Milestone 2 Mødes tirsdag 19. april kl 10 over Discord.

DESCRIPTION

After cleaning and processing our data in the first milestone, Milestone 2 will focus on how to efficiently represent the data in a database. Like last time, the milestone takes the form of a short jupyter notebook. It must be handed in on Friday, April 29, at 16:00 (in groups), and it is a requirement for attending the exam (it will be evaluated as passed/fail).

Martin

Task 1. Deadline 20. April MySQGL eller PGAdmin4 (en lille elefant)

The first task is to demonstrate that you have a working database containing the FakeNewsCorpus dataset. Explain your choice of schema design. You have been working on this task on a small subset of the data during the TA-sessions. For this milestone, demonstrate that your database contains a larger number of rows (e.g. one million - or however many you can reasonably work with on your available hardware), and that it supports simple queries.

Monika

Task 2. Deadline 25. april Har brug for Task 1.

List the relations you have created in your database. For each relation:

1. list its attributes
2. list its functional dependencies.
3. list all the primary keys.

Is each relation in BCNF form? If not, show how to transform the tables in BCNF and explain why it might be better (or not) to use the BCNF relations in your database.

Gabrielle

Task 3. Deadline 25. april Har brug for Task 1, men kan laves uden Task 2 (dog kan der være opstå problemer med BCNF).

Once your database is loaded, you can start issuing queries to better understand the characteristics of the data. Formulate the following queries in the database languages requested (in the square brackets following each item) and briefly discuss what you observe when you execute them over your database:

1. List the domains of news articles of reliable type and scraped at or after January 15, 2018. NOTE: Do not include duplicate domains in your answer. [Languages: relational algebra and SQL]
2. List the name(s) of the most prolific author(s) of news articles of fake type. An author is among the most prolific if it has authored as many or more fake news articles as any other author in the dataset. [Languages: extended relational algebra and SQL]
3. Count the pairs of article IDs that exhibit the exact same set of meta-keywords, but only return the pairs where the set of meta-keywords is not empty. [Language: SQL]

Jakob **Task 4.** Dealine 25. April Har brug for Task 1, men kan laves uden Task 3 og 4.

Now that we have our data in a database, let's revisit the "interesting observations" task from Milestone 1 - but now using queries to the database. The idea is to write database queries (e.g. using GROUP BY and COUNT) that explore features of the data set that are relevant to the fake news prediction task: outliers, artefacts. It's OK to investigate the same issues as in Milestone 1 (now using database queries) - but you are also very welcome to come up with completely new queries. You should write at least 3 such queries.

Task 5. Peerfeedback deadline 6. maj

Just like last time, after the hand in deadline, each group will be asked to evaluate the work of three other groups, based on a short list of criteria that you can find within the peergrade system. Again, this will only work well if everyone puts some effort into providing constructive comments, so please allocate some time to do this properly. It is an opportunity to get some feedback that can help you improve your final project. The deadline for giving feedback is a week after the hand-in deadline