String encoding in Python

Simon Funke, Center for Biomedical Computing, Simula Research Laboratory & Dept. of Informatics, University of Oslo

Sep 22, 2015

Introduction

Seen this before?

```
UnicodeDecodeError: 'ascii' codec
can't decode byte 0xc4 in position
10: ordinal not in range(128)
```

Then you are not handling strings correctly in Python!

These slides are based on: Unicode In Python, Completely Demystified

String handling in Python

Lets read a UTF-8 file with some special characters (the word Bokmål)

```
#!/usr/bin/env python

import sys

# wget http://fil.nrk.no/yr/viktigestader/noreg.txt

f = open("noreg.txt", "r")
s_utf8 = f.readline().split("\t")[12]

print type(string_utf8)
s_utf8
```

- s_utf8 is a string