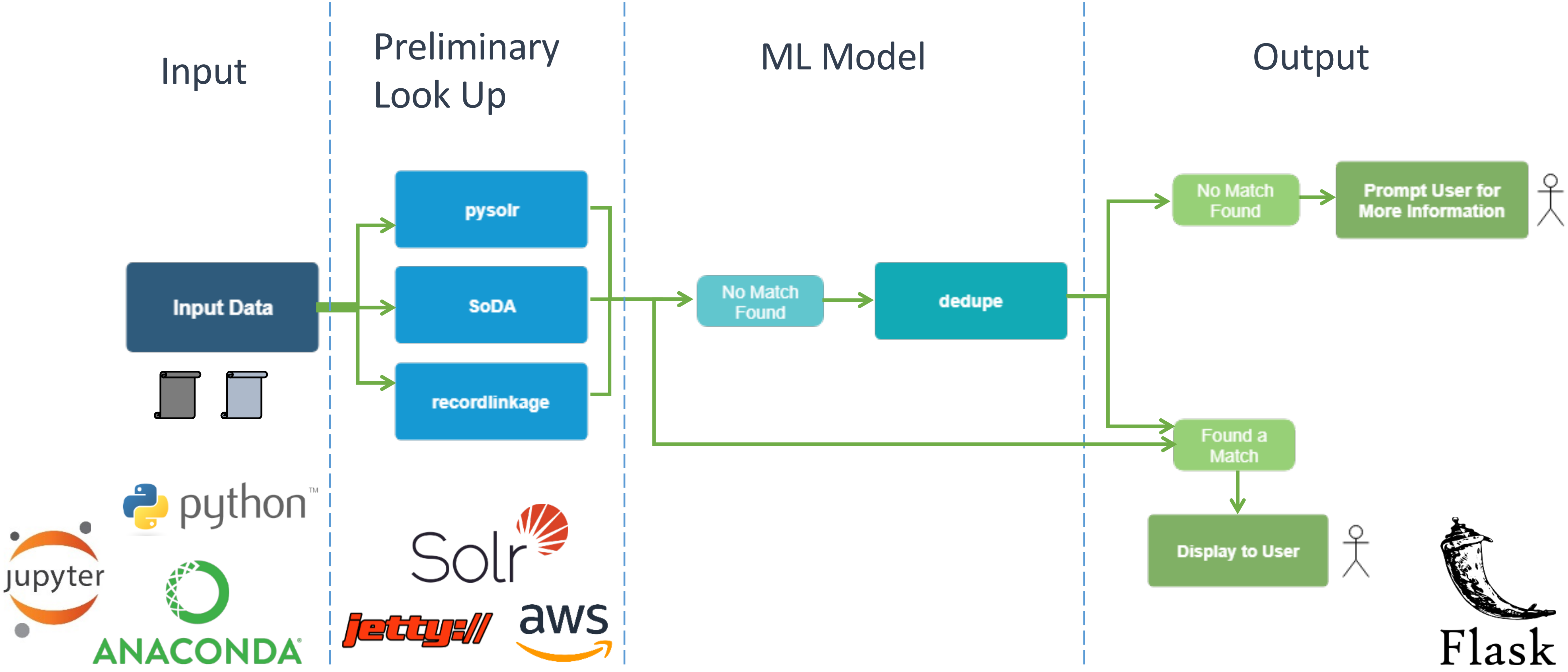
Natural Language Processing

Taylor Kramer and Anne Moshyedi

Entity Resolution Workflow



UI Demo

NLP Platform

Search for...

Logout

NLP Home

Search the Dataset

Tables

NLP Platform / Home

Welcome to your
Natural Language Processing Platform!

