Introduction

1. Problem. I chose to solve the **Challenge #1 Address Extraction** which implies writing a program that extracts all valid addresses that are found on a list of company websites in the following format: country, region, city, postcode, road and road numbers.

2. Reasoning. First, I decided to use entity recognition to identify addresses in a string containing the contents of a website. Secondly, the reasoning I used is divided in two phases: in the first phase I extract the full address from the string which contains the content of a website and in the second phase I extract the street number, street name, city, region, and ZIP code from a string which contains a full address (obtained in the first phase).

The reason why I decided not to extract the street number, street name, city, region, and ZIP code directly from the website is that it is possible that a street name or city appears in a website in a different context that its address, resulting in the identification of multiple cities in a single webpage, case in which we cannot say which is part of the address.

I trained two models, one for recognizing the address in a string containing the content of a website, and one for recognizing the street number, street name, city, region, and ZIP in an address string.

3. Tools used. I used Python with spaCy (plus some additional tools like beautifulsoup for web scraping, openpyxl to work with .xlsx files) to solve this task.

4. Results. Out of ~2500 websites (which some worked some didn't and some contained addresses, and some didn't) I obtained 509 addresses out of which, 260 are complete (meaning that they have street number, street name, city, region and ZIP) and 249 are incomplete (meaning that either street number or street name or city or region is missing).

Phase 1. Identify the full address.

In this phase, we train a model to identify a full address in a string which contains the content of a website.

Step 1. Perform the webscrape. Input: a csv file with ~100 websites. Output: a list with strings where each string contains the content of a webpage (without html tags).

```python
def get_pages_strings(website_list):
    string_list = []
    for website in website_list:
        txt = scrape_page(website)
        if txt is not None:
            string_list.append(website + "\n\n" + txt)
    return string_list

def get_de_websites(file_path):
    website_list = []
    try:
        with open(file_path, newline='') as csvfile:
            reader = csv.reader(csvfile)
            for row in reader:
                website_list.append(row[0])
        return website_list
    except Exception as e:
        print(f"Error reading CSV file: {e}")
        return None

def scrape_page(url):
    print(f"Scraping {url}")
    try:
        response = requests.get(url)
        html_content = response.content
        soup = BeautifulSoup(html_content, "html.parser")
        text_content = soup.get_text(" ")
        text_content = ' '.join(text_content.split())
        return text_content
    except:
        return None

# Extract the websites from the CSV
website_list = get_de_websites("address_websites.csv")

# Scrape each website
website_strings = get_pages_strings(website_list[])
```
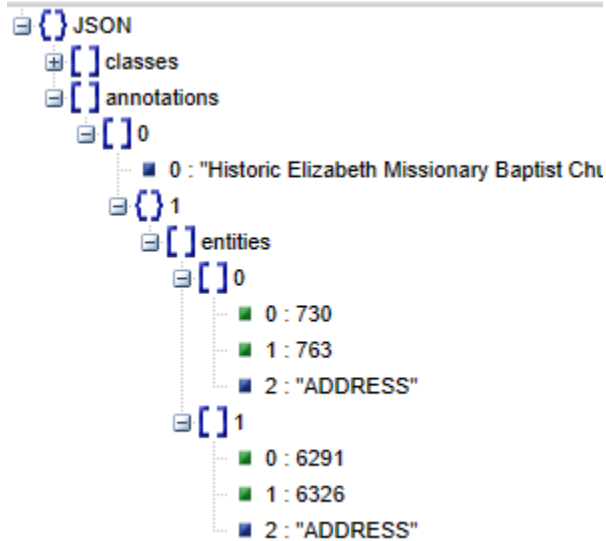
Step 2. Create the dataset. Export the strings obtained from the previous step to .txt files. Then, manually annotate the address using [NER Annotator for SpaCy (tecoholic.github.io)](#) which results in a JSON file where each entity is defined by a start index, end index and label. Then, convert the JSON to the format required by spacy.

Here is an example of JSON exported by the NER annotator:

```
JSON
    classes
    annotations
        0
            0 : "Historic Elizabeth Missionary Baptist Chu
            1
                entities
                    0
                        0 : 730
                        1 : 763
                        2 : "ADDRESS"
                    1
                        0 : 6291
                        1 : 6326
                        2 : "ADDRESS"
```

Below is the dataset I used for training where **Text** contains the content of a website, **Start_Index** represents the index where an address starts and **End_Index** represents the index where an address ends.

