JOB RECOMMENDER FOR GITHUB USER

AKHIL GONNA #1610658
RAVALI VARANASI #1610598
UNIVERSITY OF ALBERTA

Abstract: GitHub becomes the largest open source community in the world. Besides sharing code, GitHub is also a social network, in which developers can follow others or keep track of their interested projects. Considering the multi-roles of GitHub, integrating heterogenous data of each developer to identify experts is a challenging task. We can also keep tracking on the projects and the languages the users used. So, based on this concept we have proposed a concept called "Job Recommender for GitHub User". This is more likely an interactive project where we take input from the user and display results. Our project has generally two parts: 1) Dynamic job recommender and 2) Job recommendation from a dataset. We are trying to match the similarity of the languages used and the languages required for the job postings. This project also can explain about the expertise programing skill of the user.

1. Introduction:

In this project, we are using the data from GitHub by using GitHub API's. The main aspects are the user's location and the repositories of the username given dynamically. We are analysing the repositories by extracting the languages used in the repositories and sorting based on the usage.

Application Programming Interface (API):

- An application program interface (API) is a set of routines, protocols, and tools for building software applications.
- Methods to access data and workflow from an application without using the application itself.

Why do we use API: In building applications, an API simplifies programming by abstracting the underlying implementation and only exposing objects or actions the developer needs. Each developer has different specifications and requirements on a specific application. So, they can filter the extracted data and then analyse, process as per their needs.

We used "REST API v3" to extract the data in json format. All requests are done by using: https://api.github.com. Throttling limit for this API is 5000/hour. We can create an application by using GitHub account and generate a secret key to use the API.

We get the user bio and all the repositories. From the data of repositories, we fetch the languages used in each repository used. Then we store all the languages in one list and then analyse on them. We sort all the languages based the number of times used.

From here the project has two parts for the job recommendation.

1.1 Dynamic Job Recommendation:

In this step, we are fetching the data from Stack Overflow jobs website by scraping from python using beautiful soup.

Beautiful Soup: Beautiful Soup is a Python library for pulling data out of HTML and XML files. It works with your favorite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. We have used this library to get data from job posting websites.

In the next step; we pass the location of the user and languages used of repositories into the website and fetch the data by web scraping. Initially we ask the user whether he has any specification of the location. Then based on the user's input we display the recommendations. We also fetch the job posting from the nearest locations if there are any recommended jobs for the user in his preferred location.

1.2 Job Recommendation from dataset:

In this method, we are using a dataset from a website 'Kaggle' which is having fields: job title, job link, job type, skills, company name. etc. Mainly we are concentrating on the skills field of the dataset. We have checked the similarity match for this skills field of the dataset and the skill sets of the user. We get the ranks for the jobs and display the top 50 ranked jobs for the user.

2 RELATED WORK:

As we all know GitHub is providing a wide range of datasets to perform some social network analysis. We have lots of researches done by the expertise based on the languages. 'Behavior-based Expert Identification' [1] explained in the GEMiner based on the programming behaviour of the user. John and Gail [2] presented an empirical evaluation of two approaches to determining implementation expertise from the data in source and bug projects. When determining experts for fixing bugs, they consider two factors such as bug reports and bug networks. However, the networks in bug reports is less complex than networks in GitHub.

Code reviewer recommendation in GitHub based on cross-project and technology experience [4] have done some work on estimation code review expertise of a developer for a pull request by analyzing her past work experience. 'Recommending GitHub Projects for Developer Onboarding' [3] have used language similarity match and proposing GitHub project for the new developer of the project. There are few rank algorithms performed on them to get recommendations.

3 Concept:

To get job recommendation for the user we are initially taking the location of the user and fetching the jobs based on location. If there are no jobs in that location for the user, then we are fetching the jobs near by that location for the user. There is also an issue for getting data from GitHub API because there is a throttling limit of 5000 requests/hour. To avoid this, we can register and get an access token from GitHub account.

In the first step, we ask for the GitHub username as an input and search in GitHub for that username. If we do not have any matches, then ask to re-enter the username. If there are more than one matches then, we confirm the user whether the input is correct or not. If user confirms yes, then we display the details of the user. We also fetch all the details of the user repositories and make some analysis on the languages used. We also display the user a bar graph of the language analysis.

