7CCSMDM1 (Data Mining)

Coursework 2

due 24th April 2020 @ 4pm GMT

(version 1.1 – corrected typos, indicated in red)

Overview

The goal of this coursework is to learn how to work with two types of data sets: a **text corpus** and an **image library**. The coursework is divided into 2 major sections: one for working with text using **Natural Language Processing (NLP)** and one for working with images using **Image Processing**.

Marks for each are distributed as indicated below, for a total of 100 marks for the coursework. The coursework is worth 10% of your overall mark for the module.

You must write some Python code to process the data as directed below and then report and explain your results in a PDF document. You will need to submit three things:

- a PDF file that contains your answers and explanations for how you obtained each answer;
- SOURCE CODE files in Python (***.py** files) to justify your solutions; and
- a README file that explains which source code files were used to generate answers to which questions.

Be sure to follow the **submission instructions** on the last page. Failure to follow these instructions can result in a penalty of up to 25% of your mark for this assignment.

The work you submit must be strictly your own! It is fine to discuss your *solution strategies* with your classmates, but the code you write and the report you submit must be yours and yours alone. Failure to submit individual work may result, in the best case, in a hearing with the Misconduct Committee and 0 for the whole assignment; and in the worst case, in expulsion from the College.

Obtain your data

Start by downloading the data from the KEATS page for this module: **cw2-data.zip**.

Unzip the archive, and you will find two directories in this data set: **txt** and **img**.

In the **txt** directory, you will find 10 text files. These are plain text files of books downloaded from the Gutenberg Project web site (`http://www.gutenberg.org/wiki/Main_Page`). There 5 books by Beatrix Potter: *The Tale of Benjamin Bunny*, *The Tale of Ginger and Pickles*, *The Tale of Jeremy Fisher*, *The Tale of Peter Rabbit* and *The Tale of Squirrel Nutkin*. And there are 5 books by Rudyard Kipling: *The Jungle Book*, *Just So Stories*, *Kim*, *The Man Who Would Be King* and *Puck of Pook's Hill*. Note that the data is not cleaned, but may contain tags, such as **[Illustration]**. These should be ignored if you use stopwords correctly (*Hint:* Edit the stopwords file to include the tags.). Use this data to answer the questions in Part 1.

In the **img** directory, you will find 6 image files. These are images taken of plastic plants (from Ikea) for a plant identification exercise. Use this data to answer the questions in Part 2.

1 WORKING WITH A TEXT CORPUS – NLP

(50 marks total)

There are 10 files in this corpus.

HINTS:

- Refer back to the NLP Practical (week 7) for tools to handle text data.
- Don't forget to remove **stop words** before performing any text analysis, as you did in the NLP Practical.

Your mission is to analyse the data using **scikit-learn**, **nltk** and **TextBlob**. First, compute the following statistics for each book:

(a) **polarity**

(b) **subjectivity**

(c) **word count** (without stop words) in the book

(d) **most frequent term (word)** in the book

(e) **normalised frequency of most frequent word** (normalised by the word count)

(f) **term frequency** of the most frequent word in the book

(g) **inverse document frequency** of the most frequent word in the book

(h) **TF-IDF** for the most frequent word in the book

In your report, include a table in which you report all these statistics, clearly indicating which statistics go with which book.
Also indicate which TF-IDF formula you use (see Lecture notes from week 7 about TF-IDF).
(40 marks for computing the above statistics for all 10 books, 5 marks per statistic)

Next, mine the data for patterns. The books can be labelled by author (Potter and Kipling). Can you detect any patterns in any of the statistics between the two authors? Can you use the patterns to train a classifier that will accurately determine which author wrote a given book? **In your report, clearly describe the pattern(s) you have identified, which technique(s) you applied to find the pattern(s), how you trained a classifier, which classifier you used and your classification results.**
(10 marks for explanation)

2 WORKING WITH AN IMAGE DATA SET – IMAGE PROCESSING

(50 marks total)

The 5 raw images which you downloaded are shown in Figure 1. Your mission is to analyse the image data using **scikit-image**, as follows:

(a) *(5 marks)* Generate a **greyscale** version of each image.

(b) *(5 marks)* Generate a **black-and-white** version of each image.

