Word comparator Project

Manuela Giansante

This project has been elaborated by me, but my thanks go to Lucia Camenisch who has helped me gain new perspective on some of the issues I encountered.

Files handed it:

a. project.py
b. main_function.py
c. output word_list,pdf
d. output word_list_tt.pdf

Contents of project.py

The file contains the 3 main functions built to be utilised in the main_function.py file.

i) read_reference_text

This function opens the T file, that must be saved in the same environment as project.py. It processes the file, reading it line by line, separating the words by punctuation and blank spaces.

Then, it lowers the case of all the words, through the method map. It stores all the words/strings that make up a sentence in the original file into a list. Finally, all these lists are collected in a list.

ii) make_word_vector

The function takes the arguments w and txt, so a single word and the list of lists we generated from before.

It builds a semantic vector for the word we are considering, this vector is a dictionary that has as keys the words in the sentence with w (even w itself) but not any of the words in the set S or words with length smaller than 3.

The values connected to the keys are the frequencies by which they appear in the same sentence as w.

iii) sim_word_vec

It takes as arguments two semantic vectors, built with the precedent function and calculates the cosine similarity.

$$cosine\ similarity(v_1, v_2) = \frac{v_1 * v_2}{\sqrt{(v_1 * v_1) * (v_2 * v_2)}}$$

The scalar products are computed with the dictionaries values.

Contents of main_function.py

i) word_comparator

The functions in project.py are re-called and used inside the word comparator function. The arguments are the word list of words we want to compare and the text file we want to compare them by.

So the function reads and processes the file, it builds the vectors for every word in the list. Finally, it prints the words duos who are the most similar, given the context of the text file.

Contents of output word_list and output word_list

PDF prints of the output for the two lists of words we were provided with, the latter is the sample list we test the function on.