Assignment 0

Obligatory to submit (not graded)

The purpose of this assignment is to prepare the basis for the "true assignments" that will follow soon. You are supposed to form teams of 3 students, work on the task, produce a short report and submit it to the BS system. This assignment is not graded, but you will get a "first impression" of what is expected from you and how the whole submission procedure work.

Suppose you generate *n* random integers between *1* and *N* (with *n* possibly bigger than *N*). How many *distinct values* do you expect to get? (Note that your random numbers might repeat, so even after generating *N* integers you may see less than *N* distinct values!) Find a formula that expresses this expected number of distinct values: *distinct_values(n, N)* = ?

If you have no idea how to start, imagine *N* empty buckets and *n* coins that you throw, one by one, at random, to the buckets. Select a bucket (e.g., bucket number 7) and answer the following questions: what is the chance that your bucket will stay empty after throwing the first coin? And after throwing the second coin? And in general, after throwing all *n* coins? You should get a formula that resembles the classical definition of the constant *e* which is defined as the limit of the sequence $(1+1/k)^k$.

If you are still unable to find the answer, study section 1.3.5 of the textbook "Mining of Massive Data Sets" (you can download it from http://www.mmds.org/#ver21). In the unlikely case that after reading this section you still don't know how to proceed, study sections 4.3.2 and 4.3.3 of this book.

Experimentally verify your findings: implement (in Python) the whole experiment and run it several times with various values of *n* and *N*. Are the results of your experiments consistent with your formula?

Don't ask us about the values of n and N that you are supposed to play with! The decision is up to you – in "real life" it is a data scientist who makes such decisions.

Deliver a single Jupyter notebook that documents your theoretical considerations, experiments, and conclusions. Submission should be done through the BrightSpace platform (*Assignments*).

Deadline: Monday, September 25, 23:59.