Empath is a set of dictionaries spanning 194 different topics (e.g., "car", "leisure", "tool", real_estate", etc.), originally described in Fast et al. (2016), "Empath: Understanding Topic Signals in Large-Scale Text". In this work, we'll explore using empath to characterize texts and also use it as a jumping off point to think about **validity**.

The Empath *category* "help", for example, is a dictionary that contains the following *dictionary terms*:

help = {help, chore, responsible, help, grateful, maid, housekeeping, helpful, stabilize, servant, benefit, financial, aide, supportive, assistance, favor, tend, favor, encourage, wheelchair, nurse, patient, honor, protection, oversee, guide, hospitality, duty, advisor, carry, trust, obligation, rely, support, escort, friend, treat, offer, serve, cooperate, encouragement, promote, volunteer, counsel, kindly, crutch, aid, nursing, helper, request, rescue, provide, protect, generously, housework, advise, temporary, assist, entrust, prepare }

When applied to text, we can count which *tokens* have lemmas that are *dictionary terms*, indicating that it is indicative of that corresponding *category*. In the following text, the tokens that have lemmas corresponding to "help" dictionary terms have been highlighted:

(1) "The doctor prescribed a ***wheelchair*** rather than ***crutches*** to help heal the broken leg of the ***patient***. The hospital bill, however, was a significant ***financial*** burden to the ***patient***."

In [1]:
```
import spacy
from collections import Counter
```

In [2]:
```
nlp = spacy.load('en_core_web_sm', disable=['ner,parser'])
nlp.remove_pipe('ner')
nlp.remove_pipe('parser')
```

Out[2]:
```
('parser', <spacy.pipeline.pipes.DependencyParser at 0x246b593a100>)
```

First, let's read in the Empath dictionaries and create two mappings: one mapping categories to the dictionary terms within it, and one mapping dictionary terms to the categories they belong to (words can belong to multiple categories).

In [3]:
```
def read_dictionaries(filename):
    category_to_lemmas={}
    lemma_to_categories={}
    with open(filename, encoding="utf-8") as file:
        for line in file:
            cols=line.rstrip().split("\t")
            category=cols[0]
            category_to_lemmas[category]=set(cols)
            for lemma in cols:
                if lemma not in lemma_to_categories:
                    lemma_to_categories[lemma]={}
                lemma_to_categories[lemma][category]=1
    return lemma_to_categories, category_to_lemmas
```

In [4]:
```
lemma_to_categories, category_to_lemmas=read_dictionaries("../data/empath_categories.tx
```

Now let's use it to count up the Empath categories present in an input text.

In [5]:
```
def count_empath_categories(text, lemma_to_categories):
    category_counts=Counter()
    tokens=nlp(text.lower())
    for word in tokens:
        lemma=word.lemma_
        if lemma in lemma_to_categories:
            for cat in lemma_to_categories[lemma]:
                category_counts[cat]+=1

    for k,v in category_counts.most_common():
        print(v, k)
```

We'll run it on the following text from CNN.

"An oil spill that originated from Syria's largest refinery is growing and spreading across the Mediterranean Sea, and could reach the island of Cyprus by Wednesday, Cypriot authorities have said.

Syrian officials said last week that a tank filled with 15,000 tons of fuel had been leaking since August 23 at a thermal power plant on the Syrian coastal city of Baniyas. They said they had been able to bring it under control. Satellite imagery analysis by Orbital EOS now indicates that the oil spill was larger than originally thought, covering around 800 square kilometres (309 square miles) -- an area around the same size as New York City. The company told CNN Tuesday evening that the oil slick was around 7 kilometers (4 miles) from the Cypriot coast. The Cypriot Department of Fisheries and Marine research said that, based on a simulation of the oil spill's movements and meteorological data, the slick could reach the Apostlos Andreas Cape "in the next 24 hours." The department posted the statement at around 11 a.m. local time (4 a.m. ET) on Tuesday. It also said it would be willing to assist in tackling the spill."

In [6]:
```
text="""An oil spill that originated from Syria's largest refinery is growing and sprea

Syrian officials said last week that a tank filled with 15,000 tons of fuel had been le
Satellite imagery analysis by Orbital EOS now indicates that the oil spill was larger t
The Cypriot Department of Fisheries and Marine research said that, based on a simulatio
It also said it would be willing to assist in tackling the spill."""
```

