uc3m | Universidad Carlos III de Madrid

OPENCOURSEWARE
ADVANCED PROGRAMMING
STATISTICS FOR DATA SCIENCE

Ricardo Aler

ADVANCED PROGRAMMING
MASTER IN STATISTICS FOR DATA SCIENCE.
PYTHON PROGRAMMING ASSIGNMENT: PROGRAMMING BAG-OF-WORDS FEATURE EXTRACTION

2.5 POINTS

Introduction

The aim of this assignment is to program in Python 3.7 the transformation from unstructured data (text) to structured data (a matrix) known as bag-of-words. This transformation is widely used in text analytics (such as automatic classification of texts into categories). There are libraries (such as NLTK) that automatize the process but in this assignment, you will have to program it in Python both as a way of understanding the process and a practicing Python programming.

What to do:

1. Read the notebook in Aula Global.
2. Fill in with your code the empty cells.
3. Hand in your complete notebook in Aula Global. No report is needed but you can write comments in the notebook.