NLP Coursework (1) [**Reassessment**]

Dell Zhang
Birkbeck, University of London

2021/22

Part 1 of the NLP coursework is worth 10 marks.

1. **(2 marks)**

Reconstruct the document collection from the following positional inverted index.

```
cold, 2:       [ 1,1:<6>; 4,1:<4> ]
days, 1:       [ 3,1:<2> ]
eat, 1:        [ 6,1:<1> ]
hot, 2:        [ 1,1:<3>; 4,1:<8> ]
in, 3:         [ 2,1:<3>; 4,2:<1,5> ]
lot, 1:        [ 6,1:<3> ]
nine, 1:       [ 3,1:<1> ]
old, 1:        [ 3,1:<3> ]
peas, 5:       [ 1,2:<1,4>; 2,1:<1>; 5,2:<1,3> ]
porridge, 5:   [ 1,2:<2,5>; 2,1:<2>; 5,2:<2,4> ]
pot, 3:        [ 2,1:<5>; 4,2:<3,7> ]
the, 4:        [ 2,1:<4>; 4,2:<2,6>; 6,1:<2> ]
```

2. **(2 marks)**

Write regular expressions for the following language.
By "word", we mean an alphabetic string separated from other words by white-space, any relevant punctuation, line breaks, and so forth.

(a) The set of all strings with two consecutive repeated words (e.g., "to to" and "be be" but not "to be" or "be to be").

(b) The set of all strings that start at the beginning of the line with an integer and end at the end of the line with a word.

3. **(2 marks)**

Compute the Levenshtein edit distance between two words '`drive`' and '`divers`' (with insertion cost 1, deletion cost 1, substitution cost 1).
Show your work using the edit distance grid, and also represent the final result as an alignment between those two words indicating the editing operations required to convert the former to the latter.

4. **(2 marks)**

What is the γ-code that encodes the following postings list?
`16, 22, 25, 26`.

5. **(2 marks)**

Compute the cosine similarity between each pair of the following document vectors, using just the raw numbers in those vectors without any TF-IDF weighting.

$$d_1 = \begin{pmatrix} 6 \\ 6 \\ 0 \\ 0 \\ 3 \end{pmatrix} \quad d_2 = \begin{pmatrix} 1 \\ 1 \\ 2 \\ 3 \\ 1 \end{pmatrix} \quad d_3 = \begin{pmatrix} 0 \\ 2 \\ 1 \\ 0 \\ 2 \end{pmatrix}$$