SWIFT

NLP PLATFORM

NLP Platform

Logout

NLP Home

Search the Dataset

Tables

Dashboard / Tables

Data Table of Addresses

Show 10 entries

Search:

ID	Company Name	Street Address	City	Country	Postal Code
1	1 MOBILE LIMITED	30 CITY ROAD	LONDON	UK	EC1Y 2AB
2	1 TECH LTD	57 CHARTERHOUSE STREET	LONDON	UK	EC1M 6HA
3	23 SNAPS LIMITED	16 BOWLING GREEN LANE	LONDON	UK	EC1R 0BD
4	2E2 SERVICES LIMITED	200 200 ALDERSGATE ALDERSGATE STREET	LONDON	UK	EC1A 4HD
5	2E2 UK LIMITED	200 ALDERSGATE ALDERSGATE STREET	LONDON	UK	EC1A 4HD
6	40 50 MEDIA LTD	145-157 ST JOHN STREET	LONDON	UK	EC1V 4PW
7	4D DATA CENTRES LIMITED	30 CITY ROAD	LONDON	UK	EC1 2AB

NLP Platform

Search for...

Logout

NLP Home

Search the Dataset

Tables

NLP Platform / Upload Page

Search the Dataset

Company Name

Enter Company Name

Address

Enter Street Address

City

Enter City Name

Country

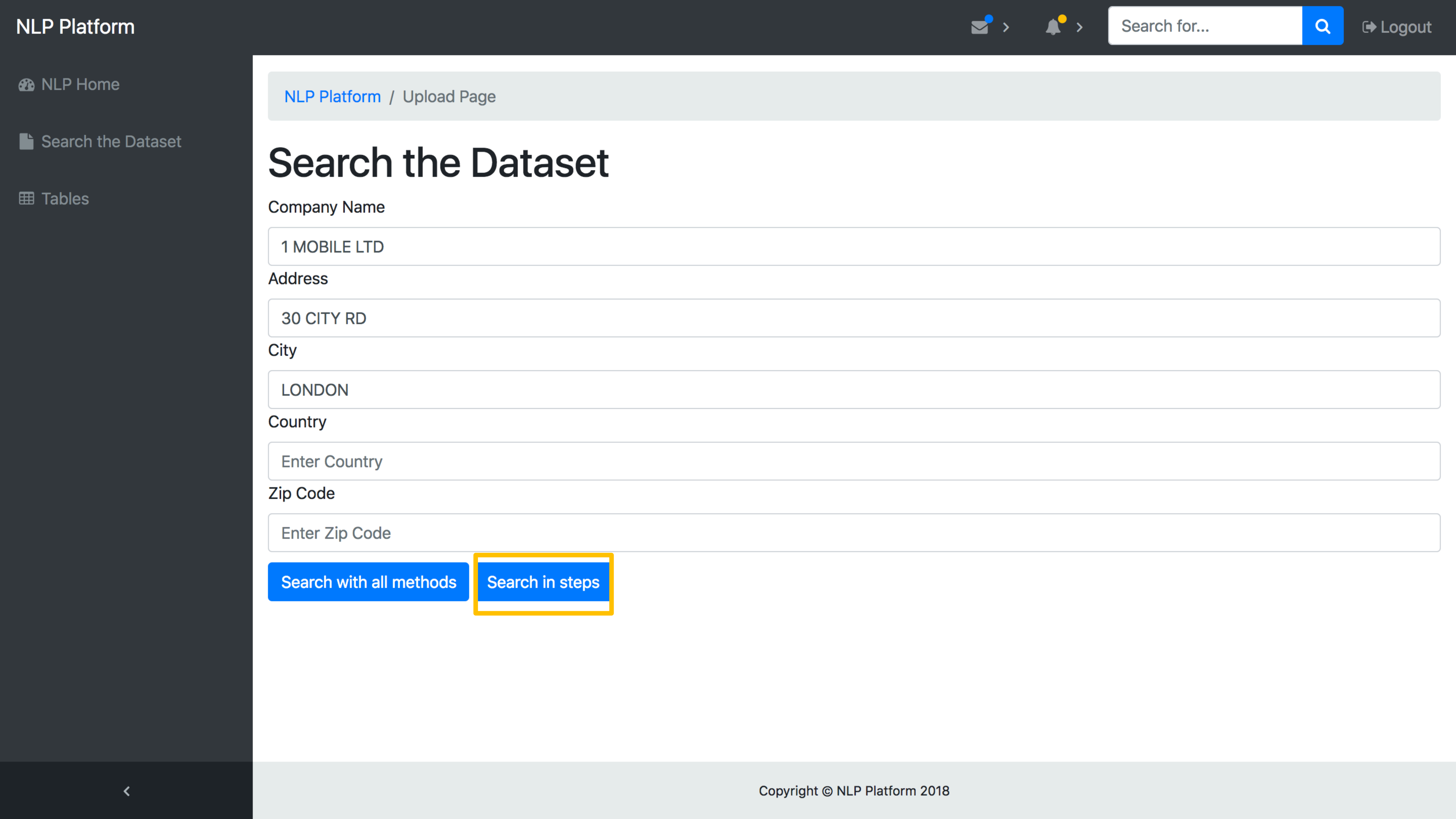
Enter Country

Zip Code

Enter Zip Code

Search with all methods

Search in steps


NLP Platform
Search for...
Logout
NLP Home
Search the Dataset
Tables
NLP Platform / Upload Page
Search the Dataset
Company Name
1 MOBILE LTD
Address
30 CITY RD
City
LONDON
Country
Enter Country
Zip Code
Enter Zip Code
Search with all methods
Search in steps
Copyright © NLP Platform 2018

NLP Home
Search the Dataset
Tables

NLP Platform / Upload Page

Running Pysolr...

Address found with Pysolr:

1 MOBILE LIMITED 30 CITY ROAD LONDON UK EC1Y 2AB