In [7]:
```
count_empath_categories(text, lemma_to_categories)
```

```
8 liquid
7 speaking
6 shape_and_size
4 fire
4 beach
4 ocean
4 business
3 water
3 ship
3 sailing
3 power
3 warmth
```

```
3 work
3 morning
2 legend
2 leader
2 order
2 clothing
2 strength
2 vacation
2 technology
2 journalism
2 science
2 fabric
2 driving
2 college
2 internet
1 swimming
1 exotic
1 masculine
1 dominant_heirarchical
1 law
1 wedding
1 zest
1 magic
1 healing
1 plant
1 tourism
1 giving
1 computer
1 communication
1 leisure
1 party
1 military
1 war
1 school
1 reading
1 movement
1 superhero
1 social_media
1 real_estate
1 urban
1 optimism
1 help
1 office
```

Remember that dictionaries operate at the type level -- *every* instance of the word "financial", for instance, evokes the Empath "help" category, even though specific tokens of "financial" in context may not. Let's first identify what tokens in a text are evoking specific Empath categories, so we can examine them for their correctness.

Q1: Write a function that identifies the *tokens* corresponding to specific *dictionary terms* for an input *category* present in a given input text. This function should highlight those specific tokens in context by wrapping them in ***. Taking the category "help" and the input text given in (1) above, your output should look like the following:

The doctor prescribed a ***wheelchair*** rather than ***crutches*** to help heal the broken leg of the ***patient***. The hospital bill, however, was a significant ***financial*** burden to the ***patient***.

In [8]:
```
def print_empath_tokens_in_context(text, category_to_lemmas, category):
    # your code goes here
    #category_counts=Counter()

    line = []
    tokens=nlp(text.lower())

    for word in tokens:
        lemma=word.lemma_

        if lemma in category_to_lemmas[category]:
            #print('***' + lemma + '***')
            line.append('***'+ lemma + '***')
        else:
            #print(lemma)
            line.append(lemma)
    return ' '.join(line)
```

In [13]:
```
print_empath_tokens_in_context(text, category_to_lemmas, "liquid")
```

Out[13]:
```
'an ***oil*** ***spill*** that originate from syria \'s large refinery be grow and sprea
d across the mediterranean sea , and could reach the island of cyprus by wednesday , cyp
riot authority have say . \n\n syrian official say last week that a tank fill with 15,00
0 ton of fuel have be leak since august 23 at a thermal power plant on the syrian coasta
l city of baniyas . -PRON- say -PRON- have be able to bring -PRON- under control . \n sa
tellite imagery analysis by orbital eos now indicate that the ***oil*** ***spill*** be l
arge than originally think , cover around 800 square kilometre ( 309 square mile ) -- an
area around the same size as new york city . the company tell cnn tuesday evening that t
he ***oil*** slick be around 7 kilometer ( 4 mile ) from the cypriot coast . \n the cypr
iot department of fisheries and marine research say that , base on a simulation of the *
**oil*** ***spill*** \'s movement and meteorological datum , the slick could reach the a
postlos andreas cape " in the next 24 hour . " the department post the statement at arou
nd 11 a.m. local time ( 4 a.m. et ) on tuesday . \n -PRON- also say -PRON- would be will
ing to assist in tackle the ***spill*** .'
```

Q2. Use the function you just wrote to find all tokens identified by the "liquid," "fire," "beach" and "ocean" categories and use them to fill out the table below. Judge whether or not each token in context actually belongs to that category. Include a rationale if you think the decision would be contestable.

Category	Token in Context	Label	Rationale (if needed)
liquid	the mediterranean ***sea***, and could	Correct	N/A

You have a total of 20 rows (8 liquid, 4 fire, 4 beach, and 4 ocean, as identified above).

In [10]:
```
print_empath_tokens_in_context(text, category_to_lemmas, "fire")
```

Out[10]:
```
'an ***oil*** spill that originate from syria \'s large refinery be grow and spread acro
ss the mediterranean sea , and could reach the island of cyprus by wednesday , cypriot a
uthority have say . \n\n syrian official say last week that a tank fill with 15,000 ton
of fuel have be leak since august 23 at a thermal power plant on the syrian coastal city
of baniyas . -PRON- say -PRON- have be able to bring -PRON- under control . \n satellite
imagery analysis by orbital eos now indicate that the ***oil*** spill be large than orig
inally think , cover around 800 square kilometre ( 309 square mile ) -- an area around t
he same size as new york city . the company tell cnn tuesday evening that the ***oil***
```

```
slick be around 7 kilometer ( 4 mile ) from the cypriot coast . \n the cypriot departmen
t of fisheries and marine research say that , base on a simulation of the ***oil*** spil
l \'s movement and meteorological datum , the slick could reach the apostlos andreas cap
e " in the next 24 hour . " the department post the statement at around 11 a.m. local ti
me ( 4 a.m. et ) on tuesday . \n -PRON- also say -PRON- would be willing to assist in ta
ckle the spill .'
```

In [11]:
```
print_empath_tokens_in_context(text, category_to_lemmas, "beach")
```