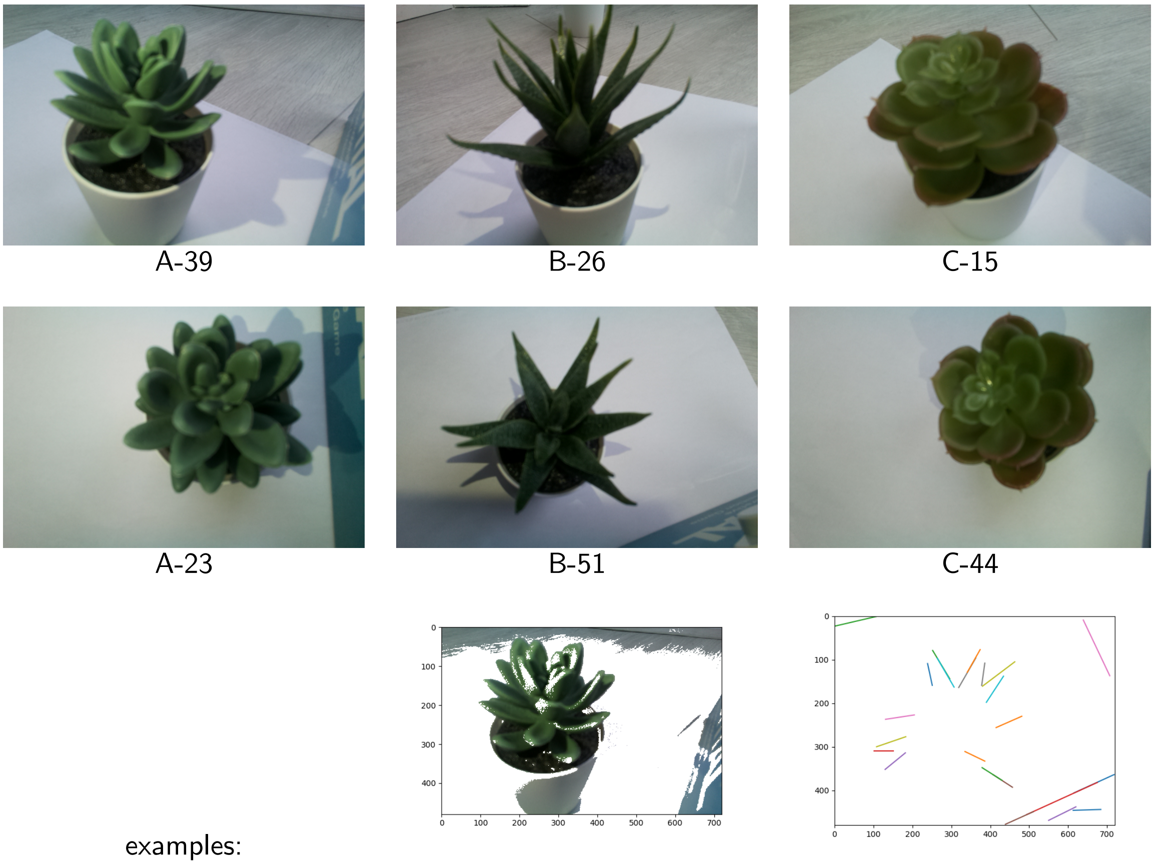
(c) *(5 marks)* Detect **edges** in each image.

(d) *(5 marks)* Detect **contours** in each image.

(e) *(10 marks)* Detect the **green** in each image

(f) *(10 marks)* Detect **straight lines** in each image using the Hough transform.

In your report, explain your analysis and how you reached your answer to the question above. For each of the 6 images, describe clearly which technique(s) you applied and show the results for each type of analysis (a-f). *(10 marks for explanation)*



A-39 B-26 C-15

A-23 B-51 C-44

examples: (2e) greens (A-39) (2f) straight lines (B-51)

Figure 1: Raw images (top and middle rows) and example answers (bottom row)

SUBMISSION INSTRUCTIONS

Generate ONE ZIP archive that contains your submission, which must consist of:

- A REPORT, **in PDF format**, with the results, tables, plots and explanations, as requested to answers the questions posed in each part (1 & 2) of the assignment
- All the Python (***.py**) files you used to generate your answers
- A README (plain text) file that explains which source code files were used to generate answers to which questions

NOTE: Your Python code MUST be saved as ***.py** files, not iPython Notebooks. **We will NOT accept notebooks. No *.ipynb files allowed!**

You could lose up to 25% off your mark if these submission instructions are not followed.

SUBMIT YOUR ZIP SUBMISSION ON THE 7CCSMDM1 KEATS PAGE, where it says "Submission Links".

BE SURE TO SUBMIT BEFORE THE DEADLINE!!!
(24th April 2020 at 4pm GMT)