TRƯỜNG ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

09/12/2022

Introduction to Data Science

Project Progress Report

Job Posts Classification for Online Recruitment Websites

Group 4 – DSAI K65

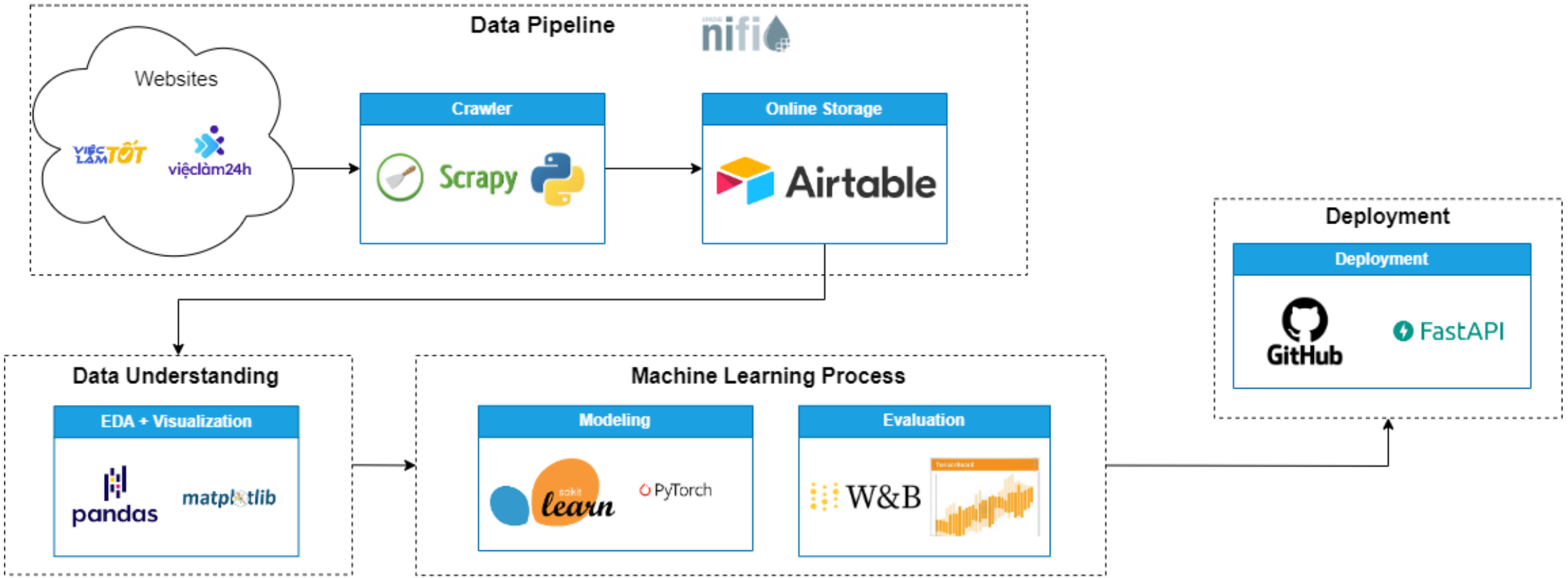
ONE LOVE. ONE FUTURE.

Table of Contents

Project objectives

- Create an automatic system to:
  - Classify online job posts based on the type of job from online platforms.
  - Determine the credibility of these posts:
    - Is the post spam or not?
    - Is the information consistent within the post?

Current project workflow



Data pipeline and EDA

What we have done:

- Using Scrapy, Apache Nifi, and Airtable to crawl and store the data
- Integrate the crawled data between two websites
- Detailed analysis of the small data subset

Difficulties:

- Limited storage to store the data

Plans:

- Maybe crawling from other websites as well
- EDA on the big dataset

id	100852566
company_id	1417234
post_time	2022-08-11 01:25:28
description	- Chỉnh sửa sản phẩm (hàng thời trang nữ) theo...
vacancies	1
min_salary	6500000
max_salary	6500000
age_range	18-
gender	Không yêu cầu
benefits	BHXH, Phép năm, thưởng sinh nhật...
education_requirements	Không yêu cầu
experience_requirements	< 1 năm
contract_type	Toàn thời gian
job_location	Phường Bến Thành, Quận 1, Tp Hồ Chí Minh
salary_type	Theo tháng
url	https://www.vieclamtot.com/viec-lam-quan-1-tp-...
created_time	2022-11-28 16:35:28
updated_time	2022-11-28 16:35:28
skills	Biết sử dụng máy may
job_type	Thợ may tại nhà
title	Công Ty Nét Việt Tuyển 01 Thợ Sửa Tại Quận 1

Data instance example

Job title wordcloud

Modelling and Deployment

What we have done:

- Baseline model with BOW + SVM for post classification

Difficulties:

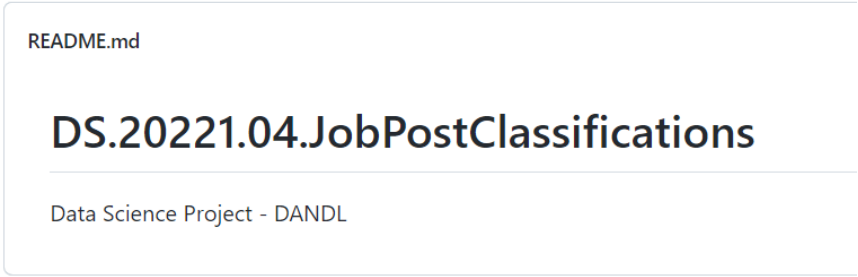
- Imbalanced classes
- Vietnamese natural language
- Choosing light-weight models for deployment

Plans:

- Use more sophisticated model (BERT, ..)
- Create web API by using FastAPI or Flask
- Publish code on Github with detailed instructions

```
Accuracy: 0.4676
Precision: 0.4901
Recall: 0.4676
F1 Score: 0.4344
```

Results of baseline appoarch



Group github

Summary

- What we have done:
  - Built the data pipeline to crawl and store job posts data
  - Analyzed the small subset of the dataset
  - Developed a simple model for job post classification
- Future plans:
  - Extract the information from text field
  - Analyze the bigger crawled dataset
  - Use a better model
  - Deployment

HUST

hust.edu.vn fb.com/dhbkhn

Thanks for Listening!