	Text	Start_Index	End_Index
1	Text	Start_Index	End_Index
2	The quick brown fox jumps over	31	63
3	Python is a 30 Memorial Drive A	12	41
4	Life is like 456 Elm Street Spring	13	47
5	The cat sat on the 777 Brockton	19	54
6	Today 123 Main Street Anytown	6	37
7	The 123 Main Street Anytown N	4	35
8	I wake up to the smell of fresh c	54	88
9	As I stroll through the park, I cor	104	139
10	I've always been fascinated by t	236	265
11	Sitting by the 789 Oak Avenue Li	15	47
12	I love taking long walks in the pa	146	177
13	After a long day at work, I like to	117	146
14	There's nothing quite like a hon	299	334
15	The sound of laughter echoes th	270	302
16	As I sit by the window, sipping a	293	327
17	The city streets are alive with ac	306	340
18	As the sun sets, the sky is painte	277	312
19	I love spending weekends explc	332	364
20	The hustle and bustle of the city	280	309
21	As the holiday season approach	296	327
22	The aroma of freshly baked cool	247	281
23	I step outside into the crisp mor	259	290
24	The city skyline is illuminated b	261	293
25	I find solace in the pages of a go	218	247
26	The scent of fresh flowers fills t	209	244
27	I take a deep breath and exhale	215	247
28	The laughter of children echoes	209	243
29	I watch in awe as the waves cras	204	233
30	I close my eyes and listen to the	217	252
31	I stand at the summit of the mol	168	199
32	The city streets are alive with th	236	271
33	I step out onto the balcony and	243	272
34	The scent of freshly brewed cof	226	257
35	I lace up my running shoes and I	242	274
36	I spend the afternoon exploring	257	291
37	The gentle rustle of leaves fills t	209	241
38	I sit by the fire, wrapped in a bla	207	242
39	I watch as the sun dips below th	186	215
40	I stroll through the park, admirir	215	249
41	I stand on the beach, watching a	196	227
42	The aroma of freshly brewed co	254	286
43	I lace up my running shoes and I	254	283
44	The soft melody of a piano fills t	268	303
45	I open the door to my favorite b	277	308
46	The sun sets in a blaze of fiery c	232	266
47	I step into the art gallery and an	261	293
48	The city skyline sparkles in the r	197	226
49	I sit by the fireplace, wrapped ir	204	239
50	The scent of freshly baked brea	265	296
51	I step outside into the crisp autu	255	289
52	The gentle sound of rain tappin{	322	353
53	I step onto the sandy beach, fee	326	358
54	The city streets are alive with th	355	384
55	I stand at the edge of the forest	345	380
56	The city skyline stretches out be	365	399
57	I step onto the stage, the spotlig	295	326
57	I step onto the stage, the spotlig	295	326
58	I lace up my hiking boots and se	329	361
59	The city streets are quiet as I ma	340	369
60	I wake up to the sound of birds :	373	408
61	The sun sets in a blaze of colors,	302	336
62	The smell of freshly cut grass fil	266	298
63	I sit by the window, watching th	249	283
64	The city comes alive at night, illi	240	269
65	I step onto the dance floor, the	219	254
66	The scent of fresh flowers fills t	209	240
67	I wake up to the sound of birds :	152	184
68	The city skyline stretches out be	253	287
69	I step onto the sandy shores of t	215	244
70	The streets are alive with the hu	263	298
71	I sit in the quiet solitude of my {	256	287
72	The sound of laughter echoes th	238	270
73	I step into the cozy warmth of m	296	330
74	The city streets are alive with th	254	283
75	I step onto the stage, the spotlig	220	255
76	The sun sets in a blaze of colors,	230	261
77	I wake up to the sound of birds (	341	373
78	The city skyline stretches out be	272	306
79	I step onto the sandy shores of t	387	416
80	The streets are alive with the hu	361	396
81	I sit in the quiet solitude of my {	351	382
82	The crisp morning air fills my lur	201	233
83	I step into the bustling market,	298	332
84	The city skyline looms on the hc	237	266
85	I sit on the porch swing, rocking	251	286
86	The sound of laughter fills the a	260	291
87	I step onto the balcony, greeted	227	259
88	The scent of fresh flowers fills t	260	294
89	I wake up to the sound of birds (	341	370
90	The city streets are alive with th	366	401
91	I sit in the quiet solitude of my {	351	382
92	The gentle lapping of waves aga	348	380
93	The aroma of sizzling barbecue	353	387
94	I step into the cool darkness of t	363	392
95	The city streets are alive with th	404	439
96	I sit on the porch swing, a mug c	324	355
97	The sound of children's laughter	322	354
98	I step onto the dance floor, the	281	315
99	The city skyline stretches out be	311	340
100	I sit by the window, a cup of ste	271	306
101	I step into the garden, surround	328	359
102			

Step 3. Train the model. Train the spacy model using the dataset created previously.

Step 4. Identify addresses. Finally, we perform the web scraping upon the entire input (~2500 websites of which, some don't work and some don't have an address) to obtain a string with its content. Then, we pass these strings as input to the trained model to identify addresses. By inspection, I noticed errors which contained long strings. Also, I noticed that the last word in most addresses contains both a digit and a letter. Therefore, I cleaned the results by removing long strings and strings where the last word doesn't contain both a digit and a letter.

```python
def predict_addresses():
    wb = openpyxl.Workbook()
    ws = wb.active

    model = spacy.load('PhaseOne/model/model-best')
    with open('websites_list.csv', mode='r', newline='', encoding='utf-8') as
file:
        csv_reader = csv.reader(file)
        #row_count = 0
        excel_row = 1
        flag = False
        skip_flag = False
        for index, row in enumerate(reversed(list(csv_reader)[:10])):
            website_url = f"https://{row[0]}"
            txt = scrape_page(website_url)
            if txt is not None:
                doc = model(txt)
                for ent in doc.ents:
                    ws.cell(row=excel_row, column=1).value = website_url
                    ws.cell(row=excel_row, column=2).value =
sanitize(ent.text)
                    excel_row = excel_row + 1
                    print("Excel row:", excel_row, "Website:", website_url,
"Address:", ent.text, ent.label_)

        wb.save('prediction_result_test.xlsx')

def clean_full_address():
    def truncate_string(input_string, max_length):
        if len(input_string) > max_length:
            return input_string[:max_length]
        return input_string

    # Load the Excel workbook
    wb = openpyxl.load_workbook('prediction_result_test.xlsx')
    sheet = wb.active

    # Iterate over rows in column B and truncate strings
    for row in sheet.iter_rows(min_row=2, min_col=2, max_col=2):
        cell_value = row[0].value
        if cell_value:
            row[0].value = truncate_string(cell_value, 93)

    # Iterate through rows and remove those where the last word in column B
```

```
doesn't contain a digit
    rows_to_delete = []
    for row in sheet.iter_rows(min_row=2, max_col=2, max_row=sheet.max_row):
        cell_value = row[1].value
        if cell_value:
            last_word = cell_value.split()[-1]
            if not any(char.isdigit() for char in last_word):
                rows_to_delete.append(row)

    # Delete rows from the worksheet
    for row in rows_to_delete:
        sheet.delete_rows(row[0].row)

    # Save the modified workbook
    wb.save('prediction_result_test_clean.xlsx')
```

After cleaning, I obtained 508 addresses. Here are the first 57 and the last 54 addresses.

#	URL	Address
1	https://sgrcs.co.uk	62 Stagshaw Close Maidstone, Kent ME15 6TN
2	https://henshaw.uk.com	Russell Road, Edinburgh EH11 2LS
3	https://exchequersolutions.co.uk	1 St John Street, Chester, CH1 1DA
4	https://exchequersolutions.co.uk	1 St John Street, Chester, CH1 1DA
5	https://burrowfieldautos.com	2 Woodfield Road, Welwyn Garden City, England, AL7 1J
6	https://vcut.co.uk	8 Dundas Close Portsmouth PO3 5RB
7	https://vcut.co.uk	75 Turnmill Street London EC1M 5SY
8	https://missionlimited.com	2 Upper Teddington Road, KingstonUponThames, Surrey
9	https://lodestar.org.uk	LodeStar Harvey House, Lode, Cambridge, CB25 9HB
10	https://quadranet.co.uk	High Wycombe Bucks, HP14 3UH
11	https://cornerfridge.com	4 Brunel Close, Harworth, Doncaster, South Yorkshire, D
12	https://porterfieldbrakes.co.uk	16B Norman Way, Severn Bridge Ind Est, Caldicot, NP26
13	https://clarkfoysterwines.co.uk	15 South Ealing Rd, London, W5 4QT
14	https://djrmarketing.co.uk	5 Shaw Park Business Village Shaw Road Wolverhampto
15	https://flackleyashhotel.co.uk	843 Finchley Road, London. NW11 8NA
16	https://coralbay.tv	99 Mabgate Road Leeds LS9 7DR
17	https://royalegraphics.co.uk	9 Archer Road, Stapleford, Nottingham, Nottinghamshir
18	https://maritimedevelopments.con	7 Wilson Street, Peterhead, AB42 1UD
19	https://browntroutinn.com	209, Agivey Road Aghadowey near Coleraine Co. Londor
20	https://cardinalpointstrategies.com	Road Watford Hertfordshire WD17 2JP
21	https://trlibdems.org.uk	1 Vincent Square, London SW1P 2PN.
22	https://schooladvice.co.uk	11 Maisies Way South Normanton Derbyshire, DE55 2DS
23	https://cloud2.co.uk	331 Styal Rd Manchester M22 5LW
24	https://cloud2.co.uk	7 Savannah Way Leeds LS10 1AB
25	https://cloud2.co.uk	Runcorn Cheshire WA7 4QX
26	https://cablemonkey.co.uk	500 Avenue West, Braintree, CM77 7AA
27	https://inoapps.com	6 Queens Road, Aberdeen, AB15 4ZT
28	https://inoapps.com	6 Queens Road, Aberdeen, AB15 4ZT
29	https://letterboxdistribution.com	65, Lancaster Road Industrial Estate, Lancaster Road, Bar
30	https://bossdriving.co.uk	66 King Edward Avenue Rainham, RM13 9RJ
31	https://jbrcapital.com	773 Finchley Road, London, NW11 8DN
32	https://bluemonkeyweb.com	14 St Mary's Street Whittlesey, Peterborough, PE7 1BH
33	https://depinna.com	20 Birchin Lane London EC3V 9DU
34	https://depinna.com	20 Birchin Lane London EC3V 9DU
35	https://depinna.com	20 Birchin Lane London EC3V 9DU
36	https://priorandwilliisantiques.co.ul	47 Salts Road West Walton Wisbech PE14 7EJ
37	https://cerberusnetworks.co.uk	9 Walmgate Road, Greenford, Middlesex, UB6 7LH
38	https://mrchippyyork.co.uk	Church Street York YO1 8BA
39	https://lenzieunion.org	65 Kirkintilloch Rd Kirkintilloch Lenzie G66 4LD
40	https://pollock.co.uk	10 Blackburn Road, Bathgate, EH48 2EY
41	https://pollock.co.uk	10 Blackburn Road, Bathgate, EH48 2EY
42	https://dukesbrewandque.com	33 Downham Road Haggerston London N1 5AA
43	https://baks.co.uk	30 Seamoor Rd, Westbourne, Bournemouth BH4 9AS
44	https://marlinservices.co.uk	26 Sarum Complex, Uxbridge, Middlesex, UB8 2RZ
45	https://raxel.co.uk	4 Ruston Road, Alma Park Industrial Estate, Grantham, N
46	https://southamptoncruisecentre.c	298 Highfield Road Blackpool Lancashire FY4 3JU
47	https://southamptoncruisecentre.c	62 Wolverhampton Street, Walsall WS2 8DD.
48	https://hyalofemme.co.uk	15079, Dunblane, FK15 5BA
49	https://albaamicorum.com	12 Kinnerton Yard Belgravia London SW1X 8EB
50	https://thesummitclinic.co.uk	40 Highgate West Hill, London N6 6LS
51	https://thesummitclinic.co.uk	206 Upper Richmond Road West, East Sheen, London SW
52	https://thesummitclinic.co.uk	40 Highgate West Hill, London N6 6LS
53	https://coxhire.co.uk	Halesowen Road, Netherton, West Midlands, DY2 9NW
54	https://etc-inter.net	21 Church Road, Parkstone, Poole, BH14 8UF.
55	https://victoriafalls-zambiatravel-sa	6081 Hamilton Blvd, Suite 600 Allentown, PA 18106
56	https://foamtechniques.co.uk	39 Booth Drive Park Farm Industrial Estate South Wellin
57	https://pupilrewardpoints.co.uk	11 Waldegrave Way, Lawford, Manningtree, Essex, CO1

