Milestone 1

DESCRIPTION

Since the final project spans over quite a long period, we have split the work into three parts: two milestones and a final assignment. The milestones are intended just to check that everyone is on track, and to make sure that the workload for the course is spread evenly over the course.

In this first milestone, you will work on retrieving data, structuring data and cleaning data. You will document your work in a jupyter notebook, which must be handed in on Friday, March 4, at 16:00. The milestone will be assessed as pass/non-pass, and passing the milestone is a requirement for attending the exam. Note that you should hand in this milestone as a group.

Task 1

If you haven't already done so, the first task is to form a group of 2-4 people. There is an announcement on Absalon describing how to do this. Make sure that you list the names of all members of the group at the top of the jupyter notebook along with your group number.

Task 2

For our fake news predictor, we will be using the FakeNewsCorpus dataset as our primary dataset. It is available from this github repository: https://github.com/several27/FakeNewsCorpus, where you can also find information about how the data is collected, the available fields, etc. In this first milestone, we will work only on a small subset of the FakeNewsCorpus dataset. Your first task is to retrieve this subset from https://raw.githubusercontent.com/several27/FakeNewsCorpus/master/news_sample.csv and structure/process/clean it. Describe which procedures (and which libraries) you used and why they are appropriate.

Task 3

Now try to explore the FakeNewsCorpus dataset. Make at least three non-trivial observations/discoveries about the data. These observations could be related to outliers, artefacts, or even better: genuinely interesting patterns in the data that could potentially be used for fake-news detection. Examples of simple observations could be how many missing values there are in particular columns - or what the distribution over domains is. Be creative! :).

Task 4

In this task you should create your very own news data set by scraping it from the web. We will be looking at the "Politics and Conflict" section of the Wikinews site (https://en.wikinews.org/wiki/Category:Politics_and_conflicts), which contains news articles sorted by the first letter in their title. Since we want the different groups to have slightly different experiences with this data, each group should try to extract the articles for a specific range of letters - given by the python expression:

"ABCDEFGHIJKLMNOPRSTUVWZABCDEFGHIJKLMNOPRSTUVWZ"
[group_nr%23:group_nr%23+10]

where group_nr is your group number (according to Task 1). The data set you produce should contain fields corresponding to the content of the article, in addition to some metadata fields like the date when the article was written. Describe the tools you used, and any challenges that you faced, and report some basic statistics on the data (e.g. number of rows, fields, etc). Note that there are no fake/no-fake labels in this dataset - we will consider it as a trusted source of only true articles. Assess whether this is a reasonable choice.

Task 5 (handing in)
You will hand-in your jupyter notebook by submitting it to the peergrade system. Please make sure that you submit it as a group and specify all group members within the peergrade system.

Task 6 (post hand in)
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. 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.