In the current stage of interview process you are given a programming/analytics task.

You are expected to implement your solutions using Python programming language on Jupyter Notebook. The description of the task is given below:

You are given two datasets:

* articles.csv
* interactions.csv

The articles dataset contains the details of a number of articles. The columns are:

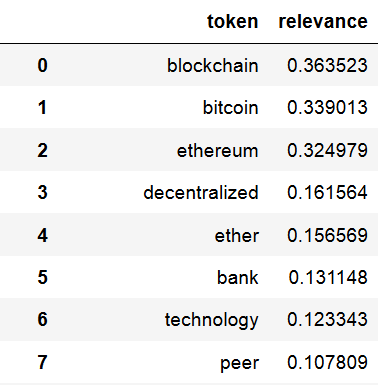
* **contentid**: unique id of an article
* **url**: web url of an article
* **title**: title of the article
* **text**: content of the article
* **lang**: language of the article

The interactions dataset contains the interactions between different people and articles. For example, some people just view an article, some other follow and bookmark other articles etc. The column names are:

* **personid:** Unique id of a person
* **contentid:** unique id of an article (can be joined with articles dataset)
* **eventType:** type of the interaction ( view, like, bookmark, follow, comment created )

**The task is:**

* Construct a list of 5000 tokens ( a token may be composed of a word or a group of words ) which will characterize the content of the articles in the best possible way.
* Represent each person in terms of that 5000-token vector. In other words, you will represent each person as a [1, 5000] vector. The values of this vector will hold the relevance score of the related token to that person. Just to visualize, for each person you are expected to come up with a vector like the following:

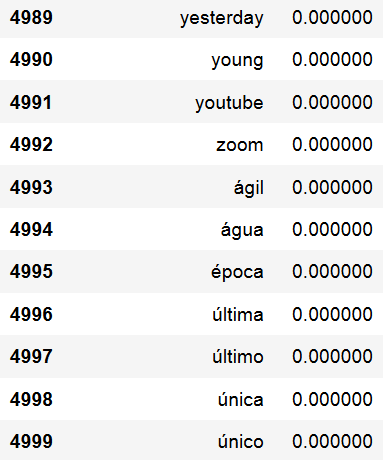


.

.

.

.



Good luck and thank you for your time.