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 [157…
```
import spacy
from collections import Counter
```

In [158…
```
nlp = spacy.load('en_core_web_sm', disable=['ner,parser'])
nlp.remove_pipe('ner')
nlp.remove_pipe('parser')
```

Out[158… ('parser', <spacy.pipeline.dep_parser.DependencyParser at 0x1362606a0>)

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 [159…

```
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 [160…

```
lemma_to_categories, category_to_lemmas=read_dictionaries("../data/empath_
```

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

In [161…

```
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 [162…

```
text="""An oil spill that originated from Syria's largest refinery is grow

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

In [163…

```
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 [164…

```
def print_empath_tokens_in_context(text, category_to_lemmas, category):
    # your code goes here
    tokens = nlp(text.lower())
    final=[]
    for word in tokens:
        lemma = word.lemma_
        if lemma in category_to_lemmas[category] or word in category_to_ler
            word="***"+str(word)+"***"
            print("",word,end="")
        else:
            # don't add space before a punctuation mark
            if(str(word) in [",","'","\".","."]):
                print(word,end="")
            else:
                print("",str(word),end="")
```

In [165…

```
print_empath_tokens_in_context(text, category_to_lemmas, "liquid")
```

```
 an ***oil*** ***spill*** that originated from syria 's largest refinery is
growing and spreading across the mediterranean sea, and could reach the isl
and of cyprus by wednesday, cypriot authorities have said.

 syrian officials said last week that a tank filled with 15,000 tons of fue
l had been leaking since august 23 at a thermal power plant on the syrian c
oastal city of baniyas. they said they had been able to bring it under cont
rol.
 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 aro
und 7 kilometers ( 4 miles ) from the cypriot coast.
 the cypriot department of fisheries and marine research said that, based o
n 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***.
```

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
liquid	an ***oil*** spill	Correct	N/A
liquid	an oil ***spill***	Correct	The word "spill" does evoke the imagery of a liquid.
liquid	the ***oil*** spill was larger than originally thought	Correct	N/A
liquid	an oil ***spill*** was larger than originally thought	Correct	N/A
liquid	the ***oil*** slick	Correct	Oil slick still refers to the layer of liquid oil on the water.
liquid	simulation of the ***oil*** spill	Incorrect	This sentence is in the context of a simulation, and doesn't really fall under the category of "liquid".
liquid	simulation of the oil ***spill***	Incorrect	This sentence also is in the context of a simulation, and doesn't really fall under the category of "liquid".
liquid	tackling the ***spill***	Correct	N/A
fire	an ***oil*** spill	Incorrect	The context here is about an oil spill, there were no fires involved.
fire	the ***oil*** spill was larger than originally thought	Incorrect	The context here is about an oil spill, there were no fires involved.
fire	the ***oil*** slick was around 7 kilometers	Incorrect	The context here is about an oil spill, there were no fires involved.
fire	based on a simulation of the ***oil*** spill's movements	Incorrect	The context here is about an oil spill, there were no fires involved.
beach	across the mediterranean ***sea***	Incorrect	The main subject of this sentence is still the oil spill. We haven't reached the part where shores/coasts are being spoken about.
beach	could reach the ***island*** of cyprus by wednesday	Correct	N/A
beach	power plant on the syrian ***coastal*** city of baniyas	Correct	N/A
beach	from the cypriot ***coast***	Correct	N/A

ocean	spreading across the mediterranean ***sea***	Correct	N/A
ocean	could reach the ***island*** of cyprus by wednesday	Correct	The oil spill in the ocean is being spoken about. Plus, islands invoke the imagery of oceans as well.
ocean	on the syrian ***coastal*** city of baniyas	Correct	N/A
ocean	(4 miles) from the cypriot ***coast***	Correct	N/A

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

In [167…

```
print("fire:")
print_empath_tokens_in_context(text, category_to_lemmas, "fire")
print("\n\n")
print("beach:")
print_empath_tokens_in_context(text, category_to_lemmas, "beach")
print("\n\n")
print("ocean:")
print_empath_tokens_in_context(text, category_to_lemmas, "ocean")
```

```
fire:
 an ***oil*** spill that originated from syria 's largest refinery is growi
ng and spreading across the mediterranean sea, and could reach the island o
f cyprus by wednesday, cypriot authorities have said.

 syrian officials said last week that a tank filled with 15,000 tons of fue
l had been leaking since august 23 at a thermal power plant on the syrian c
oastal city of baniyas. they said they had been able to bring it under cont
rol.
 satellite imagery analysis by orbital eos now indicates that the ***oil***
spill was larger than originally thought, covering around 800 square kilome
tres ( 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 o
n 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.


beach:
 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 fue
l 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 unde
r 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 co
mpany 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 o
n a simulation of the oil spill 's movements and meteorological data, the s
lick could reach the apostlos andreas cape" in the next 24 hours." the depa
rtment posted the statement at around 11 a.m. local time ( 4 a.m. et ) on t
uesday.
 it also said it would be willing to assist in tackling the spill.


ocean:
 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 fue
l 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 unde
r 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 co
mpany 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 o
n a simulation of the oil spill 's movements and meteorological data, the s
lick could reach the apostlos andreas cape" in the next 24 hours." the depa
rtment posted the statement at around 11 a.m. local time ( 4 a.m. et ) on t
uesday.
 it also said it would be willing to assist in tackling the spill.
```

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

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

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

Category	No. of tokens marked as correct	No. of tokens identified	Precision
Liquid	6	8	0.75
Fire	0	4	0
Beach	3	4	0.75
Ocean	4	4	1

In [ ]: