[MySQL utf-8 utf8mb4](http://www.csdn.net/tag/MySQL%20utf-8%20utf8mb4" \t "_blank)

问题描述：从新浪微博抓取消息保存到MySQL数据中，对应数据库字段为varchar，字符编码utf-8。部分插入成功，部分插入失败，报错如标题。

在网上查询，有人说是编码问题，建议修改编码格式，比如改成gbk，UTF-8，blob等等，但是几乎没有人给出更详细的答案。在一个英文网站上，才发现真正错误的原因。[链接1](http://stackoverflow.com/questions/8709892/mysql-throws-incorrect-string-value-error" \t "_blank) [链接2](http://stackoverflow.com/questions/7814293/how-to-insert-utf-8-mb4-characteremoji-in-ios5-in-mysql)

**错误原因：**我们可以看到错误提示中的字符0xF0 0x9F 0x98 0x84 ，这对应UTF-8编码格式中的4字节编码（[UTF-8编码规范](http://baike.baidu.com/link?url=YC0F2YQes6HZRQCQpApLcNGdts0HTcGXEvHLFEfdjGh3jVrIITkQAUo4c53uK8fFqo1OvmNKOGeiMc0BuFGbBa" \t "_blank)）。正常的汉字一般不会超过3个字节，为什么为出现4个字节呢？实际上是它对应的是智能手机输入法中的表情。那为什么会报错呢？因为mysql中的utf-8并不是真正意义上的utf-8，它只能存储1~3个字节长度的utf-8编码，如果想存储4个字节的必须用utf8mb4类型。不而要使用utf8mb4类型，首先要保证Mysql版本要不低于 MySQL 5.5.3。

解决方案：

1）使用utf8mb4数据类型

   要用这种策略，如果MySql版本低于5.5.3，首先要进行版本升级，然后将对应的数据类型改为utf8mb4类型。如果使用的是Connector/J 连接数据库，需要在配置中把编码格式改为utf8mb4（set character\_set\_server=utf8mb4 in the connection config）。

2）自定义过滤规则，将文本中出现的四字节UTF-8字符过滤或转化为自定义类型。

    下面是将4字节字符转化为0000的测试例子。

**[java]** [view plaincopy](http://blog.csdn.net/wskings/article/details/12857993)

1. for (int i = 0; i < b\_text.length; i++)
2. {
3. if((b\_text[i] & 0xF8)== 0xF0){
4. for (int j = 0; j < 4; j++) {
5. b\_text[i+j]=0x30;
6. }
7. i+=3;
8. }
9. }