Is this the address you are looking for?

Yes No

NLP Platform

Search for...

Logout

NLP Home

Search the Dataset

Tables

NLP Platform / Upload Page

Running SoDA...

Address found with SoDA:

1 MOBILE LIMITED 30 CITY ROAD LONDON EC1Y 2AB

Is this the address you are looking for?

Yes No

NLP Platform

Search for... Logout

NLP Home

Search the Dataset

Tables

NLP Platform / Upload Page

Running Record Linkage...

Address found with Record Linkage:

No matching address was found!

Is this the address you are looking for?

Yes No

NLP Platform

Search for...

Logout

NLP Home

Search the Dataset

Tables

NLP Platform / Upload Page

Running dedupe...

Address found with dedupe:

1 MOBILE LIMITED 30 CITY ROAD LONDON UK EC1Y 2AB

Is this the address you are looking for?

Yes No

NLP Home
Search the Dataset
Tables

NLP Platform / Upload Page

Success!

Matching Addresses:

The user entry, 1 MOBILE LTD 30 CITY RD LONDON , matched with the dictionary entry, 1 MOBILE LIMITED 30 CITY ROAD LONDON UK EC1Y 2AB, using dedupe.

NLP Platform

Search for...

Logout

- NLP Home
- Search the Dataset
- Tables

NLP Platform / Upload Page

Search the Dataset

Company Name

1 MOBILE LTD

Address

30 CITY RD

City

LONDON

Country

Enter Country

Zip Code

Enter Zip Code

Search with all methods | Search in steps

NLP Platform

Search for...

Logout

NLP Home

Search the Dataset

Tables

NLP Platform / Upload Page

Running all methods...

Address found:

Pysolr: 1 MOBILE LIMITED 30 CITY ROAD LONDON UK EC1Y 2AB Yes

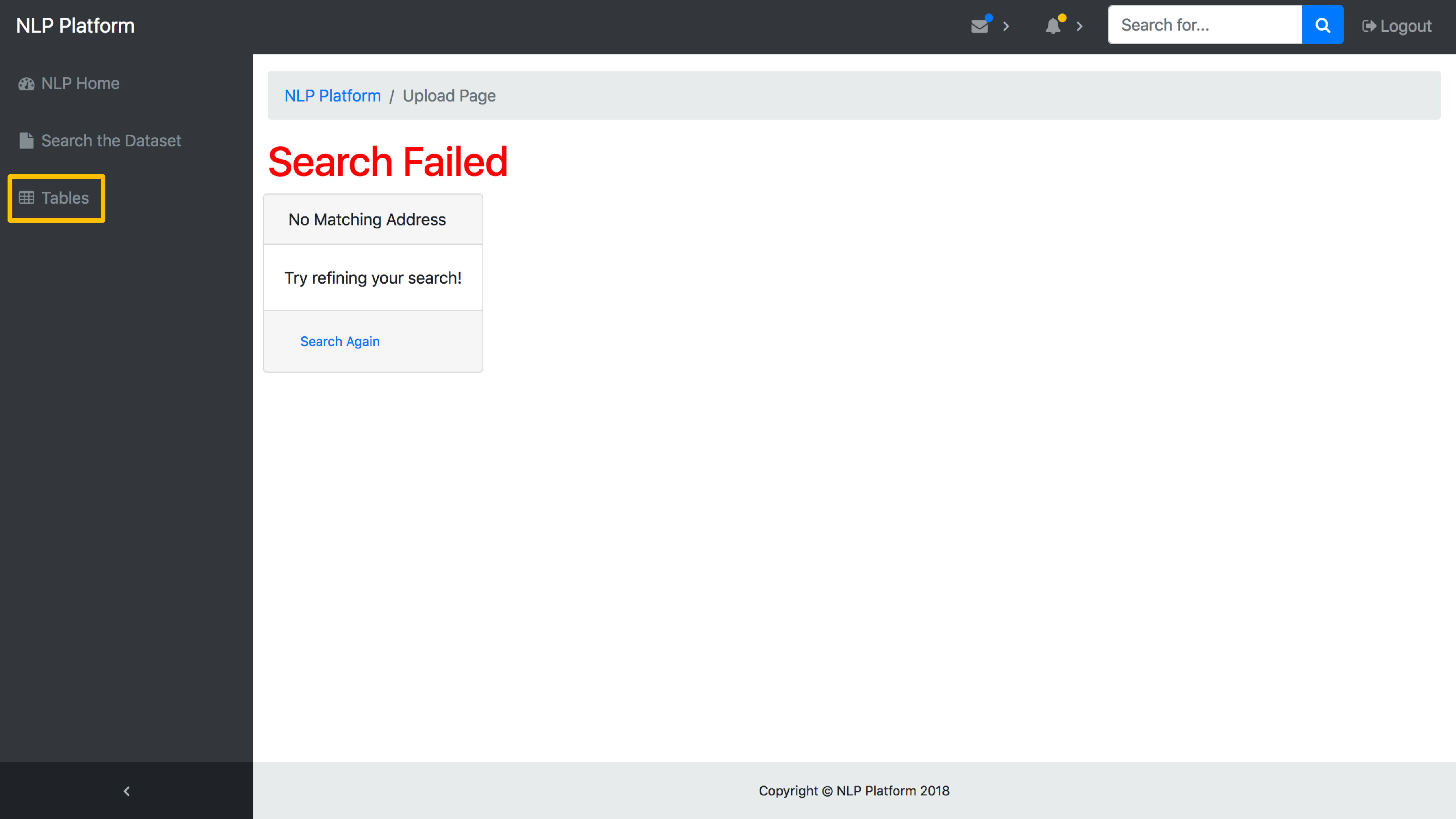
SoDA: 1 MOBILE LIMITED 30 CITY ROAD LONDON EC1Y 2AB Yes

Record Linkage: No matching address was found! Yes

Dedupe: 1 MOBILE LIMITED 30 CITY ROAD LONDON UK EC1Y 2AB Yes

Don't see the address you are looking for?

