Augmento Data Science challenge

Why you should give your best

Augmento is currently receiving many internship requests in the field of data science, machine learning and natural language processing. In order to better assess applicants, a one week data science hackathon is part of our recruitment process. Here you get the chance to convince us with your creativity, skills as well as the willingness to learn and work hard. Those characteristics are closely related to the core values of our company and will shape your internship experience.

Let's get started!

The challenge

We are providing you with a collection of tweets related to the topic "FinTech".

We challenge you to use this data in a creative way and discover insights and trends. Here are some general example questions we find interesting, but remember these are only to get you thinking and we welcome novel approaches to gaining insights.

What are people talking about FinTech on twitter?

Which fears and hopes are they expressing towards new technologies or businesses?

Which companies and locations are mentioned by twitter users?

Where are twitter users posting about FinTech coming from?

What other valuable insights about "FinTech" discussions can be extracted using Twitter user metadata?

…

Expected outcome and how we measure success

We would like to see interesting findings of any kind, ideally visualized in an appealing way. We would like to hear your thoughts, which methods you've used and how you approached challenges. Of course, we are also interested in seeing your code! You may deliver your results in any form you see fit - an IPython notebook or a PDF report will be perfect. You can use any programming language and any libraries you like. However, Python is a prefered choice, as this is what we are mostly working with at Augmento.

We encourage you to take the challenge no matter the current level of your skills–in case you do not have much experience in NLP yet your research process and the ability to dive into new topics will be evaluated as well.

Data set

The tweets are stored in a pickle file. You will need Python together with *pandas* library installed to open and use it. You can load the file like this:

```
import pandas as pd
dataframe = pd.read_pickle("fintech_cleaned.pkl")
```

Last column in the data frame - *Stakeholder* - does not come from Twitter directly. It represents a stakeholder group the user was automatically classified to (*Ambiguous* means there is no label). You can include this data in your analysis, too.

If you never used *pandas*, you can read about it here:
http://pandas.pydata.org/pandas-docs/stable/tutorials.html
Or here:
https://www.dataquest.io/blog/pandas-python-tutorial/

Be careful: the data set may contain missing data, newline characters, non-English characters, etc. It is up to you how to decide to handle it.

Good luck! :)