Empath

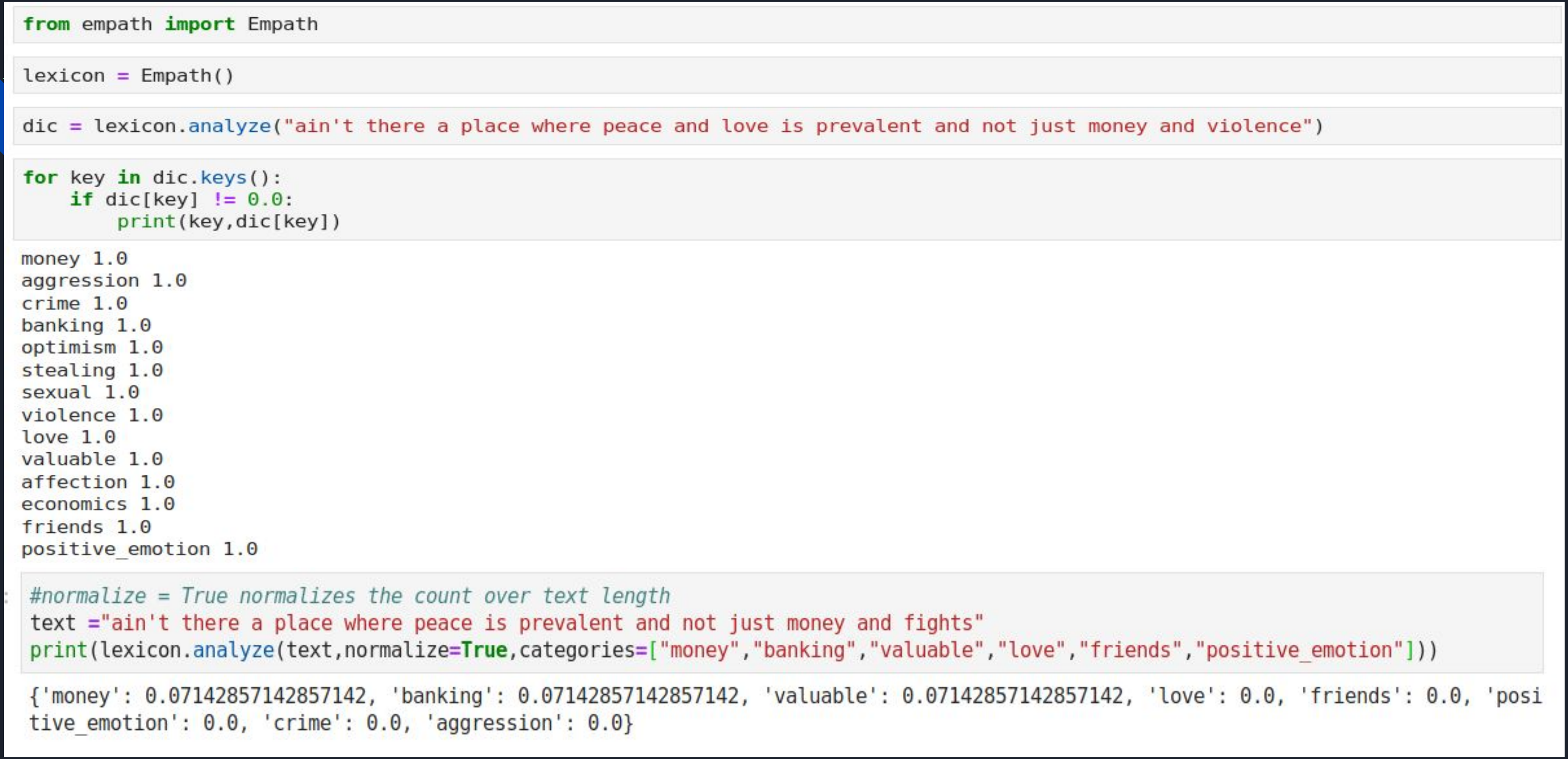
By:
-Bhoomika panwar
-Shivam Agarwal

Empath - A text analysis tool

- Empath (a living lexicon mined from modern text on the web) analyzes text across 194 gold standard topics and emotions (e.g., childishness or violence)
- can generate and validate new lexical categories on demand from a user-generated set of few seed terms.
- It uses combination of deep learning and crowdsourcing to generate and validate new categories
- For example, using the seed terms “twitter” and “facebook,” we can generate and validate a category for social media.

Empath (contd.)

