Barbara Hammer, Philip Kenneweg, Bianca Schröder Universität Bielefeld

Maschinelles Lernen im Web (WiSe 24/25)
1. Sheet

Start: Monday, 14.10.2024.
End: The worksheets should be solved using Python, in groups of 2-3 people and will be presented in the Tutorials.
Discussion: Monday or Thursday, 28.10.2024 in the Tutorials.

Information

The worksheets and necessary toolboxes will be made available in the Lernraum "392221 Maschinelles Lernen im Web (V)". Worksheets will usually be released every two weeks on Monday, and discussed during the tutorials two weeks later. In order to successfully finish the course, 50% of the available points have to be obtained and each participant has to present his/her results at least once. The week in between the release and discussion of the sheet will be used to discuss the implementation of various algorithms presented in the lecture, as well as go deeper into the relevant material.

Exercise 1: Theory

(*0 Points*)

Which of the following statements is true? This will not be graded but is similar to the final exam questions.

(a) The expected number of persons in a group of 42 persons who have birthday on the same day is between 4 and 5.

(b) Erdösz-Renyi can be modeled via a link probability p of a link in between two specific nodes or, as an alternative, as the choice of n links out of all possible connections in between N nodes. The correspondence is $p = n/N$ to give the same expectation.

(c) For a number of N nodes, the degree distribution of the Erdösz/Renyi model is a Poisson distribution and the degree distribution of the Barabasi/Albert model is a power-law distribution

(d) Nodes in a Barabasi/Albert graph have very different numbers of vertices and are inhomogeneous, because the node degree distribution can be approximated by a power law distribution which has infinite moments.

(e) The more incoming edges, the larger the pagerank of a node.

(f) When starting with random initialization, the pagerank converges to a value per node which is proportional to the probability of the number of visits of the page in a random walk through the graph

(g) Given a number of pages, one can increase the pagerank of a node in the graph beyond limits by a careful design of the link structure.

(h) ChatGPT takes into account link structure.

(i) ChatGPT is trained based on next token prediction, hence no human labeling which might be unethical is required

(j) There are no open source alternatives to ChatGPT

Exercise 2: Practical

(*10 Points*)

You can use all sources from the internet, but you must add a reference. Please add a link to your code (e.g. github / gitlab / ...). Please solve and prepare a short presentation of the methods and results. Each group needs to prepare solutions for all of the tasks and needs to present them in the tutorial. It will be scheduled which time slot the practical exercises will be discussed.

(a) *(5 Pts.)* Generate three structurally different examples for retrieval tasks where ChatGPT fails but standard web search provides an answer and vice versa. Document those. Explain why this is the case for each example, i.e. why is the answer not satisfactory and which property of the retrieval question / model prohibits that a satisfactory answer is given.

(b) *(5 Pts.)* Demonstrate the different characteristics of the distribution of the page rank for the Erdösz/Renyi model compared to the Barabasi/Albert graph model in an example (e.g. generate random graph/s with similar number of nodes and edges for each model, compute page rank and compare characteristics of the resulting value distribution such as moments).

Possibly useful codesuites are: `https://graph-tool.skewed.de/`, `https://networkx.org/documentation/stable/index.html`, `https://neo4j.com/docs/graph-data-science/current/`