PROJECT PLAN

Mentor:

Julian Brooke

MDS-CL Team:

Gurpreet Bedi

Simon Zheng

Lisa Liu

Index

Overview

Indeed is one of the best job sites in the world, with a lot of users visiting each day, either to apply for a job or lookup for specific job-related articles. Since Indeed is an open-source website, it has both categories of users: **registered** and **non-registered**. In order to enhance the user experience and personalize the user's view, we need to build an Multi-task multi-label text classification model, which would automatically predict the reader attributes based on the user activity on Indeed. This may further aid Indeed in using these attributes to recommend the user with relevant content.

Datasets

The data provided by the client refers to Indeed Career Guide site. All the data is in EN_US. There are **five** type of content (.json) files provided by Indeed, named as below:

- article.json
  - One of the Indeed pages, consisting of "free-form" articles managed by Indeed Editorial Team. The majority of the content "types" fall into this category.
  - There are a total of 13,639 distinct articles.
  - Some of the example of category in articles include:
    - Career Development or Interviewing
- careerpathpage.json
  - Content under career highlights, the type of a career a person can pursue. All the content is organized and looks similar.
  - The content here has a total of 541 distinct career paths.
  - Some of the examples include:
    - Learn about being a Pharmacologist or Learn about being a Forester
- categorypage.json
  - One of the Landing pages, categorically highlighting different types of content eg: Resume Samples. These pages themselves don't have meaningful content nor are they intended to serve a particular audience.
  - There are a total of 24 distinct categories.

- coverletter.json and resumesamplepage.json:
    - These are structured pages for sample cover letters and resumes, respectively.
    - There are 340 distinct CoverLetters and 344 distinct Resume Samples.
    - Some of the examples include:
        - Architect Cover Letter Sample or Research Assistant Resume Sample

Field Specifications

Field Name	JSON File Object Name	Description
ID	**_id** **$oid**	This corresponds to the respective document id.
Title	Possible object names: • **title** • **contentTitle** • **h1**	This displays the title of the given content type
Content	Possible object names: • **content** • **primaryContent** • **contentA** • **contentA + contentB** (concatenate the two fields)	This displays the content of the given title.
Locale	**locale**	Language used for this content
Category	**category**	Domain used to build the content route url
UrlRoute	**urlRoute**	Routes used to build the content route url

Expected Deliverables

The expectation here is to build a Multi-task multi-label text classification model, which could automatically predict the user attributes from the user activity on the Indeed pages.

Some of the major deliverables would include:

- Manual Annotation
  - We would be manually annotating on a good number of articles (around 5000) and tag these attributes to the articles respectively.
  - The list of attributes would broadly classify all the articles.
  - These attributes should be able to infer as to **"why"** the user is reading this article. We would categorize the user's attributes as user's **status**, **intention** and **experience**, with each attribute having multiple labels associated.

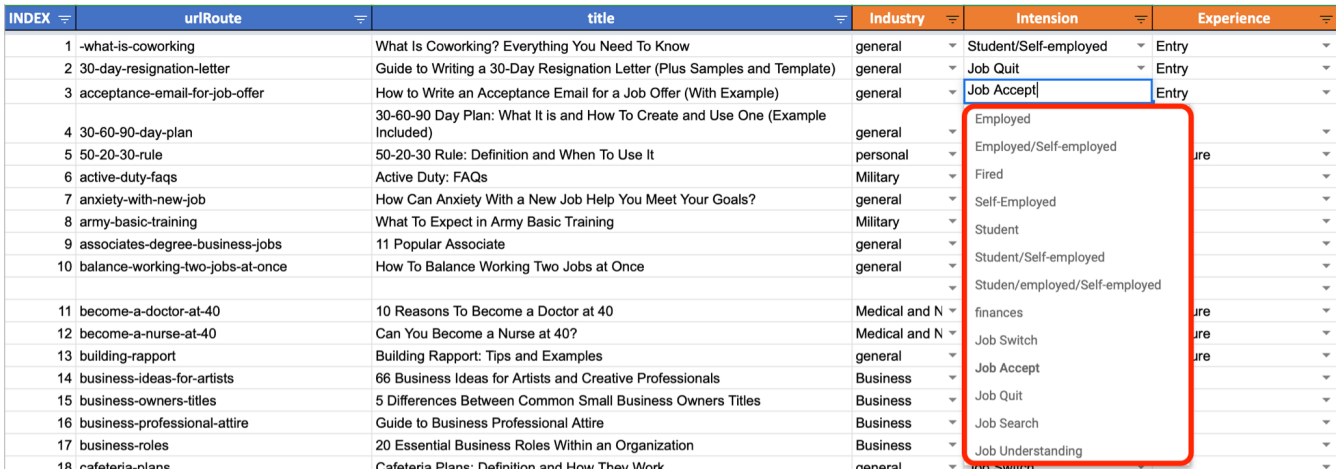
- Multi-task multi-label text classification model

  Once the attributes are manually tagged, we would then use it to train a multi-task multi-label text classification model which could classify the user attributes for untagged articles.