From here we have two parts of job recommendations. They are:

3.1 **Dynamic job recommendation:**

Based on the above data we take the location and the languages used by the user. We ask the user whether he wants only to work at his location. If yes, we display job recommendations from his location. If no, we display the most popular jobs based on the languages used in his repositories. This data is fetched from Stack Overflow jobs using a web scrapping library in python using beautiful soup. We must trim lots of fields from the data we receive from the html we receive from beautiful soup.

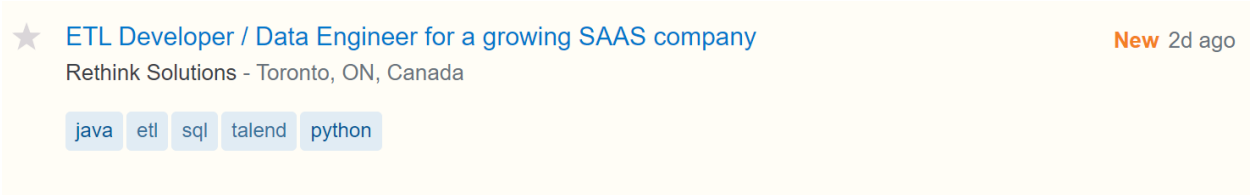

Figure 3.1: Screen shot of Stack overflow job posting

From above example we fetch the job title, job location, tags, company name and age of post by passing location and our languages used into the field of tags in html. We display the results to the user based on his GitHub profile.

3.2 Job recommendation from dataset:

In the second part of our project we are extending to fetch top 50 job recommendation from 5000 data entries of dataset. We have collected this data set from Kaggle. As we can see from below screenshot of the dataset, we have fields like job title, job link, salary, job type, skill, number of skills required, company name. etc. We are mainly filtering based on the skill field. We are ranking these jobs based

on the job skills and the languages used by the user in his repositories. We are ranking by using the formula:

$$Rank = \frac{Number\ of\ similar\ skills}{Total\ skills\ required}$$

The sample data from datset:

Job_Title	Link	Queried_Sa	Job_Type	Skill	No_of_Skill	Company	No_of_Rev	No_of_Star	Date_Since	Descriptior
Data Scient	https://ww	<80000	data_scient	['SAP', 'SQL']	2	Express Scr	3301	3.3	1	[<p><b>P
Data Scient	https://ww	<80000	data_scient	['Machine Learning', 'R', 'SAS', 'SQL',	5	Money Mart Financial Services			15	[<p><b>W
Data Scient	https://ww	<80000	data_scient	['Data Mining', 'Data Management',	9	comScore	62	3.5	1	[Va
Graduate S	https://ww	<80000	data_scient	['Certified Internal Auditor']	1	Central Inte	158	4.3	30	[<p>Full
Data Scient	https://ww	<80000	data_scient	['Statistical Software', 'Time Manage	7	Federal Res	495	4.1	30	[As
Data Scient	https://ww	<80000	data_scient	['AI', 'Quantitative Analysis', 'Data M	6	National Se	173	4.3	30	[C
Geospatial	https://ww	<80000	data_scient	['Statistical Software', 'Machine Lear	10	NYC Career	30	3.8	5	
Data Scient	https://ww	<80000	data_scient	['Machine Learning', 'R', 'SQL']	3	OM Partners			10	[<p>With
Bioinforma	https://ww	<80000	data_scient	['Linux', 'R', 'C/C++', 'Python']	4	University	233	4.2	1	[M
Data Scient	https://ww	<80000	data_scient	['JavaScript', 'Data Mining', 'TS/SCI C	6	usajobs.gov	4227	4.3	22	[
Data Scient	https://ww	<80000	data_scient	['Machine Learning', 'R', 'SPSS', 'Data	8	The Univers	541	4.2	1	[<p>The
Data Scient	https://ww	<80000	data_scient	['Machine Learning', 'R', 'SPSS', 'Data	8	The Univers	541	4.2	1	[<p>The
Data Scient	https://ww	<80000	data_scient	['Machine Learning', 'Analysis Skills',	3	Rice Univer	150	4.3	30	[<p>Salary
Jr. Data Sci	https://ww	<80000	data_scient	['Machine Learning', 'Python']	2	Elev8 Hire Solutions			6	[<p>Mid
Data Scient	https://ww	<80000	data_scient	['TensorFlow', 'Project Planning', 'Lir	7	Catalina Ma	81	3.2	30	[<p><b>W
Enrollment	https://ww	<80000	data_scient	['Machine Learning', 'R', 'Statisical A	11	Florida Poly	16	3.9	1	[<p>Locat
Junior Data	https://ww	<80000	data_scient	['Machine Learning', 'R', 'SQL', 'Analy	6	Achievement Network (ANet)			30	[<p>
Data Scient	https://ww	<80000	data_scient	['Microsoft SQL Server', 'Data Mining	12	Deloitte	7197	4	30	[<
Data Analy	https://ww	<80000	data_scient	['Machine Learning']	1	ExxonMobi	2234	4	11	[<p>Exxon
Postdoctor	https://ww	<80000	data_scient	['LMRT', 'Machine Learning']	2	MIT	278	4.2	5	[<p><b>W

