

# Disaster Mapping (08/01/2019)



**Patrick Wales-Dinan**  
**Laura Luo**  
**Andre Jimenez**  
**Antony Paulson Chazhoor**

# **Agenda**



```
graph LR; A((Agenda)) --- B((Background)); A --- C((Problem Statement)); A --- D((Approach));
```

**Background**

**Problem  
Statement**

**Approach**



# Background

## Current limitations

**Difficulties to map and identify locations of survivors needing assistance.**

**Lack of info or biased info if only rely on News Media.**

## Advantages of Social Media

**Survivors will resort to using social media to call for help or share locations.**

**Tweets posted from witnesses are usually on real-time and up-to-date.**

**Rich info comparing with news: more first-hand info**



# Problem Statement

- **Goal:** Leveraging Social Media to Map Disasters
- **Intro:**
  - Utilize **classification model** to pick emergency/disaster related tweets
  - leverage tweets to **locate hot spots** of where people are needing assistance



# Steps

**1. Data Collection  
(Twitter API)**

**3. EDA/ Clustering**

**5. Google Map  
Plotting**

**2. Build Disaster/  
Emergency Corpus**

**4. Labeling/  
Modeling**

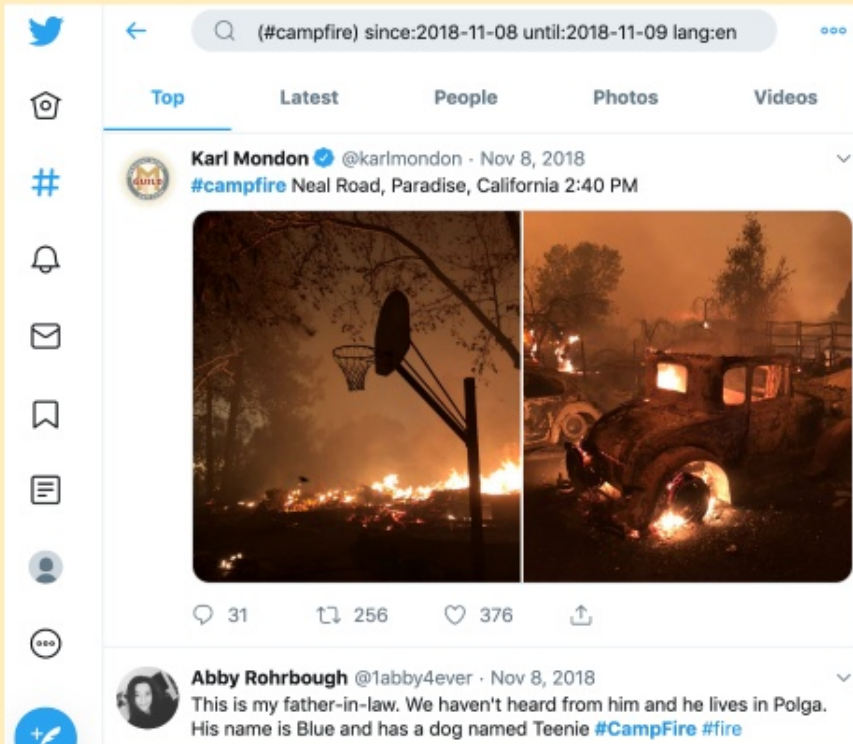
# Collecting Data (tweets)



**Topics**

**Data**

# WildFire



## #campfire

- **Date:**  
11/08/18 - 11/25/18
- **Impact zone:**  
Butte County, CA



# WildFire



## #carrfire

- **Date:**  
07/23/18 - 08/30/18
- **Impact zone:**  
Shasta & Trinity, CA



# Hurricane




## #hurricaneharvey

- **Date:**  
08/17/17 - 09/02/17
- **Impact zone:**  
South& East Texas

more...



