HP

HP Inc. is a Fortune 500 American multinational information technology company that develops personal computers (PCs), printers and related supplies, as well as 3D printing solutions. The company's first office is referred to as the birthplace of Silicon Valley. At one point HP was the largest manufacturer of PCs in the world.

Project Description

HP is targeting an audience that prefers paper documents with the hope of helping them to slowly transfer physical documents to digital documents using HP products. **This project is a continuation of HDAG's engagement from last semester where a case team worked to build an NLP model,** using TF-IDF/raw text NLP featurization to try to predict whether a form requires a signature. The previous team has already extracted the most common words (unigrams) and pairs of words (bigrams) from a sample dataset and applied Tf-Idf vectorization to implement basic NLP models to detect forms that require a signature.

The case for this semester would build upon this progress by performing two main tasks. The first task would be to expand the dataset of scanned forms so that the model has more data to work with, which would hopefully improve the accuracy of the NLP classifier. The second task would be to experiment with more classification techniques including bootstrapping and decision trees. The second part of the project has more freedom and would allow the case team to explore many different and complex ways of classifying documents.

Internal Partners: This project is under HP's Imaging AI/ML Solutions Strategist

Coding: Python and knowledge of (or ability to learn) various natural language processing and classification techniques

Specific Skills

1. Natural Language Processing
2. Basic data management with Pandas and Numpy were and scikit-learn for vectorization and the predictive model
3. Bootstrapping and decision trees

Expected Technical Difficulty: **Intermediate / Advanced**