No


NLP Platform
Search for...
Logout
NLP Home
Search the Dataset
Tables
NLP Platform / Upload Page
Search Failed
No Matching Address
Try refining your search!
Search Again
Copyright © NLP Platform 2018

NLP Platform

Logout

- NLP Home
- Search the Dataset
- Tables

Dashboard / Tables

Data Table of Addresses

Show 10 entries

Search: MOBILE

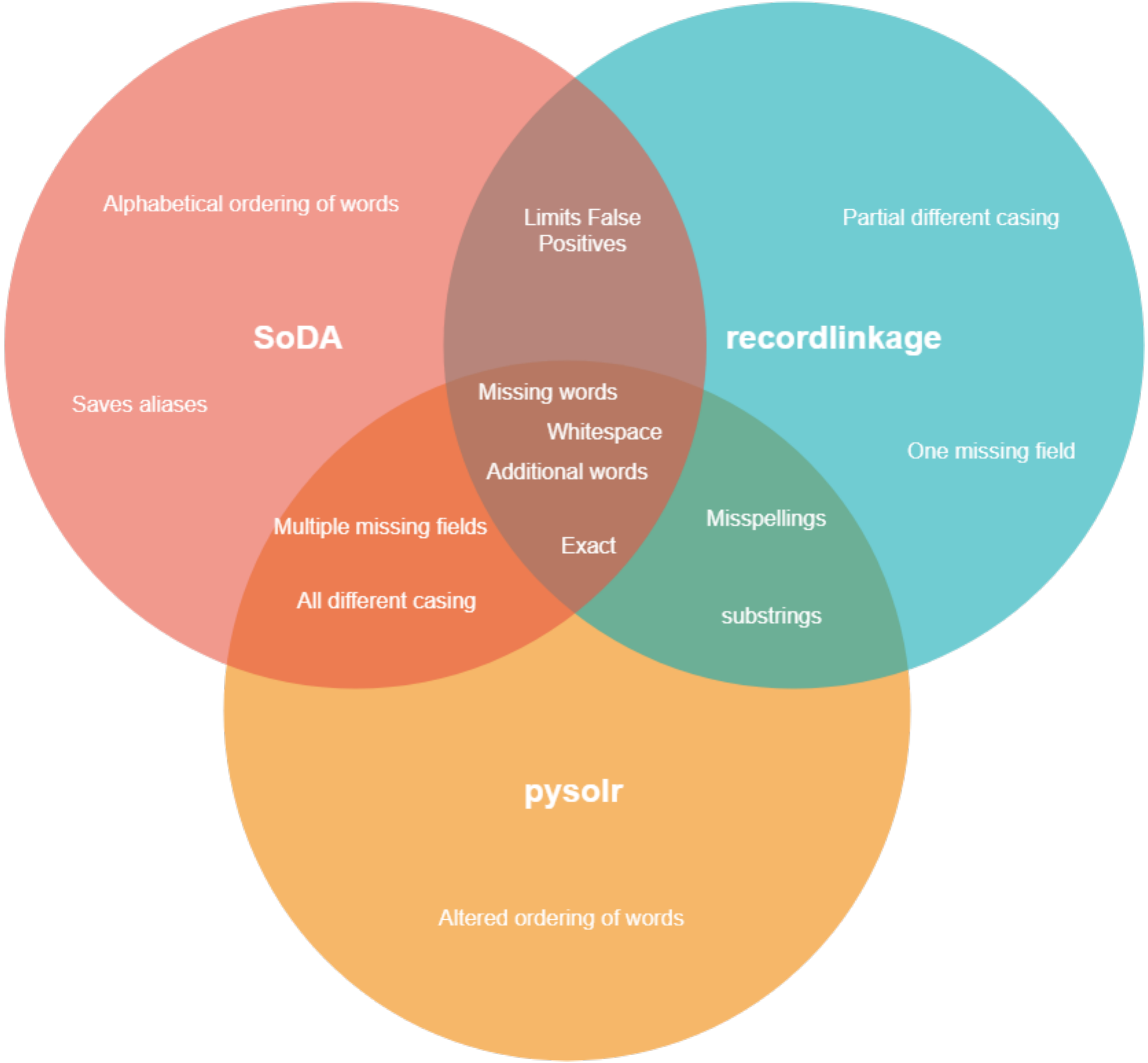
ID	Company Name	Street Address	City	Country	Postal Code
1	1 MOBILE LIMITED	30 CITY ROAD	LONDON	UK	EC1Y 2AB
8	4GETMOBILE LIMITED	152 KEMP HOUSE CITY ROAD	LONDON	UK	EC1V 2NX
116	CASH ON MOBILE LTD	145-157 ST JOHN STREET	LONDON	UK	EC1V 4PW
314	IMRAN BAIG MOBILE SOFTWARE TECHNOLOGIES LIMITED	C/O GOLDER BAQA GROUND FLOOR 1 BAKER'S ROW	LONDON	UK	EC1R 3DB
440	MOBILE TECTONICS LIMITED	KEMP HOUSE 160 CITY ROAD	LONDON	UK	EC1V 2NX