## Training set

- **1397 unique tweets**
  - **date** (1st day when disaster happens)
  - **hashtag**: #campfire, #carrfire, #harricaneharvey
  - **language**: English
- 



## Testing set

- **832 unique tweets**
  - **date** (during disasters)
  - **hashtag**: #campfire, #carrfire
  - **language**: English
  - **Coordinates**: eg. [-95.36 29.76 25mi]
- 



# **Natural Language Processing**

**Building  
Disaster  
Corpus**

# Building a disaster corpus

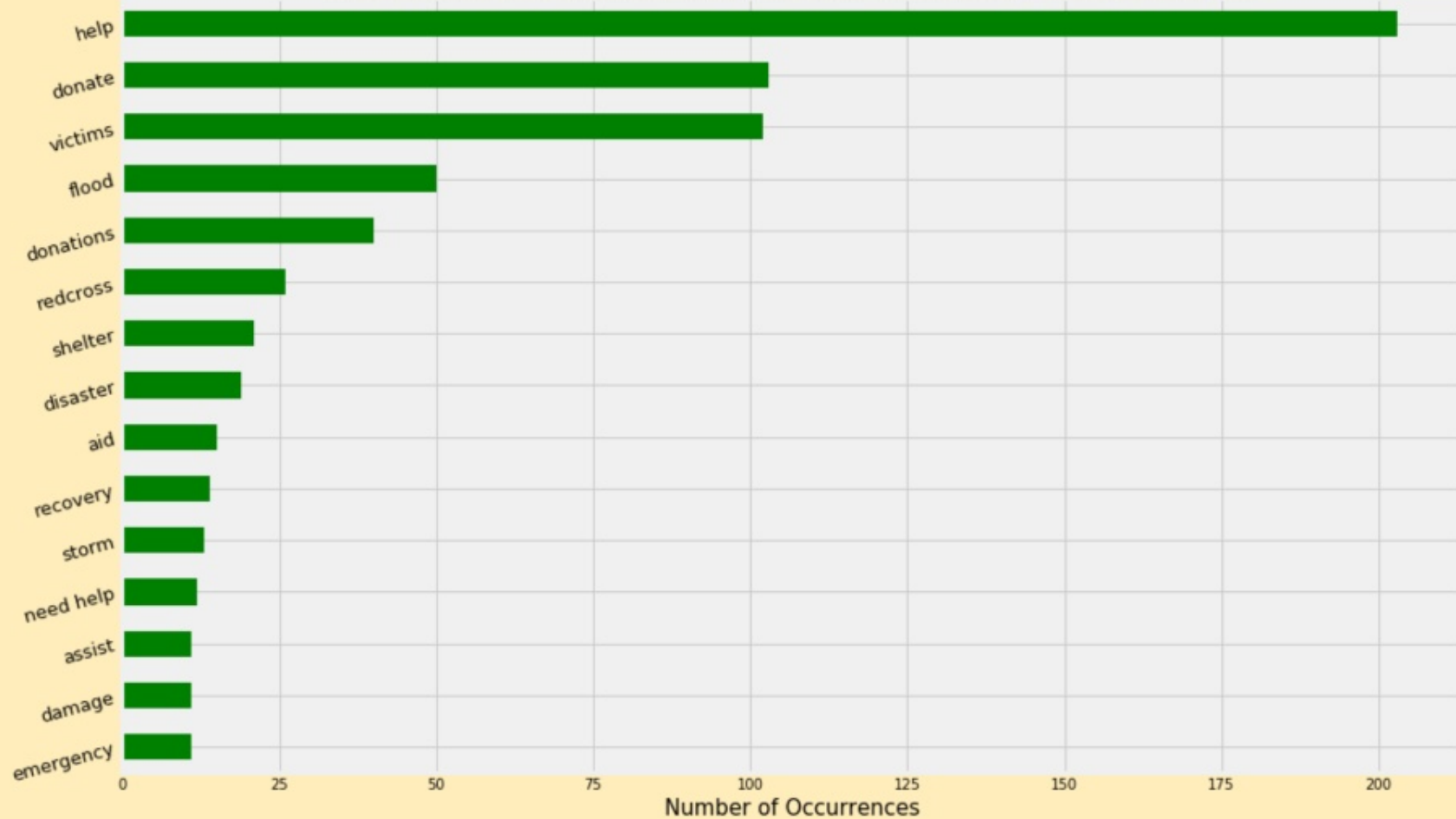
- Imported and concatenated three data frames each containing tweets that pertain to a natural disaster: Camp Fire, Carre Fire, and Hurricane Harvey.
- To clean up our data a little bit, we created an nltk WordNetLemmatizer class to serve as a parameter for our CountVectorizer function so that words can be shortened to their base forms and not be confused as completely different words
- We also created an nltk RegexpTokenizer (regular expressions tokenizer) class to serve as a parameter for our CountVectorizer function to remove punctuation & convert all words to lower case.