- Also provides an option to get normalized score for each category
- Alternative to LIWC(Linguistic Inquiry and Word Count)
- **WHY EMPATH?**
    - LIWC requires a paid license to use whereas Empath is a free software.
    - LIWC is small(80 categories and many among that have fewer than 100 words.)
    - Many potentially useful categories like violence or social media don’t exist in LIWC



```
from empath import Empath
```

```
lexicon = Empath()
```

```
dic = lexicon.analyze("ain't there a place where peace and love is prevalent and not just money and violence")
```

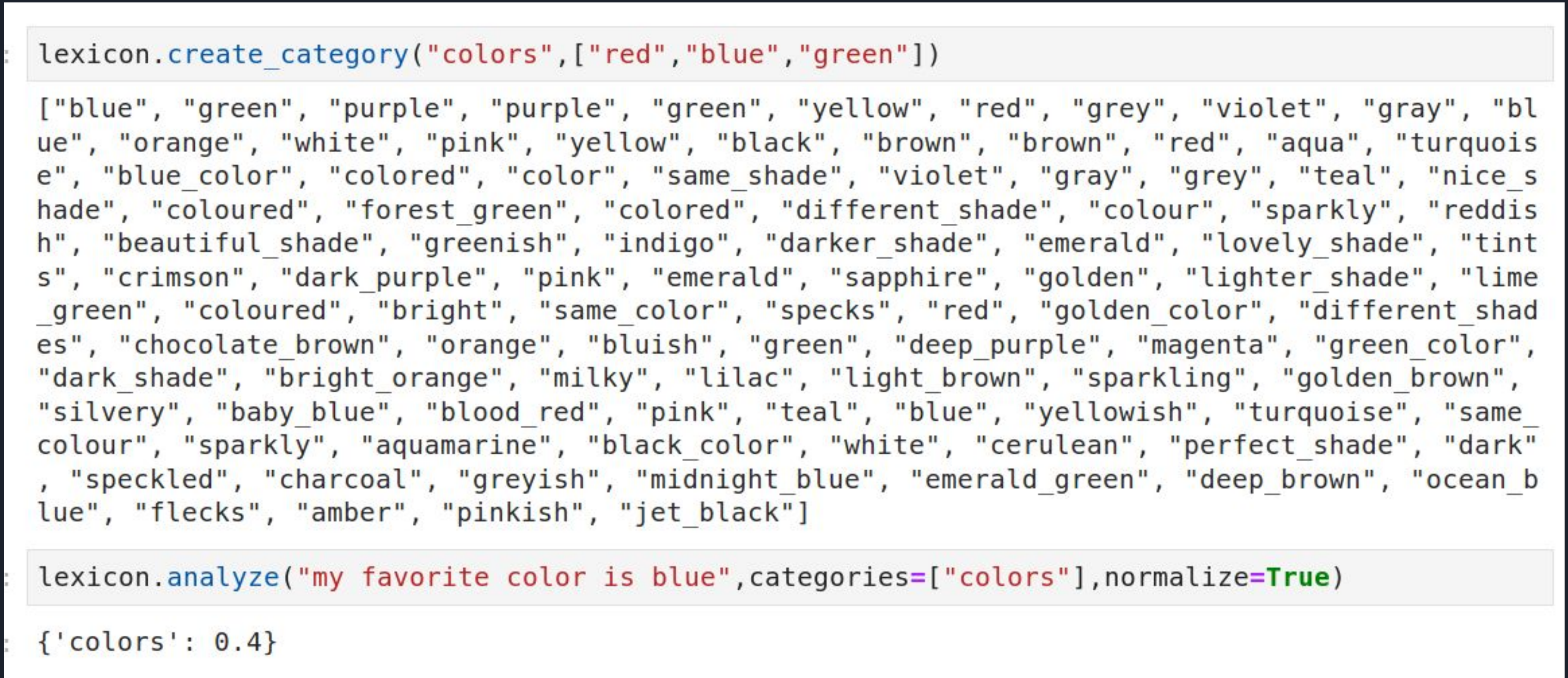
```
for key in dic.keys():
    if dic[key] != 0.0:
        print(key,dic[key])
```

```
money 1.0
aggression 1.0
crime 1.0
banking 1.0
optimism 1.0
stealing 1.0
sexual 1.0
violence 1.0
love 1.0
valuable 1.0
affection 1.0
economics 1.0
friends 1.0
positive_emotion 1.0
```

```
#normalize = True normalizes the count over text length
text ="ain't there a place where peace is prevalent and not just money and fights"
print(lexicon.analyze(text,normalize=True,categories=["money","banking","valuable","love","friends","positive_emotion"]))
```

```
{'money': 0.07142857142857142, 'banking': 0.07142857142857142, 'valuable': 0.07142857142857142, 'love': 0.0, 'friends': 0.0, 'posi
tive_emotion': 0.0, 'crime': 0.0, 'aggression': 0.0}
```