Multi-Tasks

INDEX	urlRoute	title	Industry	Intension	Experience
1	-what-is-coworking	What Is Coworking? Everything You Need To Know	general	Student/Self-employed	Entry
2	30-day-resignation-letter	Guide to Writing a 30-Day Resignation Letter (Plus Samples and Template)	general	Job Quit	Entry
3	acceptance-email-for-job-offer	How to Write an Acceptance Email for a Job Offer (With Example)	general	Job Accept	Entry
4	30-60-90-day-plan	30-60-90 Day Plan: What It is and How To Create and Use One (Example Included)	general	Job Accept	Entry
5	50-20-30-rule	50-20-30 Rule: Definition and When To Use It	personal	finances	Not Sure
6	active-duty-faqs	Active Duty: FAQs	Military	Job Switch	Entry
7	anxiety-with-new-job	How Can Anxiety With a New Job Help You Meet Your Goals?	general	Job Switch	Entry
8	army-basic-training	What To Expect in Army Basic Training	Military	Job Understanding	Entry
9	associates-degree-business-jobs	11 Popular Associate	general	Student	Entry
10	balance-working-two-jobs-at-once	How To Balance Working Two Jobs at Once	general	Employed	Mid
11	become-a-doctor-at-40	10 Reasons To Become a Doctor at 40	Medical and Nurs	Job Understanding	Not Sure
12	become-a-nurse-at-40	Can You Become a Nurse at 40?	Medical and Nurs	Job Understanding	Not Sure
13	building-rapport	Building Rapport: Tips and Examples	general	Employed	Not Sure

Multi-Labels



INDEX	urlRoute	title	Industry	Intension	Experience
1	-what-is-coworking	What Is Coworking? Everything You Need To Know	general	Student/Self-employed	Entry
2	30-day-resignation-letter	Guide to Writing a 30-Day Resignation Letter (Plus Samples and Template)	general	Job Quit	Entry
3	acceptance-email-for-job-offer	How to Write an Acceptance Email for a Job Offer (With Example)	general	Job Accept	Entry
4	30-60-90-day-plan	30-60-90 Day Plan: What It is and How To Create and Use One (Example Included)	general		
5	50-20-30-rule	50-20-30 Rule: Definition and When To Use It	personal		
6	active-duty-faqs	Active Duty: FAQs	Military		
7	anxiety-with-new-job	How Can Anxiety With a New Job Help You Meet Your Goals?	general		
8	army-basic-training	What To Expect in Army Basic Training	Military		
9	associates-degree-business-jobs	11 Popular Associate	general		
10	balance-working-two-jobs-at-once	How To Balance Working Two Jobs at Once	general		
11	become-a-doctor-at-40	10 Reasons To Become a Doctor at 40	Medical and N		
12	become-a-nurse-at-40	Can You Become a Nurse at 40?	Medical and N		
13	building-rapport	Building Rapport: Tips and Examples	general		
14	business-ideas-for-artists	66 Business Ideas for Artists and Creative Professionals	Business		
15	business-owners-titles	5 Differences Between Common Small Business Owners Titles	Business		
16	business-professional-attire	Guide to Business Professional Attire	Business		
17	business-roles	20 Essential Business Roles Within an Organization	Business		
18	cafeteria-plans	Cafeteria Plans: Definition and How They Work	general		

Employed
Employed/Self-employed
Fired
Self-Employed
Student
Student/Self-employed
Studen/employed/Self-employed
finances
Job Switch
Job Accept
Job Quit
Job Search
Job Understanding

Methods

Capturing User Attributes

Starting with annotation, the list of some of the attributes which could be applicable to the Indeed articles is as follows:

- Status: **Employed, Self-employed, Not employed**, etc.
- Intention: **Job Search, Job Switch, Job Accept, Job Quit**, etc.
- Experience: **Entry, Mid, Senior**, etc.

Below are the **three** annotation approaches to be discussed with Indeed:

1. **UBC annotates and Indeed provides sign-off**
    - For training the model, a dedicated team member would annotate around 5000 articles, to produce a list of articles with labelled attributes.
    - We then as a team, would internally perform the first level of review of the articles with labelled attributes.
    - Afterwhich, Indeed would perform a second level of review where they review attribute tagged articles, suggest updates (if required) and provide an approval.

