**Report**

Using the programme on the text.txt questions, originally, the percentage of questions on which *most\_similar\_word()* guessed the answer correctly was 62.5%, building a semantic descriptor from Swann’s Way and War and Peace. However, we discovered that the *get\_sentence\_lists()* function was not separating words that ran over the line. Thus, the last word of a line and the first word of the next line were not recognised as two independent words, limiting the size of the dictionary and, therefore, words available in *semantic\_descriptors*. The function was resolved to account for this using **\n**.

Also, the 62.5% result included words from test.txt that were not present in *semantic\_descriptors* (i.e. in the novels), for which the semantic similarity could not be computed (we also modified *most\_similar\_word()* to print ‘ERROR’ should this occur). Therefore, to make the percent more reflective of the accuracy of a programme, working ideally, we also tested the code on a modified version of test.txt that had these words removed, giving rise to 75%. The accuracy was also limited by titles and the recurrence of the word ‘chapter’ (especially in War and Peace), which would also be included in sentences and thus the semantic descriptors. So, to further improve, novels were simplified to remove headings and, as much as possible, only include novel content and then the programme was rerun but the processing time proved to be very long.