Figure 3.2: Dataset screen shot

For example, if the job requirement skills are [SAP, SQL] and the user skills from repositories are [SQL, JAVA, C]. Then the rank of this job posting will be ½ = 0.5. Thus, we derive ranks for all the job postings and display top 50 postings to the user.

4 Experiments and Results:

We have evaluated for the GitHub username ‘petertodd’ for our analysis and the job recommendations. After entering the username, the information is displayed as shown in the below screen shot.

```
Enter your GitHub username:petertodd
We have more than one result on petertodd
Please enter yes or no for this username yes
url: https://api.github.com/users/petertodd
Information about user 'petertodd':
ID: 7042
Email: None
Location: Toronto
Public repos: 119
About: Applied Cryptography Consultant (what the cool kids call 'blockchain tech')
```

Figure 4.1: Basic Info of the user

We have verified this by manually opening the account link and the results are verified as we can see in the below screenshot.

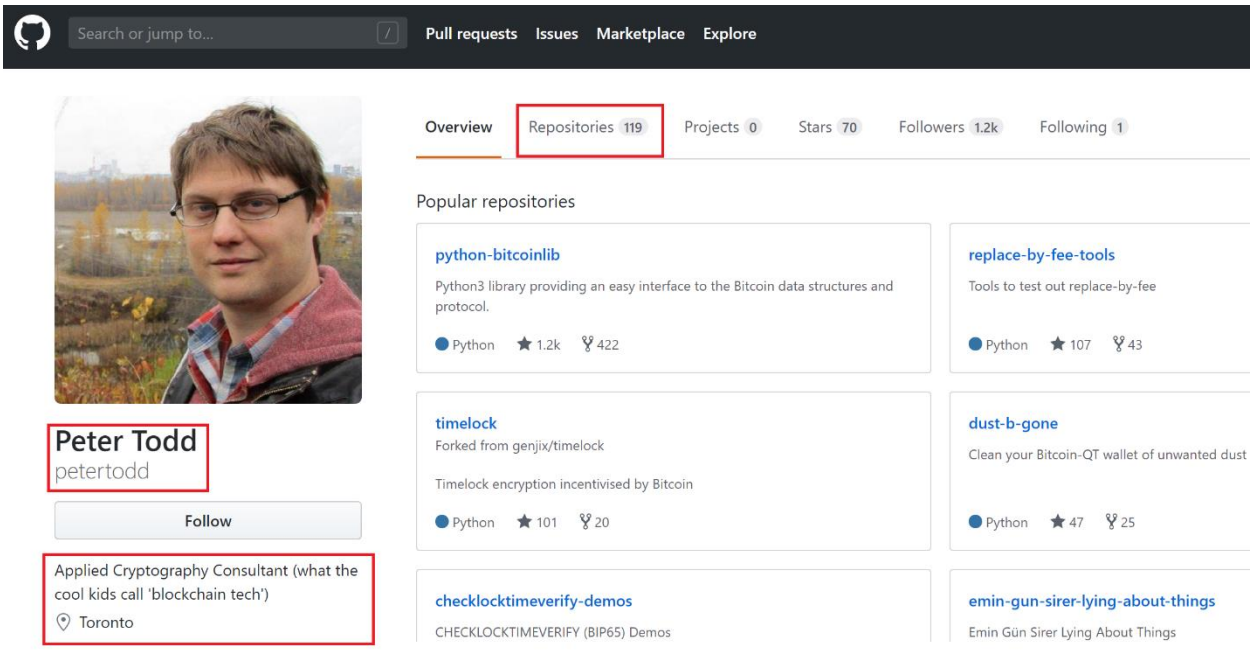

Figure 4.2: Profile of the user

In the next step, we analyse the languages and rank them based on the number of times used in the repositories. The results are as follows:

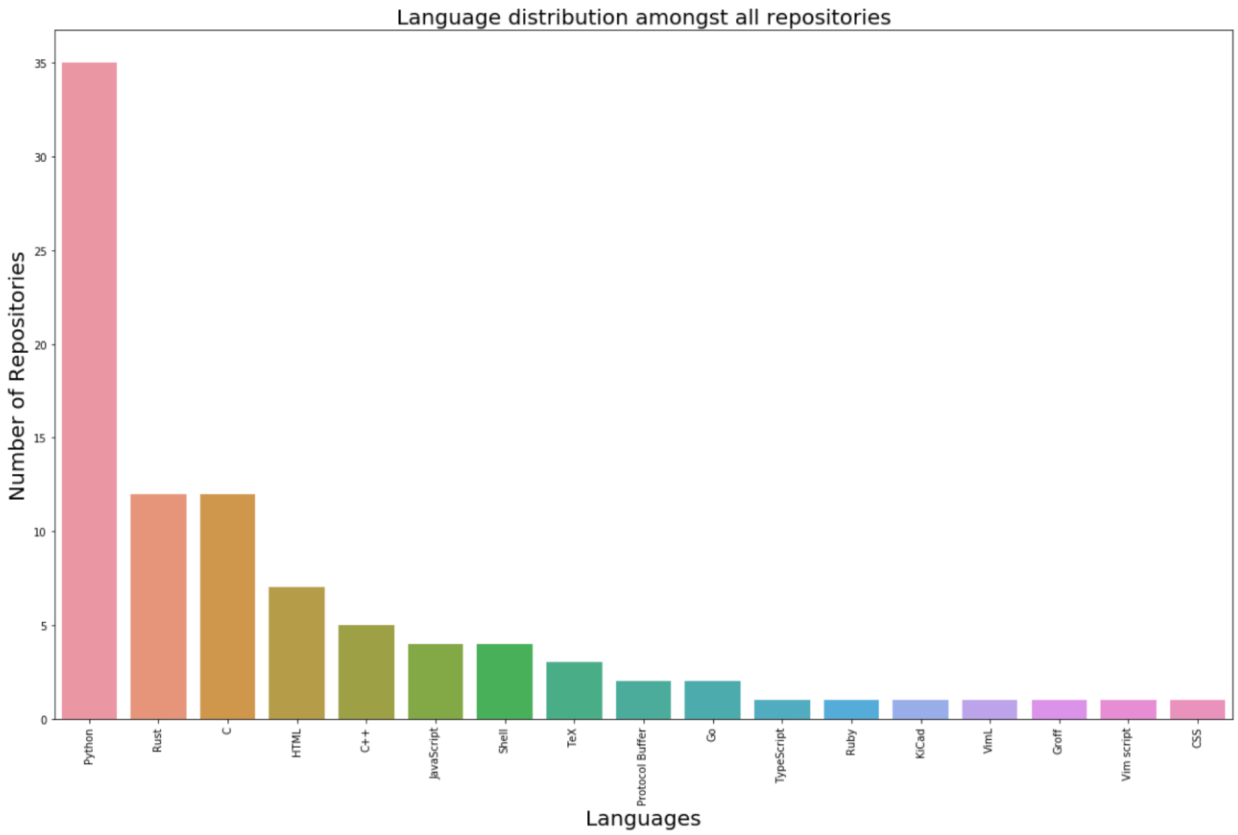

Figure 4.3: Language Analysis

Here in the first part of job recommendation we have the following results from stack overflow results:

```
Do you want to work only in Toronto yes
https://stackoverflow.com/jobs?l=Toronto&d=100&u=Km&tl=Python+Rust+C+HTML+C+++JavaScript+Shell+TeX+ProtocolBuffer+Go+TypeScr
ipt+Ruby+KiCad+VimL+Groff+Vimscript+CSS+&sort=i
Results found: 23 jobs

1 Job title: Java Developer
Company Name:  Critical Mass
Company Location:  Toronto, ON, Canada
Tags:  ['java', 'rest', 'javascript', 'sql', 'css']
Post day:  29d ago
Job ID:  305606
Post Link:  https://stackoverflow.com/jobs/305606/java-developer-critical-mass

2 Job title: Software Development Engineer - AWS Aurora
Company Name:  Amazon
Company Location:  Toronto, ON, Canada
Tags:  ['c++', 'mysql', 'amazon-rds-aurora', 'c']
Post day:  2d ago
Job ID:  296543
Post Link:  https://stackoverflow.com/jobs/296543/software-development-engineer-aws-aurora-amazon

3 Job title: ETL Developer / Data Engineer for a growing SAAS company
Company Name:  Rethink Solutions
Company Location:  Toronto, ON, Canada
Tags:  ['java', 'etl', 'sql', 'talend', 'python']
Post day:  3d ago
Job ID:  316146
Post Link:  https://stackoverflow.com/jobs/316146/etl-developer-data-engineer-for-a-growing-saas-rethink-solutions

4 Job title: Full Stack Developer
Company Name:  Top Hat
Company Location:  Toronto, ON, Canada
Tags:  ['python', 'reactjs', 'javascript', 'flux', 'django']
Post day:  3d ago
Job ID:  149252
Post Link:  https://stackoverflow.com/jobs/149252/full-stack-developer-top-hat
```

Figure 4.4: Job recommendations from Stack Overflow