#	URL	Address
454	https://royersford.com	190 Royersford, PA 19468
455	https://tmgi.net	46169 Westlake Drive, Suite 240 Sterling, VA 20165
456	https://devact.com	438 Queens Road Charlotte, NC 28207
457	https://gipreit.com	401 E Jackson St Ste 3300 Tampa, FL 33602
458	https://chcseia.com	400 North 17th Street Keokuk, IA 52632
459	https://chcseia.com	1618 Main Street Keokuk, IA 52632
460	https://chcseia.com	2409 Spring Street Columbus City, IA 52737
461	https://chcseia.com	1706 West Agency Road West Burlington, IA 52655
462	https://chcseia.com	400 S. Broadway Street Burlington, Iowa 52601
463	https://chcseia.com	1706 West Agency Road West Burlington, IA 52655
464	https://enidmad.org	129 N. University Ave. Enid, OK 73701
465	https://sellersconstruction.net	115 Milledgeville Hwy Gordon, GA 31031
466	https://goldridgecompanies.com	2713 North Clairemont Avenue Eau Claire, WI 54703
467	https://mid-america.com	2500 Broadway St. NE, Minneapolis, MN 55413
468	https://mid-america.com	2500 Broadway St NE, Minneapolis, MN 55413
469	https://fresnogreekfest.com	2219 N. Orchard St Fresno, CA 93703
470	https://aaronsantosdentistry.com	14529 South Bascom Ave, Los Gatos, CA 95032
471	https://thegrindcoffeebar.com	9532 West Linebaugh Ave, Tampa, FL 33626
472	https://thegrindcoffeebar.com	9532 West Linebaugh Avenue, Tampa, FL, USA 813
473	https://habeggercorp.com	11413 Enterprise Park Drive Sharonville, OH 45241
474	https://billproctorandassociates.cor	2050 South Blvd. Unit #412 Bloomfield Hills, MI 48303
475	https://achieveservices.net	1000 Parkwood Circle SE STE 900 Atlanta, GA 30339
476	https://millerindustrial.com	621 E. Devon Ave, Elk Grove Village, IL 60007
477	https://appwinit.com	1239 Broadway. New York, NY 10001
478	https://hscpoly.com	12334 Geddes Road PO Box 80 Hemlock, Michigan 4862
479	https://hennejewelers.com	5501 Walnut Street Pittsburgh, PA 15232
480	https://roensalvage.com	180 E Redwood St Sturgeon Bay, WI 54235
481	https://unitedairconditioning.com	13150 S Belcher Rd Largo, FL 33773
482	https://panniergraphics.com	345 Oak Road, Gibsonia, PA 15044
483	https://panniergraphics.com	345 Oak Road Gibsonia, PA 15044
484	https://allenvillagejewelers.com	190 E Stacy Rd #206, Allen, TX 75002
485	https://artofthelakes.org	244, Battle Lake, MN 56515
486	https://fcllp.com	570 Lexington Avenue New York, NY 10022
487	https://purcellvilledental.com	609 East Main Street, Suite T Purcellville, VA 20132
488	https://jackiescleaning.com	116 Crowley Rd, Lewiston, ME 04240
489	https://kiddsmith.com	117 Riviera Drive Boiling Springs , SC 29316
490	https://kiddsmith.com	117 Riviera Drive Boiling Springs , SC 29316
491	https://atlanticbeach-nc.com	125 West Fort Macon Rd Atlantic Beach, NC 28512
492	https://stmatthewmedina.org	400 N. Broadway St., Medina, OH 44256
493	https://seniorcareagencyny.com	3601 Hempstead Tpke, Suite 203 Levittown, NY 11756
494	https://your-doctor.org	5 Brayford Square London, E1 0SG
495	https://glacier.chat	302 Sentinel Drive Annapolis Junction, MD 20701
496	https://neutresalon.com	101 S Market Street Frederick, MD 21701
497	https://neutresalon.com	101 S Market Street Frederick, MD 21701
498	https://paramountlandtitle.com	8904 Sony Lane, Knoxville, TN 37923
499	https://richardbrothersfinancial.cor	50 Donald B. Dean Drive, Suite 1 South Portland, Main
500	https://searchouthomes.com	1739 S. Jade Way, Suite 100 Meridian, Idaho 83642
501	https://triplecrownconstruction.cor	1525 Tilco Drive, Unit B1, Frederick, MD 21704
502	https://wyandottewinery.com	4640 Wyandotte Dr Columbus, Ohio 43230
503	https://societyfortheblind.org	1238 S Street Sacramento, CA 95811
504	https://dillonmusic.com	325 Fulton Street Woodbridge, NJ 07095
505	https://cabwhp.org	9800 S. La Cienega Blvd., Suite 905 Inglewood, CA 9030
506	https://sk4designs.com	3411 TELLURIDE, CO 81435
507	https://embcmonroe.org	503 Maurice St. Monroe, NC 28112.
508	https://embcmonroe.org	503 Maurice Street Monroe, NC 28112
509		

