

## **Big Data Analytics: questions/topics for MSc exam**

1. How can we represent a graph and what are its basic characteristics?
2. What is the Central Limit Theorem?
3. What is the Jaccard similarity?
4. What is a random variable? Give an example. What does it mean that two random variables are independent? What is the probability density function (PDF) and what is the cumulative distribution function (CDF)? What is the relation between them?
5. Formulate and explain the Bayes formula.
6. Formulate Markov inequality.
7. Explain the basic properties of the PageRank algorithm.
8. Explain K-means algorithm
9. What are the basic properties of a metric function?
10. Explain the basic assumptions of the MapReduce paradigm
11. Name basic characteristics of stream algorithms. What algorithm uses the sliding window model?
12. Give examples of procedural, object oriented and functional programming languages. Characterize them briefly.
13. What is i) eta-expansion, ii) lambda-function, and iii) currying? Give example of these constructs in Scala
14. What problem solves the HyperLogLog algorithm? ii) What gave the name to the algorithm? iii) How the accuracy of the algorithm scales with the number of counters?
15. What is entropy in physics and in information theory? Why is it important for data science?
16. Models of complex networks and their typical properties in real life
17. Classification of simple stationary points of a system of two autonomous first order differential equations
18. Explain the perceptron algorithm. What is overfitting and how to prevent it?
19. How do Support Vector Machines work?
20. Discuss dimensionality reduction techniques.
21. What is the difference between supervised and unsupervised learning? Give some examples of problems where we can use unsupervised learning.
22. What problems may occur if we build too complex decision tree relative to the amount of data, and how to deal with them.
23. What is a generative adversarial network, and how to train such a network.
24. Describe the family of exponential smoothing forecasting models.
25. Auto-regressive models in time series forecasting (ARMA, ARIMA, SARIMA, SARIMAX).
26. Multi-layer perceptrons, recurrent and convolutional neural networks for time series forecasting.
27. Methods of analysis of complex/chaotic behavior on the basis of famous nonlinear dynamics models.
28. What is the Monte Carlo (MC) method? What mathematical theorem is related to MC? Metropolis algorithm. Analysis of MC data for Ising model

29. Pseudo-random numbers and generators. Desired features of pseudo-random generators for Monte Carlo simulations. Generation of pseudo-random numbers with various probability density functions.
30. Markov chains and their applications in statistical physics.