We have done the validation, by running this evaluation for few of my friends and we found the top-k recommendation where the user is interested or not. We can understand from below line graph, among top 10 we have 7 average recommendations interested from by validation.

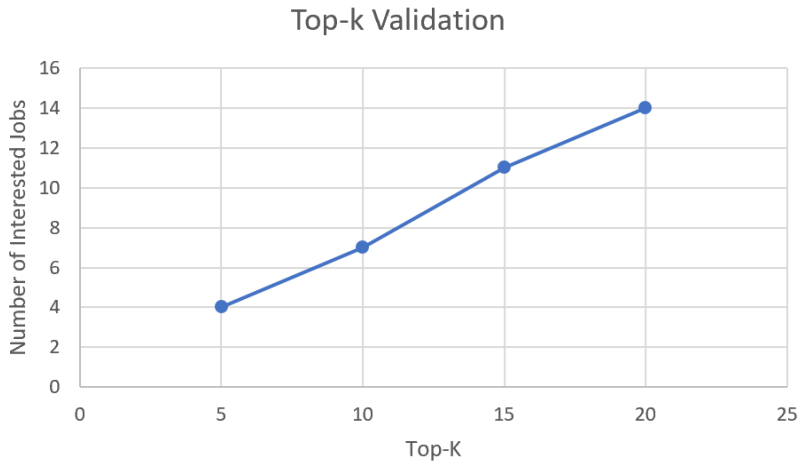

Figure 4.5: Top-k results-1

In the second part of my recommendation, we have got the results from dataset having nearly 5000 entries of data. Based on the ranking the below screen shot is the result:

	Job Title	Skill	Company	Stars	Industry
1	Data Engineer	[Ruby, Python]	Elder Research Inc	NaN	NaN
2	Digital Analytics Analyst	[JavaScript, Python]	PureCars	3.6	NaN
3	Data Analyst II	[JavaScript]	Clayton Homes	3.6	Construction
4	Artificial Intelligence Expert (Data Scientist)	[Python]	BASF	4.2	Industrial Manufacturing
5	Data Analyst (Infection Control)	[CSS]	RWJBarnabas Health	4.0	Health Care
6	Data Scientist/Database Architect	[Python]	ManTech International Corporation	4.0	Aerospace and Defense
7	Associate Data Analyst	[Python]	The Hanover Insurance Group	3.8	Insurance
8	Data Engineer	[Ruby, Python]	Booz Allen Hamilton	3.9	Consulting and Business Services
9	Data Analyst	[Python]	Child Care Resource Center – Chatsworth, CA	NaN	NaN
10	SAP Concur - Sr. Director Product Analytics	[Python]	SAP	4.3	Internet and Software
11	Senior Manager - AI/Data Scientist Pre-Sales A...	[Python]	Avanade	3.7	Consulting and Business Services
12	Senior Data Analyst	[R, Python]	ALC	3.3	NaN
13	Senior Data Scientist	[Machine Learning, Python]	CircleUp	NaN	NaN
14	Data Scientist, Computational Biology	[R, Python]	Camp4 Therapeutics Corporation	NaN	NaN
15	Intelligence Data Analyst	[R, Python]	Intelligent Waves Llc	4.1	Aerospace and Defense