EMPATH IN ACTION



```
lexicon.create_category("colors",["red","blue","green"])
```

```
["blue", "green", "purple", "purple", "green", "yellow", "red", "grey", "violet", "gray", "bl
ue", "orange", "white", "pink", "yellow", "black", "brown", "brown", "red", "aqua", "turquois
e", "blue_color", "colored", "color", "same_shade", "violet", "gray", "grey", "teal", "nice_s
hade", "coloured", "forest_green", "colored", "different_shade", "colour", "sparkly", "reddis
h", "beautiful_shade", "greenish", "indigo", "darker_shade", "emerald", "lovely_shade", "tint
s", "crimson", "dark_purple", "pink", "emerald", "sapphire", "golden", "lighter_shade", "lime
_green", "coloured", "bright", "same_color", "specks", "red", "golden_color", "different_shad
es", "chocolate_brown", "orange", "bluish", "green", "deep_purple", "magenta", "green_color",
"dark_shade", "bright_orange", "milky", "lilac", "light_brown", "sparkling", "golden_brown",
"silvery", "baby_blue", "blood_red", "pink", "teal", "blue", "yellowish", "turquoise", "same_
colour", "sparkly", "aquamarine", "black_color", "white", "cerulean", "perfect_shade", "dark"
, "speckled", "charcoal", "greyish", "midnight_blue", "emerald_green", "deep_brown", "ocean_b
lue", "flecks", "amber", "pinkish", "jet_black"]
```

```
lexicon.analyze("my favorite color is blue",categories=["colors"],normalize=True)
```

```
{'colors': 0.4}
```

Generating new categories using Empath.

How Empath Works?

Where do names of category come from?

When user provides “shirt” and “hat” as seed words, ConceptNet (a freely-available semantic network, designed to help computers understand the meanings of words that people use.) tells us shirts and hats are articles of clothing. So, Empath can create and validate a clothing category, using “shirt” and “hat” as seed words.

Where do category terms come from?

Empath’s model uses seed words to generate a candidate set of member terms for its categories, which we validate through paid crowdsourcing. Empath generates these category terms by querying a vector space model (VSM) trained by a neural network on a large corpus of text. This VSM allows Empath to examine the similarity between words across many dimensions of meaning. For example, given seed words like “facebook” and “twitter,’ Empath finds related terms like “pinterest” and “selfie.”

How Empath Works?(Contd.)

Why crowdsourcing to validate?

Human-validated categories can ensure that accidental terms do not slip into a lexicon. By filtering Empath’s categories through the crowd, we offer the benefits of both modern NLP and human validation: increasing category precision, and more carefully validating category contents. To validate each of Empath’s categories, we have created a crowdsourcing pipeline on Amazon Mechanical Turk.

Overview of empath’s working

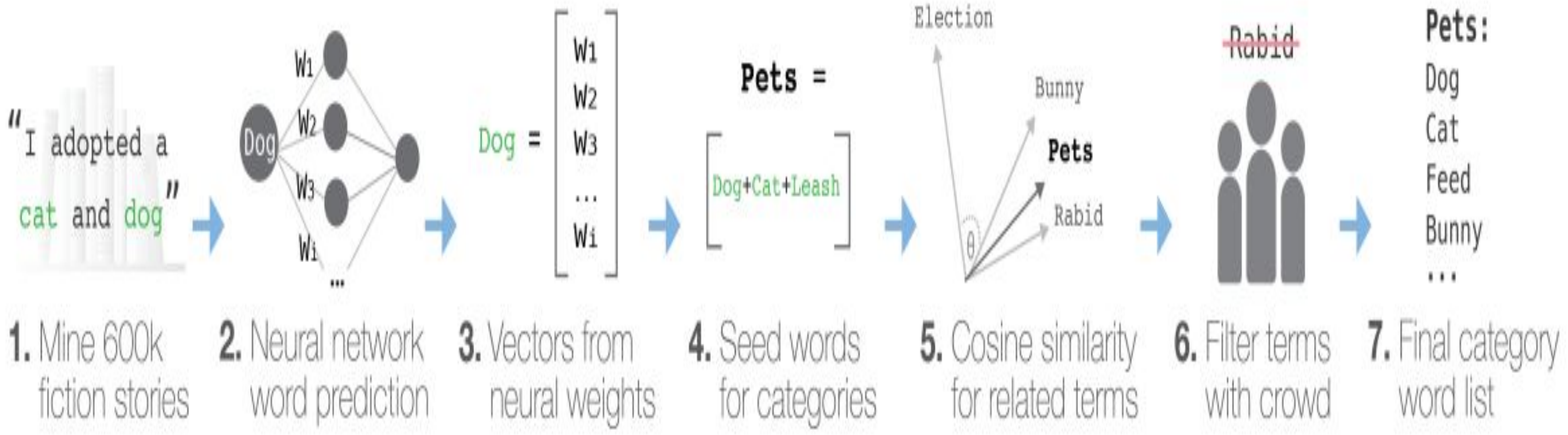

"I adopted a cat and dog"
Dog
W1
W2
W3
Wi
...
Dog = [W1 W2 W3 ... Wi]
Pets = [Dog+Cat+Leash]
Election
Bunny
Pets
Rabid
θ
~~Rabid~~
Pets:
Dog
Cat
Feed
Bunny
...
1. Mine 600k fiction stories
2. Neural network word prediction
3. Vectors from neural weights
4. Seed words for categories
5. Cosine similarity for related terms
6. Filter terms with crowd
7. Final category word list

Figure 2. Empath learns word embeddings from 1.8 billion words of fiction, makes a vector space from these embeddings that measures the similarity between words, uses seed terms to define and discover new words for each of its categories, and finally filters its categories using crowds.