Phase 2. Identify street number, street name, city, region and ZIP

In this phase, we use as input the addresses collected in the first phase and we train a model to identify the street number, street name, city, region and ZIP.

Step 1. Create the dataset. Out of 508 addresses collected in the previous phase, I used 30 UK addresses and 30 USA addresses. In these 60 addresses, I manually annotated the street number, street name, city, region and ZIP code.

```python
# Create spacy dataset
def create_train_data_spacy_address():
    nlp = spacy.blank('en')
    db = DocBin()
    for index in range(26):
        data =
json.load(open(f'./PhaseTwo/phase_two_train_data/uk_jsons/annotations
({index}).json'))
        text = data['annotations'][0][0]
        entities = data['annotations'][0][1]['entities']
        doc = nlp.make_doc(text)
        ents = []
        for start, end, label in entities:
            print(start, end, label)
            try:
                span = doc.char_span(start, end, label=label,
alignment_mode='strict')
            except:
                continue
            if span is not None:
                ents.append(span)
        try:
            doc.ents = ents
            db.add(doc)
        except:
            pass

    data =
json.load(open(f'./PhaseTwo/phase_two_train_data/usa_annotations.json'))
    for annotation in data['annotations']:
        if annotation is not None:
            text = annotation[0]
            entities = annotation[1]['entities']
            doc = nlp.make_doc(text)
            ents = []
            for start, end, label in entities:
                print(start, end, label)
                try:
                    span = doc.char_span(start, end, label=label,
alignment_mode='strict')
                except:
```

```
              continue
          if span is not None:
              ents.append(span)
      try:
          doc.ents = ents
          db.add(doc)
      except:
          pass
db.to_disk('./PhaseTwo/phase_two_train_data/train_address_data.spacy')
```

Step 2. Train the model. Then I trained a model using the 60 addresses annotated to identify the street number, street name, city, region and ZIP code.

Step 3. Identify street number, street name, city, region and ZIP. Then, I used the model trained in the previous step to identify the street number, name, city, region and ZIP in the 508 addresses collected in the previous phase.

```
# Predict address components
def predict_address_components():
    data = []
    wb = openpyxl.load_workbook('./PhaseOne/prediction_result_cleaned.xlsx')
    ws = wb.active
    for row in ws.iter_rows(values_only=True):
        website, text = row[:2]
        data.append([website, text])

    model = spacy.load('./PhaseTwo/model-best')
    result = []
    for row in data:
        doc = model(row[1])
        ents = []
        for ent in doc.ents:
            ents.append([ent.text, ent.label_])
        result.append([row[0], doc.text, ents])
    return result
```

In the result I obtained, I noticed that around 40% of addresses are not complete. Most of them are UK addresses.

Here are the first 57 UK addresses.