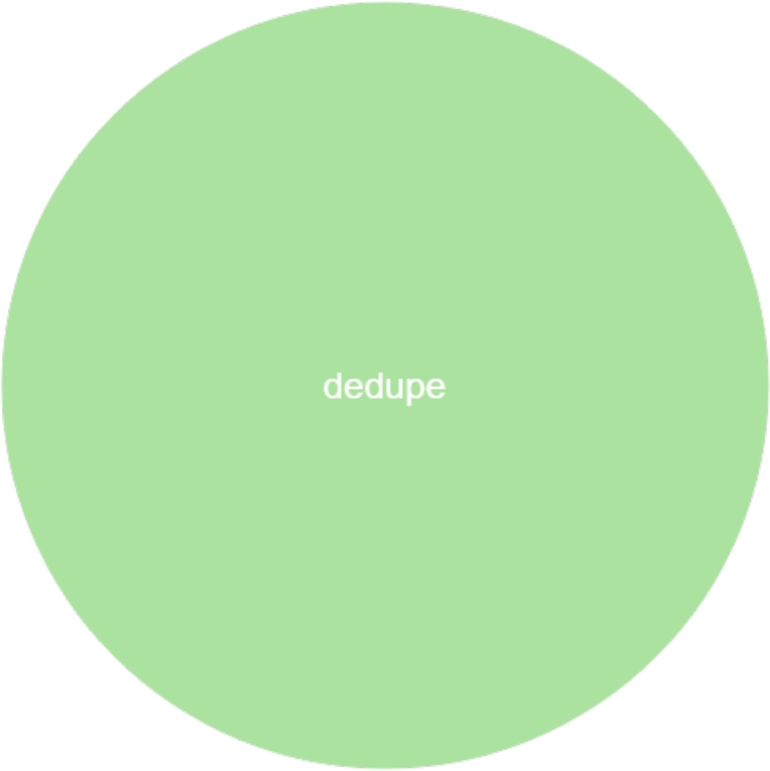
Showing 1 to 5 of 5 entries (filtered from 1,000 total entries)

Previous 1 Next

Matching Methods

Machine Learning





Demo Scenarios

Index-Based Search

User Queries

Exact Search

In [15]:
```
# Results is a list of all the potential matches, the closest match being the first in the list
results = conn.search('name:"1 MOBILE LIMITED" addr:"30 CITY ROAD"')

i = 1
for result in results:
    if i == 1:
        suggestion = (" ".join(result['name'] + result['addr'] + result['city'] + result['ctry'] + result['code']))
    i += 1

print("User entry: 1 MOBILE LIMITED 30 CITY ROAD")
print("Response:", suggestion)
```

```
User entry: 1 MOBILE LIMITED 30 CITY ROAD
Response: 1 MOBILE LIMITED 30 CITY ROAD LONDON UK EC1Y 2AB
```

