Summary:

At Pitchbook we're crawling thousands of webpages every day. In order to better serve our clients, we have to make sure the data we pull from these websites are high quality.

Project Definition:

Given a list of extracted sentences from company about pages, write and deploy a machine learning classifier of your choice, that will determine whether or not the text contains descriptive content. Here are some examples of company about pages and their descriptive content:

> https://www.nuance.com/about-us.html
> *Our innovations in voice, natural language understanding, reasoning and systems integration come together to create more human technology.*
> https://stripe.com/about
> *Stripe is a technology company that builds economic infrastructure for the internet. Businesses of every size—from new startups to public companies—use our software to accept payments and manage their businesses online.*

Definitions:

Descriptive content: Any sentence from a company's website, that indicates one or more of the following

- Products/Services
- Market Segment, Market Profile or Market Location
- Value Proposition or Company Purpose

Labeled Data:

You'll notice 4 columns of information

- About page website text (sentence)
- Label (about or none)
- Article ID
- Offsets (if label is "About", where did the descriptive text occur and what were its attributes). You need not utilize this column to build your system, but the additional data can provide more context or ground for building out additional features.

Project Scope

1) Model Build: Build a machine learning model of your choice (logistic regression, CNN, etc). Please be ready to speak to all aspects of data discovery and exploration, data

normalization, model training and the quality of the model. Using existing libraries is strongly encouraged - no need to reinvent the wheel.

2) Model Deploy: Architect a deployment strategy for taking the model built in step 1, into a production environment. If you have the time and resources, deploy onto a cloud platform (AWS, Azure, etc). Be prepared in the on-site interview to discuss in depth all aspects:
    a) Interfacing with the model in the wild: as new data comes in, how will classifications happen?
    b) Code updates: if you or other engineers want to update the existing model, how will it work? Is CI/CD possible?
    c) Multiple environments: what is your recommendation for multiple environments? For example: dev / test / prod
    d) Cloud resources: What AWS/Azure or other solutions would you recommend using in order to achieve steps a-c. This could include docker containers