

Take home task - Working Student (Data Science)

Given:

1. products_titles.csv - Contains products IDs and their corresponding titles.
2. products_technical_specs.csv - Contains products IDs and their corresponding technical specifications.

Task:

1. Extract some insights from the products' technical specifications. If necessary, make sure to apply sanity checks and/or pre-processing operations.
2. For every possible combination (pair) of products, determine the similarity between them based on product titles and technical specifications individually.
(Use **ONE** of the following text vectorization approaches: Term frequency, Tf-idf, One-hot encoding)

Expected output format:

Product No A	Product No B	Similarity based on title	Similarity based on tech specs
1001	1002	<some-similarity-value>	<some-other-similarity-value>
1001	1003	*	*
*	*	*	*
*	*	*	*
*	*	*	*
1005	1004	*	*

3. Create a visualization to identify at which rows from the above table the difference between the two similarity results (titles and technical specifications) fall under the ranges [0-0.25], [0.26-0.50], [.51-0.75] and [0.76-1.0]

4. Provide a self inference (in words) based on the similarity results and comment on it.

Optional:


- For task 1, implement a function for computing one of the text vectorization approaches instead of using an Open Source library. You should only use one or more of the following libraries: Pandas, Numpy
- For task 2, in addition to the vectorization methods use one pre-trained word embedding model of your choice and comment on it.

We are looking forward to getting your solution to the task in the form of your choice (Github repository / Jupyter notebook, ...) and discussing it. Good luck.

Best regards,
Conrad Data Science Team

Links to the Data Files:

 [products_technical_specs](#)

 [products_titles](#)