Out[11]:
```
'an oil spill that originate from syria \'s large refinery be grow and spread across the
mediterranean ***sea*** , and could reach the ***island*** of cyprus by wednesday , cypr
iot authority have say . \n\n syrian official say last week that a tank fill with 15,000
ton of fuel have be leak since august 23 at a thermal power plant on the syrian ***coast
al*** city of baniyas . -PRON- say -PRON- have be able to bring -PRON- under control .
\n satellite imagery analysis by orbital eos now indicate that the oil spill be large th
an originally think , cover around 800 square kilometre ( 309 square mile ) -- an area a
round the same size as new york city . the company tell cnn tuesday evening that the oil
slick be around 7 kilometer ( 4 mile ) from the cypriot ***coast*** . \n the cypriot dep
artment of fisheries and marine research say that , base on a simulation of the oil spil
l \'s movement and meteorological datum , the slick could reach the apostlos andreas cap
e " in the next 24 hour . " the department post the statement at around 11 a.m. local ti
me ( 4 a.m. et ) on tuesday . \n -PRON- also say -PRON- would be willing to assist in ta
ckle the spill .'
```

In [12]:
```
print_empath_tokens_in_context(text, category_to_lemmas, "ocean")
```

Out[12]:
```
'an oil spill that originate from syria \'s large refinery be grow and spread across the
mediterranean ***sea*** , and could reach the ***island*** of cyprus by wednesday , cypr
iot authority have say . \n\n syrian official say last week that a tank fill with 15,000
ton of fuel have be leak since august 23 at a thermal power plant on the syrian ***coast
al*** city of baniyas . -PRON- say -PRON- have be able to bring -PRON- under control .
\n satellite imagery analysis by orbital eos now indicate that the oil spill be large th
an originally think , cover around 800 square kilometre ( 309 square mile ) -- an area a
round the same size as new york city . the company tell cnn tuesday evening that the oil
slick be around 7 kilometer ( 4 mile ) from the cypriot ***coast*** . \n the cypriot dep
artment of fisheries and marine research say that , base on a simulation of the oil spil
l \'s movement and meteorological datum , the slick could reach the apostlos andreas cap
e " in the next 24 hour . " the department post the statement at around 11 a.m. local ti
me ( 4 a.m. et ) on tuesday . \n -PRON- also say -PRON- would be willing to assist in ta
ckle the spill .'
```

Category	Token in Context	Label	Rationale (if needed)
liquid	***oil*** that spilled	Correct	N/A
liquid	***spill*** that originate from syria's large refinery	Correct	N/A
liquid	***oil*** oil spill that from satellite	Correct	N/A
liquid	***spill*** spill sizes	Correct	N/A
liquid	***oil*** slick still a liquid	Correct	N/A
liquid	***oil*** refer to a simulation	Incorrect	N/A
liquid	***spill*** refer to a simulation	Incorrect	N/A
liquid	***spill*** refer to the oil spill	Correct	N/A
fire	***oil*** spill	Incorrect	N/A
fire	***oil*** spill	Incorrect	N/A

Category	Token in Context	Label	Rationale (if needed)
fire	***oil*** spill	Incorrect	N/A
fire	***oil*** spill	Incorrect	N/A
beach	***sea*** mediterranean sea	Incorrect	N/A
beach	***island***	Incorrect	N/A
beach	***coastal*** Cypriot coastal	Incorrect	N/A
beach	***coast*** Cypriot city	Incorrect	N/A
ocean	***sea*** mediterranean sea	Correct	N/A
ocean	***island*** cyprus isalnd	Incorrect	N/A
ocean	***coastal*** syrian coastal city	Incorrect	N/A
ocean	***coast*** cypruit coast	Incorrect	N/A

You have a total of 20 rows (8 liquid, 4 fire, 4 beach, and 4 ocean, as identified above).

Q3. Using that table, calculate the precision of the "liquid," "fire," "beach" and "ocean" categories for this passage using the following equation:

$$\text{Precision(liquid)} = \frac{\#\text{ of "liquid" tokens identified by Empath that you marked as correct}}{\#\text{ of "liquid" tokens identified by Empath}}$$

You should report 4 numbers (one measure of precision for each of the 4 categories).

In [15]:
```
def precision (correct, identified):
    return correct / identified
```

In [17]:
```
#Liquid
precision(6, 8)
```

Out[17]:
```
0.75
```

In [18]:
```
#fire
precision(0, 4)
```

Out[18]:
```
0.0
```

In [19]:
```
#beach
precision(0, 4)
```

Out[19]:
```
0.0
```

In [20]:
```
#ocean
precision(1, 4)
```

Out[20]:
```
0.25
```