Figure 4.5: Job recommendations form indeed dataset

Again, the validation is the same as the last part. We have asked few of my friends to review the recommendations and asked to number the interested jobs among top 5 to top 20. As we can see a slight improvement in this part of recommendation. As we can understand from the below line graph, that 8 jobs among top 10 are the average interested jobs.

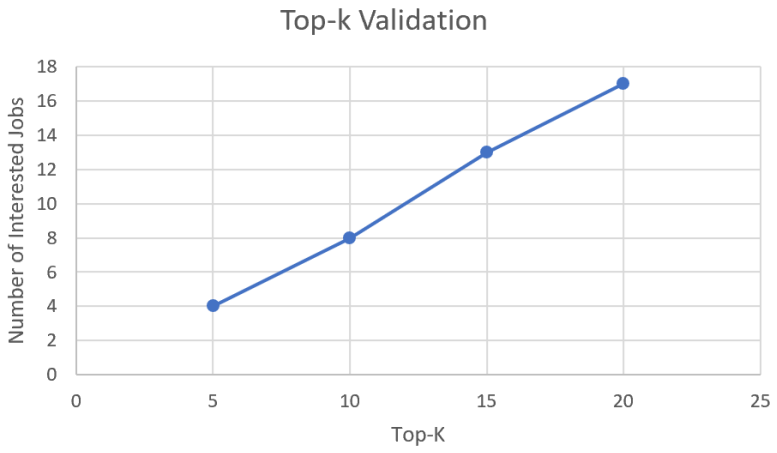

Figure 4.6: Top-k results-2

5 Conclusions and Future work:

From the above results we can conclude that we can obtain user skill analysis. Job recommendation based on the location and language experience from Stack Overflow jobs are obtained. Ranking jobs from the dataset having skill tags are obtained. We can use this method in few applications like user skill analysis, Comparison between two users can be accomplished, Automation of selecting experienced individuals based on GitHub repositories.

As we can see some high top-k results in the second part; after some validation we have found that we have many jobs having only one or two skills required whereas first part having at least four job skills required. So, we have high top-k results in part two than part one.

We also found some improvements to be done for this project to get highly précised recommendations like considering the quality of the project, popularity of the user, considering the commits done by the user. So, we need to work on this for the better results. There is also an improvement in the data to be fetched from stack overflow jobs where we have only around 3000 job postings. So, to improve this we will include indeed jobs as our future work. We also try to improve our validation skills.

References:

1. GEMiner: Mining Social and Programming Behaviors to Identify Experts in GitHub. Wenkai Mo, Beijun Shen, Yuming He, Hao Zhong School of Electronic Information and Electrical Engineering. IEEE, 2011
2. John Anvik and Gail C Murphy. Determining implementation expertise from bug reports. In Mining Software Repositories, 2007. ICSE Workshops MSR'07. Fourth International Workshop. IEEE, 2007.
3. Recommending GitHub Projects for Developer Onboarding. CHAO LIU1,DAN YANG, XIAOHONG ZHANG, BAISHAKHI RAY, MD MASUDUR RAHMAN, IEEE 2018.

4. CORRECT: Code Reviewer Recommendation in GitHub Based on Cross-Project and Technology Experience. Mohammad Masudur Rahman ; Chanchal K. Roy ; Jason A. Collins. IEEE, 2016.
5. Jun Zhang, Mark S Ackerman, and Lada Adamic. Expertise networks in online communities: structure and algorithms. In Proceedings of the 16th international conference on World Wide Web. ACM, 2007.
6. John Anvik and Gail C Murphy. Determining implementation expertise from bug reports. In Mining Software Repositories, 2007. ICSE Workshops MSR'07. Fourth International Workshop on, pages IEEE, 2007.
7. Tag recommendation in software information sites; Xin Xia ; David Lo ; Xinyu Wang ; Bo Zhou, IEEE 2013.