CODING ASSIGNMENT
(Senior) Data Scientist

Thank you for showing your interest in being a part of Talentwunder's Data Science Team. We are eager to learn about your skills and ideas that you want to contribute to Talentwunder!

In this assignment you will get the chance to showcase your technical skills, creativity as well as business sense by solving our tasks. In addition, your answers will serve as a starting point for future discussions of the interview process. We are looking forward to see your results!

Tasks

1 Serving Word Similarities

Similarity that is based on the distance between embedding vectors is a useful tool for content-based recommendation systems. For this task, you will train a model to learn word embedding vectors from textual data and deploy the model as a simple service to serve the top 10 most similar terms for a given query word.

Requirements:

- Implementation should be done in **Python 3.7+**.
- The training and service should be deployed as a docker-container (e.g. docker-compose).
- We expect to run your solutions *out-of-the-box* with a single command like „*docker-compose up*"

1.1 Model Training

Use the dataset "*data/sample_data.csv*" for training your model. Deploy the training script as a docker-container (e.g. docker-compose). Note that the quality of the word embeddings is not of importance!

1.2 Model Serving

Implement an API endpoint "/mostSimilar" that returns the 10 most similar words according to your learned word embeddings. The marketing team has provided two lists that your service needs to account for:

- "currated.json": A dictionary containing words that you should serve instead of the learned representations.
- "blocked.json": A list of query words that we don't want to serve similar words for.

These files are shared via an s3 bucket that is provided by another docker-container. (Please start the container using "*docker-compose -f docker-compose-stack.yml up*" and remember to set an appropriate endpoint-url, such as "*http://localstack:4566*" if you have joined the localstack network.)

For bonus points: Implement an endpoint to dynamically reload these files from s3.

2 Similarity Search

Representation vectors are very powerful to capture complex search criteria. However, for a large vocabulary of representation vectors (10+ million, with embedding size N=100+), a straightforward query, i.e. computing the distance between the query vector and each entry of the vocabulary, is often too expensive.

Consider Talentwunder would like to provide a search based on representation vectors to find similar profiles. A straightforward query is not feasible. **How could you help?**

Note: No coding or implementation is required. Keep your answer “high-level”.

3 Entrepreneurial thinking

In the first interview session, we introduced you to some features and data products developed at Talentwunder.

- Can you think of a data driven product/feature that would be useful in the recruiting domain?
- What data would you need to realize the project? Where do you get the data from?
- How would you measure the success of your data product?
- What kind of model would you use to implement the solution and why?

Note: Try to exclude products/features introduced at your first interview; However it is not required to come up with your own product.

FAQ

How much time do I have?

Please send us your results in 5 days from the time you get this assignment.

How much time should I spend on this?

This assignment should take no more than 8 hours in total. Time is hard to find though.

Is code quality of importance?

Yes, code quality will be a major factor. This includes documentation, error handling, unit-testing, use of OOP or other code patterns/strategies, and more. Focus on these aspects as you see fit with respect to the time constraints.

Any questions or clarification needed?

If you have any other questions, hit reply to the email this assignment was attached to or send your questions directly to: **sascha.gerloff@talentwunder.com**

Überschrift GROß

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum

Überschrift KLEIN

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum

- Talentwunder
- Talentwunder

a) Talentwunder
b) Talentwunder

Text 1	Text 2	Text 3	Text 4
Kannste was reinschreiben	Kannste was reinschreiben	Kannste was reinschreiben	Kannste was reinschreiben