Multimedia Systems and Applications 4 Coursework
Building a Photo Tag recommendation system on a Flickr Collection

Submission Deadline: 21st November Monday at 430PM   **Total marks:** 20
Please read **all** of this document before starting the task. This is an individual task.

1. Introduction
In this coursework you will be developing a photo tag recommendation system (much like those used on websites such as Flickr, Instagram) in order to suggest tags to the user based on some existing tag set.

On image sharing websites, users annotate their photographs with tags. These single word phrases are useful for organising the data set and also for searching purposes. However, users often do not give enough tags, or even appropriate tags. This is why websites such as Flickr recommends new tags to users for their images. Your system should fulfil the following:

> *Given a tag X (e.g. football), suggest related tags (e.g. ball, pitch, grass etc) based on those which exist frequently in images previously annotated with X on a Flickr dataset.*

2. Dataset
You will be given 10,000 images and tags they have been annotated with by their user on Flickr.

You will use *term co-occurrence* as a feature. Co-occurrence matrix captures term relationships and hence harnesses the semantic information of keywords co-occurring in an image collection. For example, you would expect the keyword `ice` to exist in a large number of images with `snow`. Conversely, you would expect a low co-occurrence score for `snow` and `elephant`.

Co-occurrence information can be used to recommend tags. Given the image collection, you should first build a matrix which stores the number of images that contain two given tags. For example see Figure 1:
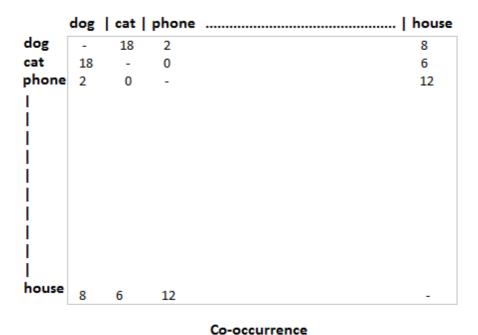

Figure 1 An example tag co-occurence matrix for a collection containing 100 images

In this example, `dog` exists in 18 images where the keyword `cat` does. `Phone` occurs in no images where `cat` exists.

3. Experiments

You may complete the coursework using **any** programming language.

Task 1 – Building a tag co-occurrence matrix

First you must build a tag co-occurrence matrix for the tags present in the 10,000 supplied images such as the one shown in Figure 1. Based on this, you must submit the following along with your report

1. The code used to build the co-occurrence matrix
2. The co-occurrence matrix in a comma separated file format

Task 2 – Recommending the most popular tags

Secondly, you must compute the top 5 tags which co-occur with the following tags:

1. Water
2. People
3. London

To do this, you should select the relevant row in your matrix, and rank the columns in descending order. Make sure to first set the co-occurrence value of "X with X" to be 0 e.g. otherwise we will recommend water for the input tag water.

Task 3 – Recommending tags based on their popularity and significance

You may notice that the recommendations offered in the top ranks are often unrelated popular tags. This is because popular tags, by their nature, co-exist highly with almost all tags in a collection. Therefore, to overcome this we must introduce a weighting factor called inverse document frequency (IDF).

This value is computed for every tag in a collection and attempts to capture its "significance", or popularity. The IDF score for tag X is computed as:

$$IDF = \log(I/I(X))$$

where I is the number of images in the collection, and I(X) is the number of images tagged with tag X. Therefore, popular tags are discounted.

In order to overcome the "popular tag recommendation" problem, when recommending tags you must first multiply each recommended tag score by its IDF value. Given this, now compute the top 5 tags for water, people and London again.