2. Indeed to perform annotation
    - Indeed to set up the annotation on their end and provide the list of annotated articles.

3. Using Amazon Mechanical Turk
    - We would suggest Indeed, if they agree to have annotation work done using Amazon Mechanical Turk.
    - We would then need to perform Inter-annotator agreement study for articles annotated using mechanical turk.

Machine learning models

Unsupervised Approach

Since we do not have any previously attribute labelled articles, one approach we could follow is unsupervised approach for classifying attributes and confidence levels:

- Zero-shot classification

We would be performing an unsupervised zero-shot classification (model from hugging face) for a certain set of articles, and document the results.

- Topic Modelling

We would be performing unsupervised LDA Topic modelling using **scikit-learn** and **gensim**, and comparing the quality of results.

Supervised Approach

In order to perform a supervised approach, we would use the manual annotated data for training the machine learning model and make predictions on the unlabelled dataset. As, not all the text could be classified using a single model, therefore, we might have to use multiple models for multi-label text classification:

- **Rule based text classification**
  - One of which could be a **Rule based text classification** model which may use regex for some of the articles that can be directly classified.
- **Multi-label text classification**
  - Some of the articles can be classified using multi-label classification.
- **Pre-trained model (BERT)**
  - We might use a pre-trained model like **BERT** (for contextualized word embeddings).

Note: The JSON files contain fields such as "title" and "content". It is comparatively faster for humans to annotate the article based on "title" and predict reader attributes, whereas a machine learning model would require both "title" and "content" information for classifying the attributes.

Evaluation

For multi-task multi-label text classification, we would use scikit-learn's **f1-score** as the evaluation metric, because here we are more concerned about false positives and false negatives (i.e. misclassification of the articles).

Libraries

Some of the libraries which may be required for this project are:

- pandas
- numpy
- collections
- json
- gensim
- transformers
- sentencepiece
- scikit-learn
- matplotlib
- wordcloud

Schedule

Below are the tasks listed for the project:

Week 1 : Analyzing Datasets

- Analyzing the dataset
    - Reading the JSON files
    - Identifying the fields associated with each JSON file.
    - Perform data analysis
- Manually annotate a small set of articles from **article.json** to identify user attributes and get this list of attributes approved from Indeed.

Week 2 : Implementing Unsupervised Approach

- Performing unsupervised approach to predict user attributes:
    - Zero-shot classification
    - Topic modelling
        - Scikit-learn LDA model
        - Gensim LDA model
- Documenting the results
    - Comparing the results for unsupervised approaches.
    - Identifying the one which has a better quality.

Week 3 + Week 4: Manual Annotation & Implementing Supervised approach

- We manually annotate **article.json** and tag a good amount of articles (around 5000) and internally review the labelled articles.
- Get user attribute list approved by Indeed.
- In parallel, we would build a multi-task multi-label text classification model with a supervised approach. In order to achieve that, we may use following models for text classification:
    - Rule based model for attribute prediction
    - Multi-label text classification model for attribute prediction
    - Pre-trained models like **BERT** model based attribute prediction

Week 5 : Hyper-parameterizing and Evaluation

- Hyper-parameterizing the machine learning model(s) for better results.
- Evaluating the results from two different approaches (unsupervised and supervised), and documenting them.

Week 6 : Integration and Code Review

- Code Review:
  - Reviewing the teammates' code.
  - Code Cleanup (if required)
- Integration
  - Integrating all the modules
- End-to-end execution of the integrated code, to ensure successful completion of the project.

Week 7 : Report writing and Presentation

- Developing a final report detailing all the project components.
- Creating a video presentation.

Weekly Stand-up's

Below is our weekly stand-up schedule:

- Meeting with Indeed partner : Thursdays, 11:00 AM.
- Meeting with mentor : Tuesdays, 3:00 PM
- Regular weekly team meetings:
  - Mondays, 11:00 AM - defining framework for the week.
  - Wednesdays, 1:00 PM - intermediate progress catch-up.
  - Fridays, 11:00 AM - review and wrap-up of weekly checkpoint.

Note: Since, the project discussion is at a very preliminary stage, therefore, the models mentioned above are just a notion of what we understand from the current dataset provided by the client. We would follow a hybrid-agile methodology which is an iterative and incremental approach for project development, driven by both what we have investigated during dataset analysis (i.e. manual and unsupervised techniques) and what the client is looking for.