DATA20021 Information Retrieval (Spring 2022)
Exercise Set 2 (Submit solutions via Moodle by 4 February.)

1. Encode the numbers 100, 200, and 400 as:
   (a) Elias-γ codes;
   (b) VByte codes.

2. Are the following statements true or false?
   (a) In a Boolean retrieval system, stemming never lowers precision.
   (b) In a Boolean retrieval system, stemming never lowers recall.
   (c) Stemming increases the size of the vocabulary.
   (d) Stemming should be invoked at indexing time, but not while processing a query.

3. The following pairs of words are stemmed to the same form by the Porter stemmer. Which pairs would you argue should not be made the same. Give your reasoning.
   (a) abandon/abandonment
   (b) absorbency/absorbent
   (c) marketing/markets
   (d) university/universe
   (e) volume/volumes

4. In the context of a web crawler, what is the URL-seen test? Describe an efficient data structure for performing the URL-seen test and describe how it is used.

5. Why is it better to partition hosts (rather than individual URLs) between the nodes of a distributed crawl system?

6. Consider the following collection of documents:
   Page 1: The smallest planet is Mercury.
   Page 2: Jupiter, the largest planet!
   Page 3: Neptune is smaller than Uranus, but is heavier.
   Page 4: Uranus is smaller than Saturn.
   Page 5: Neptune is heavier than Uranus.
   Page 6: Jupiter $>$ Saturn $>$ Uranus $>$ Neptune
   (a) Draw the inverted index for the collection after normalizing by applying case folding and removing punctuation (normalization you like, e.g., stemming). Include document frequencies (the number of times each normalized term occurs in the collection); term frequencies (the number of times each document contains the term); and also term positions. See slides from the third lecture for a guide (you can ignore the "Weights" part shown there).

7. Give an overview of the Elias-Fano encoding scheme and illustrate it for the example set $\{4, 8, 10, 12, 33\}$. Describe how the ith integer can be extracted from the encoding.