Spring 2024

Course Assignments for Advanced Visual Data Analysis (TNM098) Spring 2024

Lab Assignment 3

Content based analysis of data

Task 1 *Image comparison*

Take the zipped data file and extract the 12 images it contains. Create a visual data analysis solution that loads each image and analyses it to produce a feature vector. Suitable features might include:

1. Colour content
2. Colour distribution around the central point
3. Colour distribution around several points
4. Luminance distributions around one or several points
5. Edge positions and orientations
6. Anything else that you can think of

Use a distance measure to compute the distances of the feature vectors to each other and produce a 12x12 distance matrix (Refer to lecture 4 on computational support). Pick one image from the set and use the distance matrix to rank the other 11 in order of similarity to your chosen image.
Suitable image loader/analysis libraries can easily be found on the web. Examples of possible libraries of interest for python: opencv (ported also to JavaScript), scikit-image, scipy, pillow/PIL, NymPy.

Questions of interest: How did you make the feature vector comparison? Did you apply any weighting of features over one another? Why?

Task 2 *Text analysis*

This tasks build on data from the VAST challenge 2011, mini challenge 3 (MC3): "Investigation into Terrorist Activity".

Description of MC3: *Intelligence analysts are looking for information related to potential terrorist activity in the Vastopolis area. News reports have been provided and it is up to you to identify any potential threats and give as much detail as possible on them.*

A dataset of news reports in provided in .csv format (`TNM098-MC3-2011.csv`), where each row corresponds to a news report.
The file includes following attributes/columns per news report: ID (unique numeric id), Title (headline of news article), Date (of publication), Content (textual content of the news article).
The task is to: Create a visual data analysis solution to extract relevant information regarding the potential threat(s) to Vastopolis.

Use the approach of preference build up your solution: as an interactive visual interface, as a python notebook, using high-level machine learning platforms (knime, orange etc.). Refer to the lecture notes and the suggested list on the ivis group page for suggestions: `https://ivis.itn.liu.se/courses/resources/tools.html`
Some suggested steps for relevant analysis are below and you are free to explore additional directions/aspects of interest. Please note that, different steps can be appropriate depending on the approach and library/platform used. Refer also to the lectures of text visualization and analysis (LE 5 & 6).

1. Import and pre-process/prepare the text for upcoming analysis; remove punctuation, make everything lowercase, remove stop words, tokenize.

2. Explore the temporal distribution of you temporal data. Use the dates to get an overview of the reporting density over time.

3. Reduce your dataset to a relevant subset. Eg. filter the news reports based on a relevant set of keywords according to the challenge description (eg. threat, terrorism, dead, attack, explosion ...)

4. Explore the temporal distribution of you temporal data to get an overview of the relevant reporting density over time after filtering. (An interactive exploration alternative here could be to create a daily histogram and query the text contents of selected days; especially the peaks for example).

5. Visually compare unfiltered vs filtered temporal distributions.

6. Apply topic modelling to identify relevant topics reflecting the potentially ongoing situation.

   - Create a document-term and/or tf-idf (term frequency-inverse document frequency) matrix.
   - Build a topic model, eg. LDA
   - Display topic related words (eg. in a list, word cloud, or use pyLDAvis)

7. Look at the temporal distributions of the reports (documents) in selected retrieved topics. Relate them to the temporal distributions of unfiltered and filtered(relevant) sets of news reports (see above). This could give indications on whether any of the topics become prevalent around/after specific event timestamps.

8. Topic modelling parameters may be iteratively refined based on the (initial) insights made through the analysis process. Additional keywords maybe added to adjust the relevant filtered dataset accordingly.

9. Potentially display/visit individual identified relevant reports for getting more context. For example extract and explore more focused/narrow subsets relating to a detected topic by filtering based on particular/fewer keywords from that topic.

Question of interest: Is it typically the top (first identified) topic that is most informative with respect to the subject of analysis, or one of the follow-up topics?

Prepare a short report describing your analysis approach and your proposed solution. Include a description of your solution, a discussion of your results and their quality, of the questions of interest and of any additional insights your solution makes possible.

Present your solution to a teaching assistant in one of the scheduled lab session for Lab 3: **Monday, 22 April, 13-17** (reserve lab: Monday 29 April, 13-15).
Upload your short report (~2 pages, PDF format) under submissions Lab 3 in Lisam.
