Distant Reading (VT-23)
Course coordinator Matti La Mela, matti.lamela@abm.uu.se

Lab assignment week 4

Do the following exercises in a Jupyter Notebook file (see T0), submit the file in Studium in Studium by **Friday 10 February** at 23:59. For optional tasks, you submit also the requested output files.

Copy-paste the individual assignment "tasks" (T1, T2, T3, T4…) in Markdown cells. For programming questions, answer below in a code cell. In the code cell, use comments with hashtag where needed. Tasks that are "open questions" without coding are answered by typing your answers (you can add images).

Double-check that your python code works. Feel free to reuse as much code from the learning materials as you like, use the tutorial videos, search solutions on the internet and look for help in the reference readings. If you remain stuck or have questions, you can use our slack workspace to ask more help.

NB. Ts marked with optional are required for reaching grade VG, but are not compulsory.

Good luck, this is the last Lab assignment!

Assignments:

T0) Create a new Notebook file in Jupyter Lab. Name the file yourlastname_lab4.ipynb (e.g. La_Mela_lab4.ipynb). Copy-paste the individual assignment “questions” (Ts) in Markdown cells and your answer in a Markdown / code cell below this depending on the question.

In the first part, we use a topic model generated with Mallet on our DHQ corpus. You can find the output files used in the labs (dhq_keys.txt, dhq_composition.txt) in the folder “dhq-tm”. In the second part, which is optional, you use Mallet to build an own topic model; guidance for installation can be found in Material_4a.

1. Examining the topic model

T1) Open the file “dhq_keys.txt” either in Jupyter Lab (following Material_4a) or in Excel (or other spreadsheet program, even notepad++ is ok). In Excel, it is good to use the import data/txt tool, (“Data -> From Text/CSV”) and not to open the file directly. Browse the file as requested in T2-T4.

T2) How many topics are there in this model? What are the top words in topic number 23 (our first topic is number 0)

T3) Go through the top words in all the topics. Discuss briefly the results, do they seem valid topics?

T4) Pick five topics and give “labels” to them, e.g. describe them briefly according to what you think these topics represent. **Type your answer:** Topic X -> “my description”

T5) Choose one topic that you find interesting and “label” it (give a brief description of the topic). Open then the “dhq_composition.txt” either in Jupyter Lab (see Material 4a) or in Excel (Use “Data -> From Text/CSV”). In the dhq_composition.txt, find (by browsing the file) **three articles** that have a high representation of your chosen topic. **Type in your answer:** the number of the topic you have chosen, and then the names of the three articles and the share of the topic in the article (the share is found in “dhq_composition”).

T6) Read the abstracts of these three articles (either online, or opening the txt files in the corpus). Discuss briefly: do they seem to correspond with the topic that you have chosen? Based on this, is the “label”/name/description you have chosen for your topic suitable?

2. Topic modelling with MALLET (optional)

This second part is **optional** but recommended to everybody (at least to read and think through); it will not only teach the use of Mallet, but also instructs you again to navigating in the terminal / command line. See instructions in Material_4a.

When you submit your assignment, **remember to include the “keys” and “composition” output files** of your topics model to the submission (either zip, or attach several files).

(optional) **T7)** Copy the dhq_corpus_complete_2007_2020 to your Mallet directory, eg. to: …\mallet.2.0.8\dhq_corpus\ .

(optional) **T8)** Import the corpus directory into Mallet by using the import-dir command.

(optional) **T9)** Train a topic model on the dhq_corpus (use: bin\mallet train-topics …). **Choose 60 topics**. How much time did it take?

(optional) **T10)** Discuss and compare your results with the topic model used in part 1 of this assignment. What are the main differences? What new topics do you find? In which cases is the 60-topic model more usable?

(optional) **T11)** Select one topic that you find interesting. Based on the "composition" output, find three articles that include a high share of that topic. **Type here:** the number of your topic, the top keys (top words) of the topic, and the titles of the three articles. How could you label the topic you have chosen?

(optional) **T12)** Describe briefly a workflow (explain, no code needed), where you would generate a topic model on DHQ articles that are lemmatized and filtered only for nouns (thus explain the steps from the current DHQ corpus to the end product). In contrast to T9, how do you think the results would change when using this lemma-nouns DHQ corpus?

Submit your Notebook file and include the "keys" and "composition" output from the topic models created in T9.