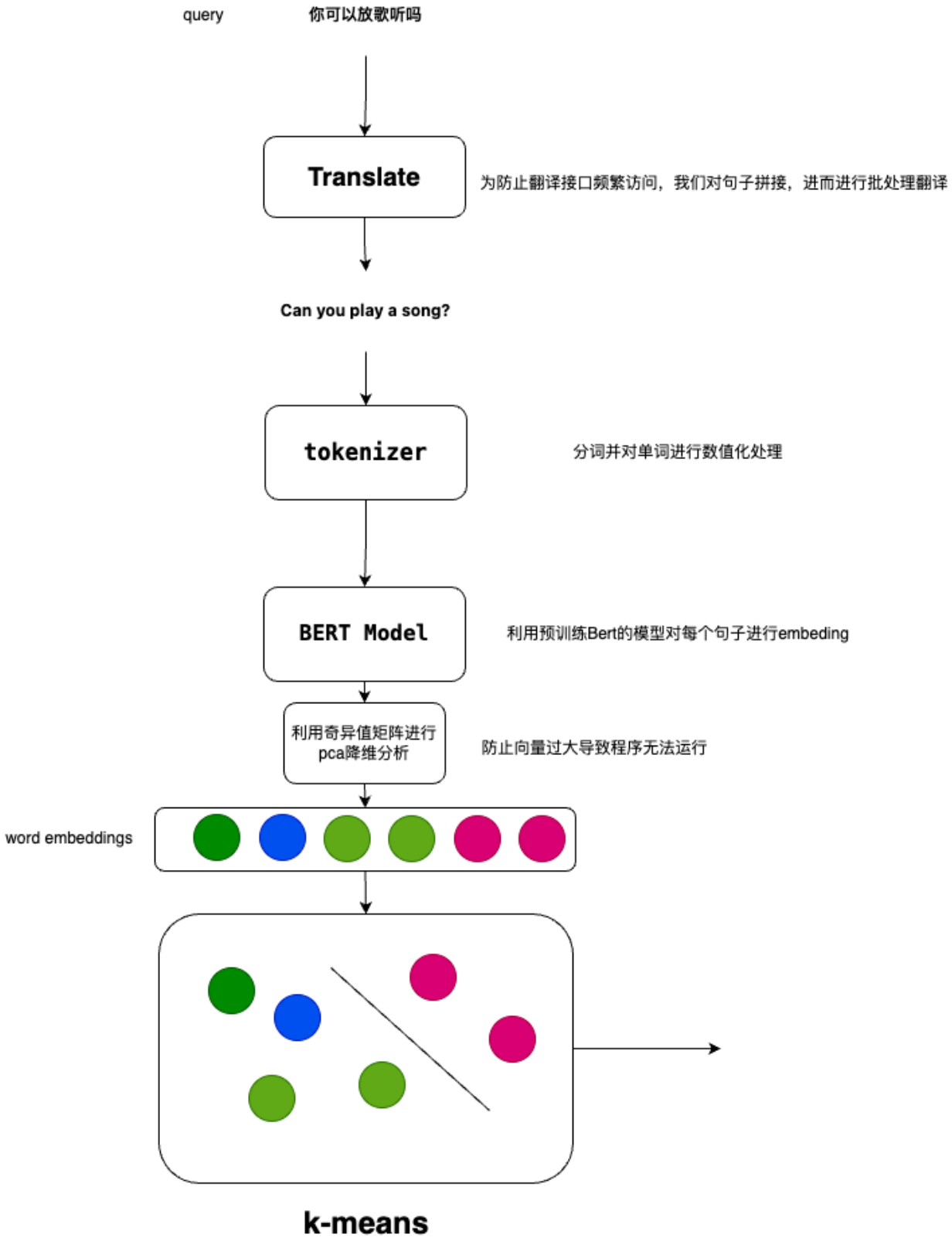
# 异常query分析算法



## 设计思路




# 思路

## Translate

对query进行翻译

主要问题：采用api接口对句子进行翻译时，由于query较多会造成网页无返回，连接超时等问题，

解决：将多个句子合并一同翻译，然后根据设置的分隔符将其拆分。

## Tokenizer

将query转化为token，但是此token只是对句子中每个单词的编号，无距离相似度意义。

## BERT Model

将token转换为带有单词间关系信息以及相似度信息的向量

## 奇异值分解

利用奇异值矩阵计算token的特征值，利用特征值去在空间上压缩特征，将其压缩到d_model大小（这里d_model设置为768）

## 聚类分析

根据k-means算法通过余弦相似度将query划分为2类，及异常query和正常query。

# 结果

我们对domain为music的类别进行划分：划分效果如下：

| 异常query | 正常query |
| --- | --- |
| 奥秘 | 请唱一首歌 |
| 学猫叫 | 放首歌 |
| 归位 | 放首歌 |
| 归位 | 唱个歌 |
| 小鸡小鸡 | 你可以放歌听吗 |
| 播放今天 | 给我唱首歌呗 |
| 怎样 | 放学的什么歌 |
| | |

| | |
|---|---|
| 漂亮 | 可以为我唱首歌吗 |
| 王八蛋 | 我懂了 |
| 回来 | 妹妹房间 |
| 你们 | 想我了吗 |
| 臭臭 | 我想你了 |
| 湿布 | 春眠不觉晓 |
| 找妈妈 | 来找我 |
| 边刷边刷 | 来找我 |
| 你会不会 | 放一首生日快乐歌 |

# 问题以及后续改进方法

问题：

- 对“播放音乐”等query存在一些划分错误的现象，划分的准确率有待提高
- 模型对query的划分过渡依赖与向量的表示。
- 模型在降维时，虽然降低了计算消耗和内存占用，但是一定程度上，丢失了数据的空间维度信息。

后续改进：

- 在特征表示上改进：利用实体抽取模型对句子中的实体关系进行抽取，然后构造一个自己的知识图谱，引入知识，来指导模型进行划分。（缺点：需要大量的数据跟时间去构建图结构，知识如何融入模型需要花费时间去构建）
- 改进聚类算法，采用密度的方法DBSACN, OPTICS, DenClue等或者普聚类算法等无监督算法对其进行改进
- 改造特征：构建特征均值、方差等信息，以及其他相似度信息作为决策变量