TDT4117 Information Retrieval - Autumn 2022
Assignment 5
The deadline for delivery is 17.11.2022

November 2022

Important notes

Please carefully read the following notes and consider them for the assignment delivery. Submissions not fulfilling these requirements will not be assessed and should be submitted again.

1. The assignment must be delivered as a PDF (Documentation/Report) and supplementary content in a ZIP file containing code and other complementary content. Other formats, such as .docx and .txt are not allowed.
2. The assignment must be **typed**. Handwritten assignments are not accepted.
3. Please be precise and to the point, and use figures/examples when it makes sense.
4. You may work in groups of a maximum of 2 students.

Task 1: Precision and recall

Imaging the two settings:

- You are in an eCommerce company, and you should promote a new product. However, before you start promoting, you want to retrieve previous promotion campaigns and information about state-of-the-art promotion strategies.
- You are a researcher, and you are developing a new method. Before you start entirely with the development and writing of a publication, you want to retrieve previous publications in the direction of your new method.

Given these two settings, answer the following questions:

- Describe what the information retrieval system should provide you.
- What operations, such as stemming, support the retrieval system in finding good matches for the case?

Task 2: Page rank and HITS

- Compare page rank and HITS and briefly describe the main ideas of both approaches and point out their differences.
- Given the graph below, compute hub and authority scores for webpages labelled as A, B, C, D and E using the HITS algorithm. Perform at least 3 iterations of the algorithm and illustrate your computations by providing formulas filled with values for at least one iteration.
- What are the scores if the self-connect in node B is ignored? Perform at least 3 HITS iterations of the algorithm and present the results.

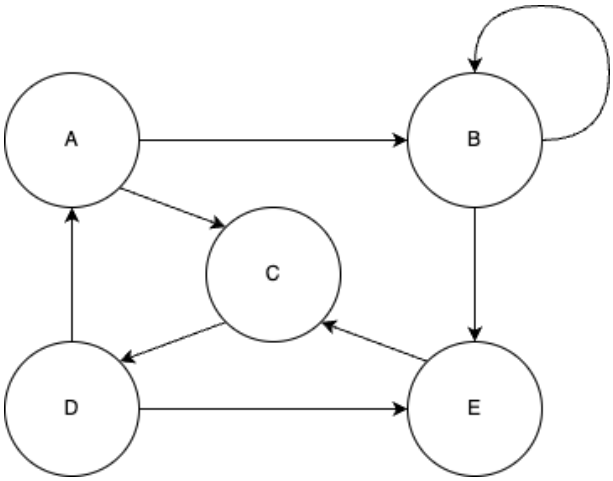

Figure 1: Graph of websites connected by links.

Task 3: Structured Indexing and Retrieval in Elasticsearch

For this task, reuse the ELK stack from assignment 4, including the Enron dataset reduced to 100,000 emails. Use the Elasticsearch API or Kibana for the discovery task (3b).

a) Index emails into more structured

- Check the structure and define what is interesting to be mapped to a specific field for the index.

- Note what you are going to map.
- Index the short list of emails (100,000) with ELK with your specified mapping after parsing each email.

b) Discovery

Use the discovery and dashboard option in Kibana.

- State some basic statistics:
  - Who sent most of the emails?
  - What is the most common subject?
  - How many emails are not from members of the *Enron* company?
- Further questions regarding "debra.perlingiere@enron.com":
  - What are the top 5 contacts Debra is mostly communicating with?
  - What are the top 5 contacts Debra is mostly communicating with, with no subject?
  - How many emails did Debra send with no subject?
- Further questions regarding Howard University:
  - What subject is mostly used in any email sent by Howard University and contains "University" in the body? (Hint date information should be ignored)
  - What do the emails contain to Howard University that contain "University" in the body?
- Explore the dataset and report on additional findings, such as "What institutions are part of the emails?".