The assignment has 4 Questions, enumerated using Roman Numerals (I - IV).
Questions II and III does not involve any coding. You have to answer the quizzes uploaded in Moodle.
Question I has 2 sub parts of which I.1 has a machine (read Deep) learning component in it.
Question IV also includes a learning component.

I. The question mainly focuses on two simple nlp tasks where we are supposed to deal with word vectors. (40 Marks)

1. **Analogy Task :** The analogy prediction task is defined as follows. Given a pair of words a, b you need to find out the pair of words among five given pair of words, which is more appropriate as per as analogy is concerned . Learn a deep learning model for the task. Report the accuracy of the model after performing 5-fold cross validation.
    a. For training you may use the new training files in **wordRep.zip**. In the files, all the pairs belonging to same category is given in a single file. Use different combinations from the corresponding files to generate positive and negative examples required for training.
    b. Please note that, you need not change the word vectors, but you are supposed to learn a new model (or a function) that performs the analogy task.
    c. You may use the **Word-analogy-dataset** only for validating your model

    e.g. - ‘`sandal:footwear`’ is analogically appropriate to ‘`watch:timepiece`’, `compare to other pairs like ‘monarch:castle’, ‘child:parent’, ‘volume:bookcase’, ‘wax:candle’.`

2. **Similarity Task:** For a given input word you need to find out the most similar word among the 4 options given
    a. The task involves no learning.
    b. You are supposed to use the following metrics
        i. Cosine similarity
        ii. Euclidean distance
        iii. Manhattan distance
    c. Amongst the 40 entries given in the word similarity dataset, Report the number of entries which gave the highest score to the correct answer
    d. Also, report the MRR for each of the distance measure

    e.g. - ‘approve’ is more similar to the word ‘support’ compare to ‘boast’ , ‘scorn’, ‘anger’

Resources:
1> **glove.6B.300d.txt.gz** - contains billion of words with corresponding 300 dimensional vector.
2> **Word-analogy-dataset** - contains 100 questions with answers to validate.
3> **Word-analogy-dataset-format -** contains the format of the previous file

4> **Word-similarity-dataset** - contains 40 questions with answers to validate.
5> **Word-similarity-dataset-format -** contains the format of the previous file
6> **wordRep.zip -** Training instances for the analogy task
NOTE: If you don't get the vector of any word (from the two datasets) in the **glove.6B.300d.txt.gz** file, ignore the question.

II) Quiz 1 - Source and Derived Words (updated in moodle) (20 Marks)

III) Quiz 2 - Word Pairs (20 Marks)

IV) **Derivational word vector generation (20 Marks)** - A new word in a language can be formed from an existing word and an affix (generally suffixes). Such words are called derivational words. For example *Indian* is derived from *India*, *industrialist* is derived *industry* etc.

You have to learn a model that generates vectors for the derived words, when given the vector for source word and the target affix. You can learn a separate model for each affix or you can learn a single model for all the affixes. The derived word vectors are also provided in the dataset for training and validation. Report the accuracy of the model after performing 5-fold cross validation.
Resources:
Vector_lazaridou.txt - Word vectors for source and derived words as per the distributional space described in "Compositional-ly Derived Representations of Morphologically Complex Words in Distributional Semantics"

fastText_vectors.txt - Word vectors for source and derived words a per the fastText model

wordList.csv - CSV files containing the triplets **Source word, derived word and the affix**