# Building a disaster corpus

- We ran sklearn's CountVectorizer with our WordNetLemmatizer, RegexpTokenizer, stop\_words parameters to transform the lists of the tweet words into training data features that we can pass into a model. CountVectorizer created columns, where each column counts how many times each word was observed in each tweet . These features became a pandas series called tweets\_words.
- Then we sorted each word in our tweet\_words series by its count frequency and converted the series into a data frame called tweet\_words\_with\_counts to get a clear view of related words
- We briefly scanned and manually removed words (like https and rt) that were completely unrelated to help words. We then created a list of help related words based on their word frequency.

The 15 Most Common 'Help Words' in Harvey Tweet Corpus

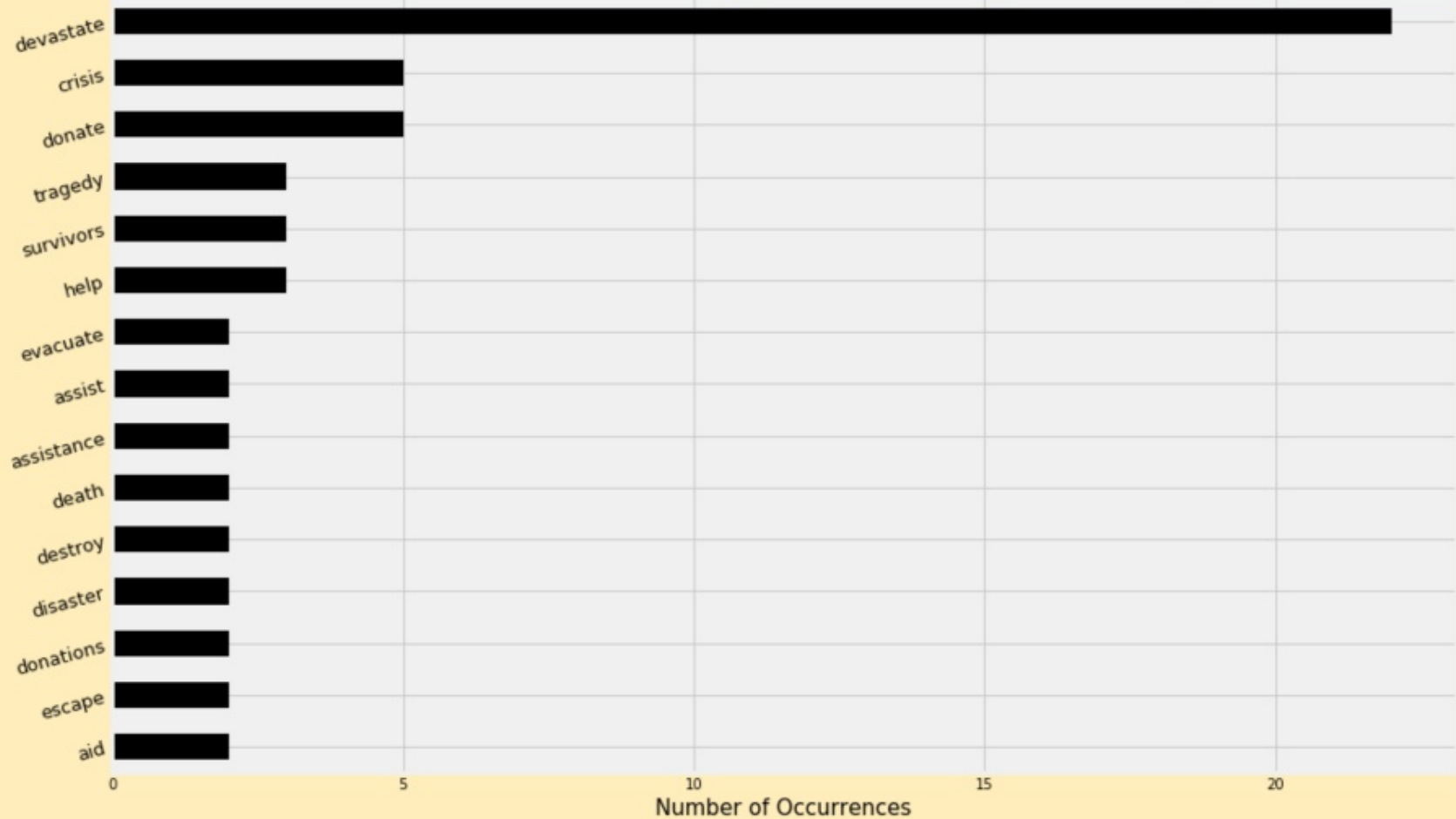




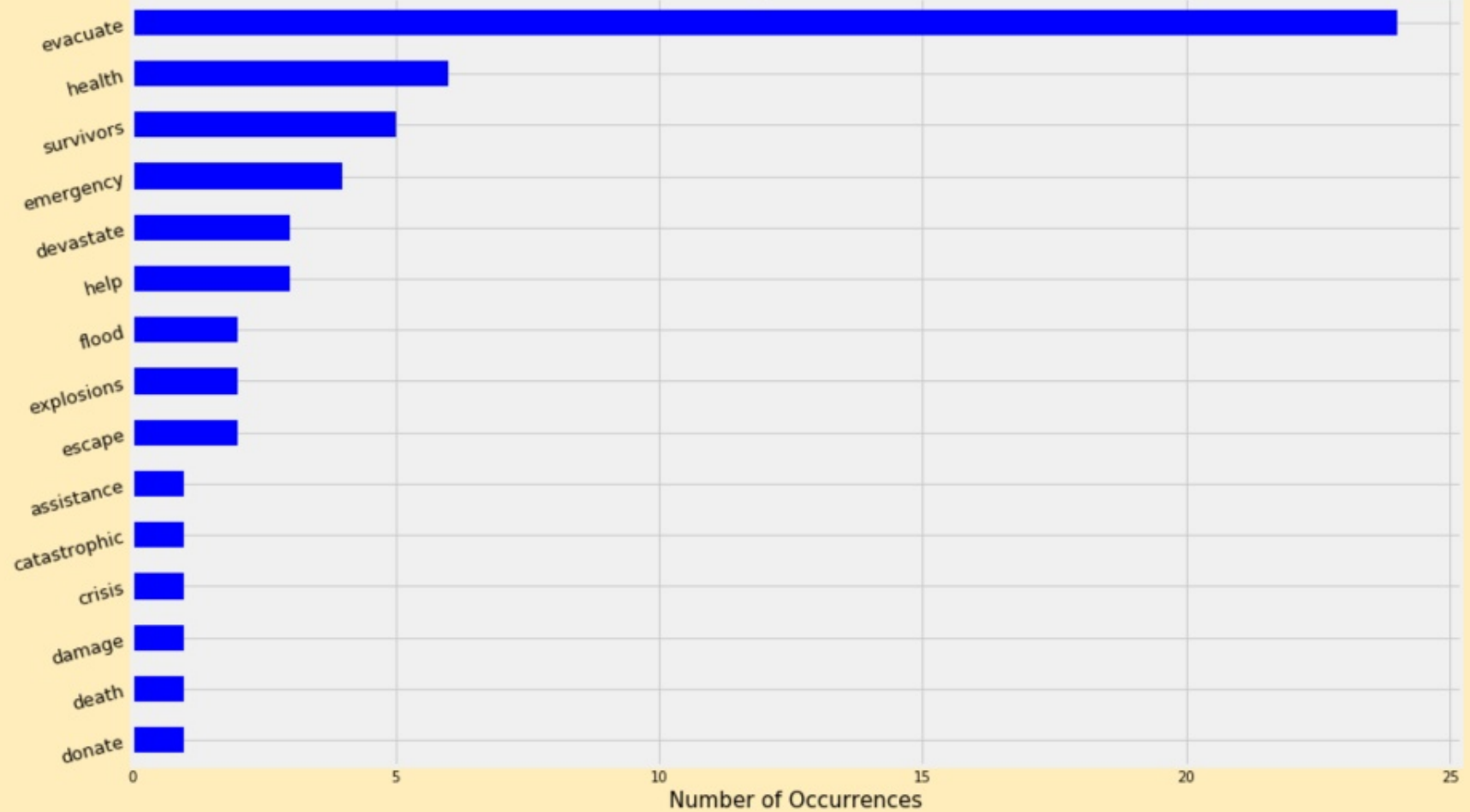
# Building a disaster corpus

- We then inputted `tweet_words_with_counts` as our corpus into a Word2Vec model so that we can use Word2Vec's `.most_similar` method to find and append more help-related words into our existing list of help related words
- To find and append more help-related words, we repeated the Word2Vec procedure once more using gensim's Wikipedia data as our new corpus. After performing Word2Vec on both corpora we ended up with an expanded list of 337 help related words. We then scanned the list one final time to handpick our final disaster words for the training data labeling phase of our project.

