Extracting Drugs from Clinical Trials

The exercise is to extract some drugs names from a data set of clinical trials by matching to a list of drug names. In addition, some drugs names follow a naming pattern which can be used to classify them.

Data Sets:

clinical_trials_2015.jsonl

Dataset of clinical trials interventions from 2015. This is in JSON lines format.

Columns

- nct_id - this is the unique trial code. A trial will typically contain multiple interventions
- intervention_type - Type of intervention. We are interested in drug interventions
- intervention_name - Intervention. List of interventions i.e. drugs. Some text strings can contain multiple drugs

drugs.csv

Dataset of drug names and synonyms extracted from Wikidata.

Columns

- itemLabel - Primary drug name
- altLabel_list - Pipe separated of synonyms for the drug.

usan_stems_c.csv

This is a dataset mapping drugs to classification. The dataset includes prefixes, infixes, suffix

E.g.

- drug name with `acetam` suffix is `nootropic agents (learning, cognitive enhancers)` e.g. `piracetam`
- drug name with `aj` infix is `antiarrythmics (ajmaline deriviatives)` e.g. `lorajmine`
- drug name with `cef` prefix is `cephalosporins` e.g. `cefazolin`

Tasks

Task 1 - Match drug names

Use the drugs.csv dataset to match drugs names in the clinical_trials_2015.jsonl dataset

E.g.

```
Natalizumab,AN100226m|Antegran|Anti-alpha4 integrin|Anti-VLA4
```

Any mention of AN100226m, Antegran, Anti-alpha4 integrin and Anti-VLA4 should be matched to Natalizumab.

There can be multiple drug names contained within the `intervention_name` field

e.g. `{ "nct_id" : "NCT01969578", "intervention_name" : "bicalutamide + triptorelin", "intervention_type" : "Drug" }`

Should result in matching two drugs

Output a JSON file of nct_id and drugs

i.e.

Expected Output Format

```
[
{"nct_id": "NCT01969578", "drugs": ["bicalutamide", "triptorelin"]}
]
```

It won't be possible to match all intervention_name strings to the drugs.csv dataset

Task 2 - Match USAN codes

Many drugs names follow a structured format. The classification can be inferred from the prefix or suffix.

```
 "ac","-ac","anti-inflammatory agents (acetic acid derivatives)","bromfenac",
"subgroup:",,,,
,"-zolac","anti-inflammatory - pyrazole acetic acid derivatives","  pyrazole acetic acid
```

For example drugs names ending with -ac are anti-inflammatory agents. In addition there can be sub-classifications. E.g. drugs ending in -zolac are anti-inflammatory - pyrazole acetic acid derivatives

Example

For rovazolac:

USAN class = anti-inflammatory agents (acetic acid derivatives) USAN sub class = anti-inflammatory - pyrazole acetic acid derivatives

Expected Output Format

There should be a JSON file with the following structure, containing a record for each drug name found in the trial which is also matched to a USAN stem pattern

`drugs_usan.json`

```
[
  {
    "drug": "rovazolac",
    "usan_codes": [
      {"description":"anti-inflammatory agents (acetic acid derivatives)", "type":"class
      {"description":"anti-inflammatory - pyrazole acetic acid derivatives", "type":"sub
    ]
  },...
]
```

Task 3 - Generate counts of trials by USAN class

Perform some aggregation for drugs which have a matching USAN description report a list of trials which match the description

Report trials which match the each of the USAN descriptions. E.g. multiple drugs will map to the same class and there may be multiple drugs per trial.

Expected Output Format

```
[
  {
    "description": "anti-inflammatory agents (acetic acid derivatives)",
    "type": "class",
    "trials": ["NCT0123", "NCT0456",...]
  },
  {
    "description":"anti-inflammatory - pyrazole acetic acid derivatives",
    "type":"subclass"
    "trials": ["NCT0123"...]
  },
  ...
]
```

Task 4 - Generate Counts of USAN class pairs

Extend the aggregation in task 3 by generating counts of trials for pairs of USAN drug classes. For this exercise the USAN sub-classes can be ignored.

Expected Output Format

The file should be sorted with in descending order of trial counts.

```
[
  {
    "description_1": "anti-inflammatory agents (acetic acid derivatives)"
    "description_2": "antimyloidotics"
    "trial_count": 4
  }
]
```