	Website	Full Address	Street N	Street Name	City	Region	ZIP
1	Website	Full Address	Street N	Street Name	City	Region	ZIP
2	https://sgrcs.co.uk	62 Stagshaw Close Maidstone, Kent ME15 6TN	62	Stagshaw Close	Maidstone	Kent	ME15 6TN
3	https://henshaw.uk.com	Russell Road, Edinburgh EH11 2LS		Russell Road	Edinburgh		EH11 2LS
4	https://exchequersolutions.co.uk	1 St John Street, Chester, CH1 1DA	1	St John Street	Chester		CH1 1DA
5	https://exchequersolutions.co.uk	1 St John Street, Chester, CH1 1DA	1	St John Street	Chester		CH1 1DA
6	https://burrowfieldautos.com	2 Woodfield Road, Welwyn Garden City, England, AL7 1JQ	2	Woodfield Road	Welwyn Garden City	England	AL7 1JQ
7	https://vcut.co.uk	8 Dundas Close Portsmouth PO3 5RB	8	Dundas Close	Portsmouth		PO3 5RB
8	https://vcut.co.uk	75 Turnmill Street London EC1M 5SY	75	Turnmill Street	London		EC1M 5SY
9	https://missionlimited.com	2 Upper Teddington Road, KingstonUponThames, Surrey. I	2	Upper Teddington Road	KingstonUponTham	Surrey.	KT1 4DY
10	https://lodestar.org.uk	LodeStar Harvey House, Lode, Cambridge, CB25 9HB		LodeStar Harvey House	Lode	Cambridge	CB25 9HB
11	https://quadranet.co.uk	High Wycombe Bucks, HP14 3UH		High Wycombe Bucks			HP14 3UH
12	https://cornerfridge.com	4 Brunel Close, Harworth, Doncaster, South Yorkshire, DN	4	Brunel Close, Harworth	Doncaster	South Yorkshir	DN11 8QA
13	https://porterfieldbrakes.co.uk	16B Norman Way, Severn Bridge Ind Est, Caldicot, NP26 5P	16B	Norman Way	Severn Bridge Ind Es	Caldicot	NP26 5PT
14	https://clarkfoysterwines.co.uk	15 South Ealing Rd, London, W5 4QT	15	South Ealing Rd	London		W5 4QT
15	https://djrmarketing.co.uk	5 Shaw Park Business Village Shaw Road Wolverhampton	5	Shaw Park Business Village	Wolverhampton West Midlands Ur		WV10 9LE
16	https://flackleyashhotel.co.uk	843 Finchley Road, London. NW11 8NA	843	Finchley Road	London.		NW11 8NA
17	https://coralbay.tv	99 Mabgate Road Leeds LS9 7DR	99	Mabgate Road	Leeds		LS9 7DR
18	https://royalegraphics.co.uk	9 Archer Road, Stapleford, Nottingham, Nottinghamshire,	9	Archer Road, Stapleford	Nottingham	Nottinghamsh	NG9 7EP
19	https://maritimedevelopments.com	7 Wilson Street, Peterhead, AB42 1UD	7	Wilson Street	Peterhead		AB42 1UD
20	https://browntroutinn.com	209, Agivey Road Aghadowey near Coleraine Co. Londond	209	Agivey Road	Londonderry		BT51 4AD
21	https://cardinalpointstrategies.com	Road Watford Hertfordshire WD17 2JP		Road Watford	Hertfordshire		WD17 2JP
22	https://trlibdems.org.uk	1 Vincent Square, London SW1P 2PN.	1	Vincent Square	London		SW1P 2PN.
23	https://schooladvice.co.uk	11 Maisies Way South Normanton Derbyshire, DE55 2DS	11	Maisies Way South	Normanton	Derbyshire	DE55 2DS
24	https://cloud2.co.uk	331 Styal Rd Manchester M22 5LW	331	Styal Rd	Manchester		M22 5LW
25	https://cloud2.co.uk	7 Savannah Way Leeds LS10 1AB	7	Savannah Way	Leeds		LS10 1AB
26	https://cloud2.co.uk	Runcorn Cheshire WA7 4QX			Runcorn	Cheshire	WA7 4QX
27	https://cablemonkey.co.uk	500 Avenue West, Braintree, CM77 7AA	500	Avenue West	Braintree		CM77 7AA
28	https://inoapps.com	6 Queens Road, Aberdeen, AB15 4ZT	6	Queens Road	Aberdeen		AB15 4ZT
29	https://inoapps.com	6 Queens Road, Aberdeen, AB15 4ZT	6	Queens Road	Aberdeen		AB15 4ZT
