

# MySQL, Python e UTF-8

Contribuição de Thomas Lopes  
07 de July de 2008  
Última Atualização 07 de July de 2008

Esse final de semana, lá estava eu, trabalhando em meu TCC, coletando dados da Internet... quando de repente, me deparo com uma barreira não prevista: problemas de conjunto de caracteres.

Até então, estava tudo correndo bem. O BeautifulSoup (um parser [X]HTML/XML que ajuda a processar conteúdo web) trazendo tudo para Unicode, e a escrita de arquivos correndo perfeita. Então, parti para a escrita dos dados no banco, MySQL (banco de dados free, open-source, recentemente comprado pela SUN), e aí começaram os problemas: Percebi que alguns conteúdos estavam entrando no banco com caracteres bizarros no meio das palavras, como 'notÃ-cias', 'ConteÃ°do', etc... dependendo da fonte (página) onde se encontra o texto (quando as páginas não tinham a codificação correta no cabeçalho, principalmente).

Mesmo utilizando o decode certo ( `string.decode('utf-8')` ), continuava a receber erros e caracteres estranhos. As tabelas e campos já estavam com seu collation type configurado corretamente para esse caso (`utf8_general_ci`), e mesmo assim, o problema persistia

Pronto, só restava: revisão do código.

...Nenhum problema. Aí resolvi testar escrevendo os dados num arquivo texto normal, antes de gravar no banco. Resultado: os arquivos eram gravados normalmente, em UTF-8, sem problemas: Conclusão: problema com o Banco de dados, ou com o conector utilizado (MySQLdb). Tentei de tudo, vários decodes, encodes, e continuava obtendo erro:

```
UnicodeEncodeError:'latin-1' codec can't encode character ...
```

Até que encontrei esse post do blog Dasprid's, onde o autor mostra como sanar esse mesmo problema. Simples de tudo. basta setar algumas variáveis de ambiente do MySQL para que funcione corretamente com a codificação UTF-8 (aliás, recomendo usar UTF-8 em tudo viu. Recomendo a leitura desse artigo: Tudo Sobre Python e Unicode ). Para se ter uma idéia, basta fazer os seguintes comandos:

```
db.set_character_set('utf8')
dbc.execute('SET NAMES utf8;')
dbc.execute('SET CHARACTER SET utf8;')
dbc.execute('SET character_set_connection=utf8;')
```

Onde db é o seu objeto connect, e dbc o objeto cursor. Muito simples, e resolveu um problema que arranca cabelos de muitos. E mais umavez, o Python demonstrando como as soluções podem ser simples, simples mesmo.