Insensitive to Capitalization

In [18]:
```
results = conn.search('name:"1 MobILE LimITED" addr:"30 CITY ROAD"')

i = 1
for result in results:
    if i == 1:
        suggestion = (" ".join(result['name'] + result['addr'] + result['city'] + result['ctry'] + result['code']))
    i += 1

print("User entry: 1 MobILE LimITED 30 CITY ROAD")
print("Response:", suggestion)
```

```
User entry: 1 MobILE LimITED 30 CITY ROAD
Response: 1 MOBILE LIMITED 30 CITY ROAD LONDON UK EC1Y 2AB
```

Insensitive to Excessive White Space

In [20]:
```
results = conn.search('name:"1 MOBILE            LIMITED" addr:"30 CITY ROAD"')

i = 1
for result in results:
    if i == 1:
        suggestion = (" ".join(result['name'] + result['addr'] + result['city'] + result['ctry'] + result['code']))
    i += 1

print("User entry: 1 MOBILE            LIMITED 30 CITY ROAD")
print("Response:", suggestion)
```

```
User entry: 1 MOBILE            LIMITED 30 CITY ROAD
Response: 1 MOBILE LIMITED 30 CITY ROAD LONDON UK EC1Y 2AB
```

Substrings

In [25]:
```
results = conn.search('name:"MOBILE" addr:"30 CITY ROAD"')

i = 1
for result in results:
    if i == 1:
        suggestion = (" ".join(result['name'] + result['addr'] + result['city'] + result['ctry'] + result['code']))
    i += 1

print("User entry: MOBILE 30 CITY ROAD")
print("Response:", suggestion)
```

```
User entry: MOBILE 30 CITY ROAD
Response: 1 MOBILE LIMITED 30 CITY ROAD LONDON UK EC1Y 2AB
```

Special Characters

In [24]:
```
results = conn.search('name:"#1 MOBILE LIMITED" addr:"30 CITY ROAD"')

i = 1
for result in results:
    if i == 1:
        suggestion = (" ".join(result['name'] + result['addr'] + result['city'] + result['ctry'] + result['code']))
    i += 1

print("User entry: #1 MOBILE LIMITED 30 CITY ROAD")
print("Response:", suggestion)
```

```
User entry: #1 MOBILE LIMITED 30 CITY ROAD
Response: 1 MOBILE LIMITED 30 CITY ROAD LONDON UK EC1Y 2AB
```

Extra wording, prefixes, suffixes

In [40]:
```
results = conn.search('name:"1 AMOBILE LIMITEDs Company" addr:"30 CITY RD"')

i = 1
for result in results:
    if i == 1:
        suggestion = (" ".join(result['name'] + result['addr'] + result['city'] + result['ctry'] + result['code']))
    i += 1

print("User entry: 1 AMOBILE LIMITED Company 30 CITY ROADS")
print("Response:", suggestion)
```

```
User entry: 1 AMOBILE LIMITED Company 30 CITY ROADS
Response: 1 MOBILE LIMITED 30 CITY ROAD LONDON UK EC1Y 2AB
```

Exact, Unique Addresses

In [45]:
```
results = conn.search('name:"" addr:"COURTYARD SUITE 100 HATTON GARDEN"')

i = 1
for result in results:
    if i == 1:
        suggestion = (" ".join(result['name'] + result['addr'] + result['city'] + result['ctry'] + result['code']))
    i += 1

print("User entry: COURTYARD SUITE 100 HATTON GARDEN")
print("Response:", suggestion)
```

```
User entry: COURTYARD SUITE 100 HATTON GARDEN
Response: ACTURIS LIMITED COURTYARD SUITE 100 HATTON GARDEN LONDON UK EC1N 8NX
```

Dictionary Annotator

User Queries

White Spacing, Abreviations, Missing Info, Capitalization

In [83]:
```
print(main())
```

```
User entry: 1 Mobile      Limited 30 City Rd
Response: 1 MOBILE LIMITED 30 CITY ROAD LONDON EC1Y 2AB
```