30	https://letterboxdistribution.com	65, Lancaster Road Industrial Estate, Lancaster Road, Barn	65	Lancaster Road Industrial Es	Barnet	Hertfordshire.	EN4 8AS
31	https://bossdriving.co.uk	66 King Edward Avenue Rainham, RM13 9RJ	66	King Edward Avenue	Rainham		RM13 9RJ
32	https://jbrcapital.com	773 Finchley Road, London, NW11 8DN	773	Finchley Road	London		NW11 8DN
33	https://bluemonkeyweb.com	14 St Mary's Street Whittlesey, Peterborough, PE7 1BH	14	St Mary's Street	Whittlesey	Peterborough	PE7 1BH
34	https://depinna.com	20 Birchin Lane London EC3V 9DU	20	Birchin Lane London EC3V 9DU			
35	https://depinna.com	20 Birchin Lane London EC3V 9DU	20	Birchin Lane London EC3V 9DU			
36	https://depinna.com	20 Birchin Lane London EC3V 9DU	20	Birchin Lane London EC3V 9DU			
37	https://priorandwillisantiques.co.uk	47 Salts Road West Walton Wisbech PE14 7EJ	47	Salts Road West	Wisbech		PE14 7EJ
38	https://cerberusnetworks.co.uk	9 Walmgate Road, Greenford, Middlesex, UB6 7LH	9	Walmgate Road	Greenford	Middlesex	UB6 7LH
39	https://mrchippyyork.co.uk	Church Street York YO1 8BA		Church Street	York		YO1 8BA
40	https://lenzieunion.org	65 Kirkintilloch Rd Kirkintilloch Lenzie G66 4LD	65	Kirkintilloch Rd	Kirkintilloch Lenzie		G66 4LD
41	https://pollock.co.uk	10 Blackburn Road, Bathgate, EH48 2EY	10	Blackburn Road	Bathgate		EH48 2EY
42	https://pollock.co.uk	10 Blackburn Road, Bathgate, EH48 2EY	10	Blackburn Road	Bathgate		EH48 2EY
43	https://dukesbrewandque.com	33 Downham Road Haggerston London N1 5AA	33	Downham Road Haggerston London N1 5AA			
44	https://baks.co.uk	30 Seamoor Rd, Westbourne, Bournemouth BH4 9AS	30	Seamoor Rd	Westbourne	Bournemouth	BH4 9AS
45	https://marlinservices.co.uk	26 Sarum Complex, Uxbridge, Middlesex, UB8 2RZ	26	Sarum Complex,		Middlesex	UB8 2RZ
46	https://raxel.co.uk	4 Ruston Road, Alma Park Industrial Estate, Grantham, NG	4	Ruston Road	Alma Park Industrial	Grantham	NG31 9SW
47	https://southamptoncruisecentre.com	298 Highfield Road Blackpool Lancashire FY4 3JU	298	Highfield Road	Blackpool Lancashire	FY4 3JU	
48	https://southamptoncruisecentre.com	62 Wolverhampton Street, Walsall WS2 8DD.	62	Wolverhampton Street		Walsall	WS2 8DD.
49	https://hyalofemme.co.uk	15079, Dunblane, FK15 5BA	15079		Dunblane		FK15 5BA
50	https://albaamicorum.com	12 Kinnerton Yard Belgravia London SW1X 8EB	12	Kinnerton Yard Belgravia London SW1X 8EB			
51	https://thesummitclinic.co.uk	40 Highgate West Hill, London N6 6LS	40	Highgate West Hill	London		N6 6LS
52	https://thesummitclinic.co.uk	206 Upper Richmond Road West, East Sheen, London SW1	206	Upper Richmond Road Wes	London		SW14 8AH.
53	https://thesummitclinic.co.uk	40 Highgate West Hill, London N6 6LS	40	Highgate West Hill	London		N6 6LS
54	https://coxhire.co.uk	Halesowen Road, Netherton, West Midlands, DY2 9NW		Halesowen Road	Netherton	West Midland	DY2 9NW
55	https://etc-inter.net	21 Church Road, Parkstone, Poole, BH14 8UF.	21	Church Road	Parkstone	Poole	BH14 8UF.
56	https://victoriafalls-zambiatravel-safar	6081 Hamilton Blvd, Suite 600 Allentown, PA 18106	6081	Hamilton Blvd	Allentown	PA	18106
57	https://foamtechniques.co.uk	39 Booth Drive Park Farm Industrial Estate South Wellingb	39	Booth Drive Park Farm Industrial Estate South W	Northampton:	NN8 6GR

