AUTOMATIC INFERENCE OF READER ATTRIBUTES FROM BLOG CONTENT

Tony Hui, Indeed

ORGANIZATION DESCRIPTION

Indeed is one of the most popular job sites in the world, serving over 250million unique visitors per month. Our mission is to help people get jobs. The primary way we do this is to aggregate all the jobs on the internet, and allow businesses to post jobs directly. A lesser known way we help people is by providing career advice for job seekers (https://www.indeed.com/career-advice/). We commission career coaches to write helpful articles for a variety of topics ranging from interviewing tips, workplace conflict, and anything else that's work related.

BRIEF PROJECT DESCRIPTION

In order to better help people get jobs, we want to know more about our users. Currently, we do this from understanding their resumes as well as their site activity. However, for users who come to Indeed Career Advice (and are not logged in), we don't know anything about them and therefore can only serve generic information and calls to actions.

However, based on the content they are reading, it should be possible to infer *why* they're reading the page, and from that infer some attributes (structured data) about these users to better serve them relevant content. Additionally, we can use these attributes to improve their experience on the rest of the site once they create an account.

For example, if someone visits an article about "How to Write a Resignation Acceptance Letter (With Template and Examples)", we can probably assume the following (with varying levels of confidence):
- They are currently employed
- They are a manager
- They are probably going to hire to backfill their resigning employee

If the same person then visits Indeed.com, we might want to show a *Post your job!* or *Explore market salaries* call to action on the homepage as we know they probably don't want to search for jobs.

I think of this as a multi-class unsupervised topic modelling problem. Could be wrong though.

AVAILABLE DATA

All publicly available indeed career guide pages supplied in raw text (markdown format). About 7000 articles. Also includes metadata (like HTML <meta> tags, etc)

We can also supply aggregate measures of traffic (how many people visited X page in last year?) if that's helpful.

We are able to share additional data as needed.

DATA PRODUCT

1) Code to serve ML model via a REST API (or a python function that can easily be "imported" as a module or similar).
Language: Python

Arguments:
- Page content (url, full text, title, or whatever else)
- Some sort of confidence threshold

```
 JSON output:
[
   {  \
      ""attribute"":""manager"",
      ""confidence"":0.9
   },
   {
      ""attribute"":""employed"",
      ""confidence"":1
   },
   {
      ""attribute"":""hiring"",
      ""confidence"":0.4
   }
]
```

2) A 10-minute presentation
Briefly describe what was done and why the results are trustworthy

LEGAL RESTRICTIONS

UBC must have commercial and cyber insurance for data breaches

Other Information.

Yes	Will students be able to share details of the capstone in a private job interview?
Yes	Will students be able to share details of the project in a blog post?
Yes	Will students be able to give a public presentation about their work (With sensitive details removed)?
No	Will students be asked to sign an NDA?
No	Students will be required to have a background check?
Yes	Do you anticipate having any data scientist opening(s) following completion of the project?