Misordered Information

In [84]:
```
print(main2())
```

```
User entry: 1 Limited Mobile 30 City
Response: 1 MOBILE LIMITED 30 CITY ROAD LONDON EC1Y 2AB
```

Special Characters

In [88]:
```
print(main3())
```

```
User entry: #1 Mobile Ltd. 30 City Road
Response: 1 MOBILE LIMITED 30 CITY ROAD LONDON EC1Y 2AB
```

Exact UNIQUE Address

In [106]:
```
print(main4())
```

```
User entry:  COURTYARD SUITE 100 HATTON GARDEN London EC1N 8NX
Response: ACTURIS LIMITED COURTYARD SUITE 100 HATTON GARDEN LONDON EC1N 8NX
```

Information Entered in Incorrect Field

In [104]:
```
print(main5())
```

```
User entry:  1 MOBILE LIMITED 30 CITY ROAD London EC1Y 2AB
Response: 1 MOBILE LIMITED 30 CITY ROAD LONDON EC1Y 2AB
```

AWS Dictionary Annotator Implementation

```
Last login: Thu Aug  2 10:38:40 on ttys002
Annies-MacBook-Pro:~ Annie$ ssh ec2-user@18.204.201.31
ec2-user@18.204.201.31's password:
Welcome to Ubuntu 16.04.4 LTS (GNU/Linux 4.4.0-1060-aws x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

  Get cloud support with Ubuntu Advantage Cloud Guest:
    http://www.ubuntu.com/business/services/cloud

55 packages can be updated.
1 update is a security update.


*** System restart required ***

The programs included with the Ubuntu system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.


The programs included with the Ubuntu system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.

Last login: Thu Aug  2 14:38:54 2018 from 149.134.173.224
ec2-user@ip-172-31-88-92:~$ cd /mnt/ebs
ec2-user@ip-172-31-88-92:/mnt/ebs$ ls
companies_addr1.tsv  companies_city2.tsv  companies_code1.tsv  companies_ctry1.tsv  companies_fa.csv     companies_name3.tsv  project
companies_addr2.tsv  companies_cn.csv     companies_code2.tsv  companies_ctry2.tsv  companies_fa.tsv     example              solr-7.3.0.tgz
companies_city1.tsv  companies_cn.tsv     companies.csv        companies_dict.tsv   companies_name1.tsv  new_companies.csv    SolrTextTagger
ec2-user@ip-172-31-88-92:/mnt/ebs$ cd SolrTextTagger/soda/
ec2-user@ip-172-31-88-92:/mnt/ebs/SolrTextTagger/soda$ ls
build.sbt  docs  LICENSE.md  project  README.md  solr-7.3.0  solr-7.3.0.tgz  src  target
ec2-user@ip-172-31-88-92:/mnt/ebs/SolrTextTagger/soda$
```

Machine Learning Model

User Query Examples

ML model creates the rules for determining a match. They are dependent on training data and are not predetermined.

In [61]:
```
print("User entry:", user_entry)
print("Response:", response_dedupe)
```

```
User entry: MOBILE LIMITED city road 30 London UK EC1Y 2AB
Response: 1 MOBILE LIMITED 30 CITY ROAD LONDON UK EC1Y 2AB
```

In [65]:
```
print("User entry:", user_entry)
print("Response:", response_dedupe)
```

```
User entry: 1 mobile limited 30 city road Lon
Response: 1 MOBILE LIMITED 30 CITY ROAD LONDON UK EC1Y 2AB
```

In [69]:
```
print("User entry:", user_entry)
print("Response:", response_dedupe)
```

```
User entry: 1 mobile limited 30 city road London UNITED KINGDON
Response: 1 MOBILE LIMITED 30 CITY ROAD LONDON UK EC1Y 2AB
```

In [85]:
```
print("User entry:", user_entry)
print("Response:", response_dedupe)
```

```
User entry: 1 MOBILE LTD 300 CITY RD LON UK
Response: 1 MOBILE LIMITED 30 CITY ROAD LONDON UK EC1Y 2AB
```