Here are the last 55 USA addresses:

454	https://royersford.com	835 Township Line Rd. Phoenixville, PA 19460	835	Township Line Rd	Phoenixville	PA	19460
455	https://royersford.com	190 Royersford, PA 19468	190		Royersford	PA	19468
456	https://tmgi.net	46169 Westlake Drive, Suite 240 Sterling, VA 20165	46169	Westlake Drive	Sterling	VA	20165
457	https://devact.com	438 Queens Road Charlotte, NC 28207	438	Queens Road	Charlotte	NC	28207
458	https://gipreit.com	401 E Jackson St Ste 3300 Tampa, FL 33602	401	E Jackson St Ste 3300	Tampa	FL	33602
459	https://chcseia.com	400 North 17th Street Keokuk, IA 52632	400	North 17th Street	Keokuk	IA	52632
460	https://chcseia.com	1618 Main Street Keokuk, IA 52632	1618	Main Street	Keokuk	IA	52632
461	https://chcseia.com	2409 Spring Street Columbus City, IA 52737	2409	Spring Street	Columbus City	IA	52737
462	https://chcseia.com	1706 West Agency Road West Burlington, IA 52655	1706	West Agency Road	West Burlington	IA	52655
463	https://chcseia.com	400 S. Broadway Street Burlington, Iowa 52601	400	S. Broadway Street	Burlington	Iowa	52601
464	https://chcseia.com	1706 West Agency Road West Burlington, IA 52655	1706	West Agency Road	West Burlington	IA	52655
465	https://enidmad.org	129 N. University Ave. Enid, OK 73701	129	N. University Ave.	Enid	OK	73701
466	https://sellersconstruction.net	115 Milledgeville Hwy Gordon, GA 31031	115	Milledgeville Hwy	Gordon	GA	31031
467	https://goldridgecompanies.com	2713 North Clairemont Avenue Eau Claire, WI 54703	2713	North Clairemont Avenue	Claire	WI	54703
468	https://mid-america.com	2500 Broadway St. NE, Minneapolis, MN 55413	2500	Broadway St. NE	Minneapolis	MN	55413
469	https://mid-america.com	2500 Broadway St NE, Minneapolis, MN 55413	2500	Broadway St NE	Minneapolis	MN	55413
470	https://fresnogreekfest.com	2219 N. Orchard St Fresno, CA 93703	2219	N. Orchard St	Fresno	CA	93703
471	https://aaronsantosdentistry.com	14529 South Bascom Ave, Los Gatos, CA 95032	14529	South Bascom Ave	Los Gatos	CA	95032
472	https://thegrindcoffeebar.com	9532 West Linebaugh Ave, Tampa, FL 33626	9532	West Linebaugh Ave	Tampa	FL	33626
473	https://thegrindcoffeebar.com	9532 West Linebaugh Avenue, Tampa, FL, USA 813	9532	West Linebaugh Avenue	Tampa	USA	
474	https://habeggercorp.com	11413 Enterprise Park Drive Sharonville, OH 45241	11413	Enterprise Park Drive	Sharonville	OH	45241
475	https://billproctorandassociates.com	2050 South Blvd. Unit #412 Bloomfield Hills, MI 48303	2050	South Blvd. Unit #412	Bloomfield Hills	MI	48303
476	https://achieveservices.net	1000 Parkwood Circle SE STE 900 Atlanta, GA 30339	1000	Parkwood Circle SE STE	Atlanta	GA	30339
477	https://millerindustrial.com	621 E. Devon Ave, Elk Grove Village, IL 60007	621	E. Devon Ave	Elk Grove Village	IL	60007
478	https://appwinit.com	1239 Broadway. New York, NY 10001	1239	Broadway.	New York	NY	
479	https://hscpoly.com	12334 Geddes Road PO Box 80 Hemlock, Michigan 48626	12334	Geddes Road PO	Hemlock	Michigan	48626
480	https://hennejewelers.com	5501 Walnut Street Pittsburgh, PA 15232	5501	Walnut Street	Pittsburgh	PA	15232
481	https://roensalvage.com	180 E Redwood St Sturgeon Bay, WI 54235	180	E Redwood St	Sturgeon Bay	WI	54235
482	https://unitedairconditioning.com	13150 S Belcher Rd Largo, FL 33773	13150	S Belcher Rd	Largo	FL	33773
483	https://panniergraphics.com	345 Oak Road, Gibsonia, PA 15044	345	Oak Road	Gibsonia	PA	15044
484	https://panniergraphics.com	345 Oak Road Gibsonia, PA 15044	345	Oak Road	Gibsonia	PA	15044
485	https://allenvillagejewelers.com	190 E Stacy Rd #206, Allen, TX 75002	190	E Stacy Rd #206	Allen	TX	75002
486	https://artofthelakes.org	244, Battle Lake, MN 56515	244		Battle Lake	MN	56515
487	https://fcllp.com	570 Lexington Avenue New York, NY 10022	570	Lexington Avenue	New York	NY	
488	https://purcellvilledental.com	609 East Main Street, Suite T Purcellville, VA 20132	609	East Main Street	Purcellville	VA	20132
489	https://jackiescleaning.com	116 Crowley Rd, Lewiston, ME 04240	116	Crowley Rd	Lewiston	ME	04240
490	https://kiddsmith.com	117 Riviera Drive Boiling Springs , SC 29316	117	Riviera Drive Boiling Springs ,		SC	29316
491	https://kiddsmith.com	117 Riviera Drive Boiling Springs , SC 29316	117	Riviera Drive Boiling Springs ,		SC	29316
492	https://atlanticbeach-nc.com	125 West Fort Macon Rd Atlantic Beach, NC 28512	125	West Fort Macon Rd	Atlantic Beach	NC	28512
493	https://stmatthewmedina.org	400 N. Broadway St., Medina, OH 44256	400	N. Broadway St.	Medina	OH	44256
494	https://seniorcareagencyny.com	3601 Hempstead Tpke, Suite 203 Levittown, NY 11756	3601	Hempstead Tpke	Levittown	NY	11756
495	https://your-doctor.org	5 Brayford Square London, E1 0SG	5	Brayford Square London, E1 0SG			
496	https://glacier.chat	302 Sentinel Drive Annapolis Junction, MD 20701	302	Sentinel Drive	Annapolis Junction	MD	20701
497	https://neutresalon.com	101 S Market Street Frederick, MD 21701	101	S Market Street	Frederick	MD	21701
498	https://neutresalon.com	101 S Market Street Frederick, MD 21701	101	S Market Street	Frederick	MD	21701
499	https://paramountlandtitle.com	8904 Sony Lane, Knoxville, TN 37923	8904	Sony Lane	Knoxville	TN	37923
500	https://richardbrothersfinancial.com	50 Donald B. Dean Drive, Suite 1 South Portland, Maine 04	50	Donald B. Dean Drive	South Portland	Maine	04106
501	https://searchouthomes.com	1739 S. Jade Way, Suite 100 Meridian, Idaho 83642	1739	S. Jade Way	Meridian	Idaho	83642
502	https://triplecrownconstruction.com	1525 Tilco Drive, Unit B1, Frederick, MD 21704	1525	Tilco Drive	Frederick	MD	21704
503	https://wyandottewinery.com	4640 Wyandotte Dr Columbus, Ohio 43230	4640	Wyandotte Dr Columbus		Ohio	43230
504	https://societyfortheblind.org	1238 S Street Sacramento, CA 95811	1238	S Street	Sacramento	CA	95811
505	https://dillonmusic.com	325 Fulton Street Woodbridge, NJ 07095	325	Fulton Street	Woodbridge	NJ	07095
506	https://cabwhp.org	9800 S. La Cienega Blvd., Suite 905 Inglewood, CA 90301	9800	S. La Cienega Blvd.	Inglewood	CA	90301
507	https://sk4designs.com	3411 TELLURIDE, CO 81435	3411		TELLURIDE	CO	81435
508	https://embcmonroe.org	503 Maurice St. Monroe, NC 28112.	503	Maurice St. Monroe		NC	28112
509	https://embcmonroe.org	503 Maurice Street Monroe, NC 28112	503	Maurice Street	Monroe	NC	28112
510							

Step 4 (Optional). Remove invalid addresses. After removing the invalid addresses, I ended up with 208 complete addresses.