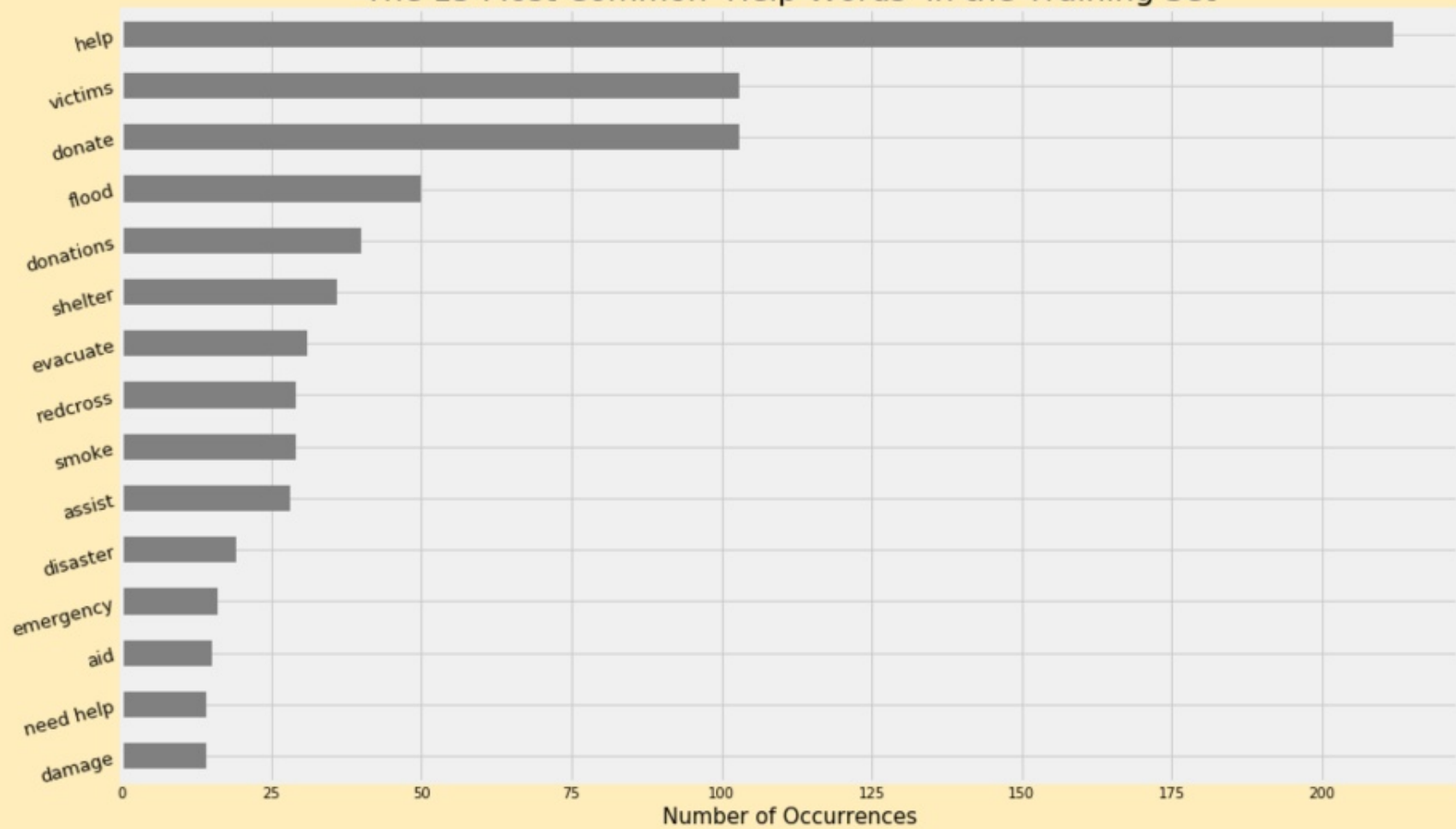
The 15 Most Common 'Help Words' in Camp Tweet Corpus



The 15 Most Common 'Help Words' in Carr Tweet Corpus



The 15 Most Common 'Help Words' in the Training Set





The diagram features a large orange circle in the center containing the text "Labeling & Modeling". To the right of this central circle are three smaller, light-yellow circles arranged vertically. The top circle is labeled "Clustering", the middle one "Labeling", and the bottom one "Modeling". The background of the slide is a light gray with a faint world map.

# **Labeling & Modeling**

**Clustering**

**Labeling**

**Modeling**

## EDA- Clustering

- Use **K-means clustering model** to check there is are any patterns before labeling
  - set number of clusters as 2 ,3 or 5
  - no patterns at all
  - most tweets are centralized in one cluster



# Labeling

```
final_disaster_words = ['needing', 'donate', 'unsafe',  
    'danger', 'sediments', 'death',  
    'disaster', 'needs', 'freezing',  
    'hypothermia', 'disasters', 'severe',  
    'injure', 'help us', 'devastating',  
    'victim', 'property', 'catastrophic',  
