NLP Coursework (1)

Dell Zhang
Birkbeck, University of London
2021/22

Part 1 of the NLP coursework is worth 10 marks.

1. (2 marks)

Build the positional inverted index for the following document collection, and show it in the format used by the IIR textbook.

- $d_1$: learning to learn using gradient descent
- $d_2$: meta learning in reinforcement learning
- $d_3$: learning to learn by gradient descent by gradient descent
- $d_4$: learning to learn without gradient descent by gradient descent
- $d_5$: meta gradient reinforcement learning

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 that represent prices in the UK, i.e., the pound sign followed by a real number which has optionally two digits after the decimal point. For example, £1500 and £19.99.

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

3. (2 marks)

Compute the Levenshtein edit distance between two words ‘Sunday’ and ‘Saturday’ (with insertion cost 2, deletion cost 2, substitution cost 3). 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. \( (2 \text { marks}) \)

What is the \( \gamma \)-code that encodes the following postings list?
12, 18, 21, 22.

5. \( (2 \text { marks}) \)

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.

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