BEHAVOX

DATA ANALYST: TEST TASK

February 2020

Test task overview and deadlines

Congratulations on advancing to the next stage in our interview process.

The test task presented below is aimed at assessing your ability to design and implement features calculations across a dataset.

The test project details are outlined below. There is no predefined due date for the submission of the implementation. Please provide an estimated time of completion for the task.

Solution design test

Deliverables:

- Implementation of an algorithm for counting values of a feature across content items.

Background: your key responsibility will be creating logic that extracts valuable insights from constant flow of data. Data can be of different types (communication, transactions, voice, registry items, meeting logs, calendar events, user activity, change management, data quality records etc.). The data processing will vary from collecting and showing statistics to calculating features and findings outliers. In general, you won't collect information - you will write tools to do it on the client side in another dataset. The tools will be installed and used by other people. Some other people, including clients, can see the code.

From a programming language standpoint, in our tools we use Behavox proprietary DSL and Groovy. NLP is almost entirely in Python, and some Python NLP tools will be available as well. We use Java for testing. Some other languages can be present as well to a limited extent.

To be successful in this role you must be able to find important features, generalize, write production quality code and document your solution. Code quality for a CRX analyst means above all fast and safe execution even on big data, handling of expected exceptions and data deficiencies even if you haven't met them in the test dataset.

You will also need to have thoroughness, clarity, and originality of thinking required to create an efficient enterprise software product.

Problem:

We provide the dataset. It consists of a "Communication samples" folder with communication items, and a "Contacts" text file. The communication includes publicly available items taken from the Enron case, and some artificial content items. The text file contains contacts of one person, also partially artificial.

Your task is to implement a tool that counts the value of one selected feature for each content item. You can select which feature you prefer to implement. Proposed variants are: **Stress**, **Closeness** and **Collusive behaviour**. The script can be implemented in Python, Groovy or Java, as you will need these for the work. Mind that we might want to test it on other communication items of the same format.

You must share the tool with us via bitbucket.

In Behavox system we have access to meta information and features extracted by NLP, including several ways to classify communication and recognized entities. For this assignment you receive the raw dataset. What features could help you with the task?

About the Behavox Platform: Behavox is a data analytics Platform processing and storing all corporate data sources including but not limited to emails, chats, audio recordings, HR data, CRM data, etc. Each Behavox client has a dedicated, secure and fully isolated instance of Behavox. Behavox products such as Behavox Compliance are deployed on top of the Platform and leverage its computational, storage and analytical foundations.

Each data point ingested into the Platform goes through a data processing pipeline responsible for running data quality checks, normalizing data, linking to relevant persons (initialized in the Platform using HR and CRM data) as well as extracting various artifacts using a suite of extractors and Machine Learning classifiers. Examples of artifacts are signatures in emails, tickers in chats, legal documents in attachment, voice-to-text transcriptions, etc.

Both raw data and derivatives calculated by Behavox analytics engine are stored in the data lake. The Platform features a native IDE, a rich set of APIs leveraging Apache Groovy as well as proprietary visualization and behavioral modeling tools.