FullContact - Business Card OCR Engine

About Us

1. We're a startup located in LoDo (23 paces away from the Wynkoop Brewery).
2. We're TechStars graduates.
3. We're solving the world's contact information problem.
4. We <3 Big Data.
5. We <3 machine learning.
6. We were founded in 2010.
7. We're lean, mean and venture-funded.
8. We're not afraid of bleeding-edge technologies.

About You

1. You want to get experience with machine learning, natural language processing, and data mining
2. You're familiar with the JVM.
3. You know Java, but you prefer other JVM languages like Groovy, Clojure, Scala, JRuby, or Jython.
4. You have a basic understanding of machine learning and/or image recognition concepts.
5. You enjoy working with and building RESTful APIs.
6. You can't think of a better version control system than Git.
7. You took more math than you needed to graduate.
8. You find code without good test coverage to be brittle and unpleasant.
9. You know your way around a *nix shell like bash or zsh.

Problem Description

FullContact is building a suite of applications that will attempt to rid the world of paper business cards. Since the transition will be slow (most business people have stashes of **thousands** of outdated business cards in their drawer), we need to build-in some very high quality OCR technology for transcribing business cards. This data will then be ported as structured data into various systems such as Salesforce and Highrise.

For some backstory, in March the FullContact team took Austin Texas by storm, clad in wingbandits at SXSW 2012. The guy on the left in this photo featured on Mashable is Colorado School of Mines' own Michael Rose, a member of the 2011 Field Session team that worked with FullContact. Our goal as a company at SXSW was to talk to as many people as we could about our product and collect contact information from the people we talked to. By the end of the week, we had collected over 500 business cards. But then it hit us, who has time to enter 500 business cards into SalesForce manually? Funny enough, a writer for CNET had the same problem.

You will research and analyze the performance of the leading open source and commercial OCR tools available on the market. Based upon your findings, you will implement a RESTful service around the ideal candidate(s) that accepts images of business cards and returns structured data such as name, company, title, street address, phone number, and email address. The ideal solution will have a high accuracy rate, will have the ability to recognize various data attributes and organize accordingly, and will operate at high transactional volumes.

You will likely have to train several machine learning algorithms and deploy several solutions for testing purposes. Initially, we'll help you build a training and test set of business card images and expected results, which you can use as an even playing field when evaluating different options.

We think this is an ideal project for a team of **2-3 people**, but **we'll listen to you** if you think it should be more or fewer.

You should choose this project because...

1. You'll get to experience life at a venture-funded startup.
2. You will get to work with some very cool machine learning algorithms and implementations.
3. You will get to work with cool non-relational data stores like Redis, HBase and MongoDB.
4. You'll get to use alternative JVM languages like Groovy and Clojure.
5. We're a fun bunch of people.
6. If all goes well, we might even hire you after the project!

Questions?

Email me at dan@fullcontact.com