W2V model training. Hyper parameter tuning. RV Coefficient.

The goal of the task is to search for and find a suitable, already pre-trained W2V Gensim model and use the gensim NLP library to further finetune/reinforce it for a specific domain by training it on a set of specific documents that Iris.ai will provide.

The goal of the software is to be able to hyperparameter tune the enriched model and evaluate it against a metric defined below.

After you choose the basys pretrained gensim model to enrich, you can use the dataset provided by Iris.ai for the enrichment training or any other dataset you find useful. Building your own dataset is an option too.

Dataset: the dataset is in a json file containing the titles and descriptions of around 24 K documents. The task includes also preprocessing the dataset data so that it can be used for the enrichment training in gensim. Preprocessing should be prepared as a module, so that it can be used on other datasets too.

For the hyperparameter tuning of the enriched model you should use the **Optunity** framework. You should set it up in a way so that a configuration json file can be provided with the desired criteria in the form of range values for each parameter.

The performance metric is **RV Coefficient** which represents the similarity between a set of learned vectors (embeddings) for a control set of words. The control set of words will be provided to you, along with the learned vectors from an already fine tuned model. A code snippet for computing the RV coefficient will also be available to you.
During the hyper parameter tuning of the model, the goal is to try to reach **maximum RV coefficient value.**

You may use internet materials, data or anything else that you find suitable for solving the task. There are no restrictions whatsoever.

The deliverable of the task would be a git repository with all the necessary code & artefacts inside. We would also discuss the solution after it is ready.

Do not hesitate to write if there are any questions from your side. The time for completion of the task is 3 full days.

Good luck!