    'sinking', 'fatal', 'please help',  
    'damage', 'debris', 'accident',  
    'support harricane', 'medical', 'helping',  
    'accidents', 'hazards', 'shelter',  
    'destroy', 'damaging', 'escape',  
    'aid', 'smoke', 'shelters',  
    'toxin', 'health', 'navigate',  
    'erosion', 'pleasehelp', 'tragedy',  
    'redcross', 'flood', 'injury',  
    'abandoned', 'casualties', 'medicine',  
    'evacuate', 'respond', 'emergency',  
    'disaster assist team', 'survivors', 'recovery',  
    'explosion', 'donations', 'fires',  
    'victims', 'floods', 'explosions',  
    'crisis', 'loss', 'breach',  
    'abandonment', 'injuries', 'help',  
    'need shelter', 'explodes', 'flame',  
    'help me', 'need support', 'painful',  
    'starvation', 'assistance', 'lightning',  
    'flooding', 'supplies', 'withdrawal',  
    'losing', 'needed', 'storm',  
    'disasterassistteam', 'lose', 'storms',  
    'needhelp', 'lose home', 'lost',  
    'foodstocks', 'devastate', 'help needed',  
    'assist', 'fear', 'serious',  
    'abandon', 'crises',  
    'sediment', 'need help',  
    'destroy home', 'droughts',  
    'send help',
```

- 100 disaster words

- Rule:

If a tweet contains one or more words in disaster words list --> label 1  
otherwise --> label 0



# Classification Models

- **Unbalanced class --> Oversampling**
  - 942 out of 1397 tweets were labeled as 1
  - 455 tweets were labeled as 0
- **Build 7 classification models on training set**



Logistic Regression



Bagging Classifier



Gradient Boosting Classifier



Random Forest Classifier



AdaBoost Classifier



Decision Tree



SVM

- **Apply models on testing set**

# Results

## • Performance Overview (training set)

	train	test	F1-train	F1-test	true_neg	fal_pos	fal_neg	true_po
lr_class	0.998401	0.868106	0.998379	0.874715	170	31	24	192
forest_class	0.901679	0.846523	0.900566	0.848341	174	27	37	179
tree_class	0.884093	0.832134	0.886807	0.843750	158	43	27	189
ada_class	0.926459	0.846523	0.924466	0.843137	181	20	44	172
bag_class	0.599520	0.522782	0.332889	0.167364	198	3	196	20
svc	0.808153	0.633094	0.759036	0.451613	201	0	153	63
grad	0.948841	0.880096	0.947282	0.881517	181	20	30	186

## • Conclusion (testing set)

- 142 out of 832 tweets were classified as pertaining to disasters. (3 or more models vote 1.)

	lr_class	forest_class	tree_class	ada_class	bag_class	svc	grad	total
496	1	1	1	1	1	0	1	6
0	1	1	1	1	0	0	1	5
130	1	1	1	1	0	0	1	5
598	1	1	1	1	0	0	1	5
769	1	1	1	1	0	0	1	5
1	1	1	1	1	0	0	1	5



# **Mapping Coordinates (Google API)**

**Plotting  
Procedure**

**Satellite  
view**

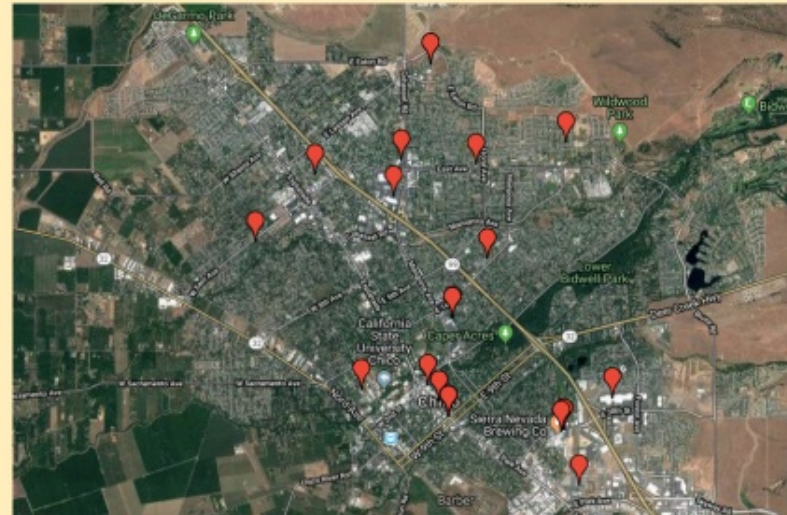
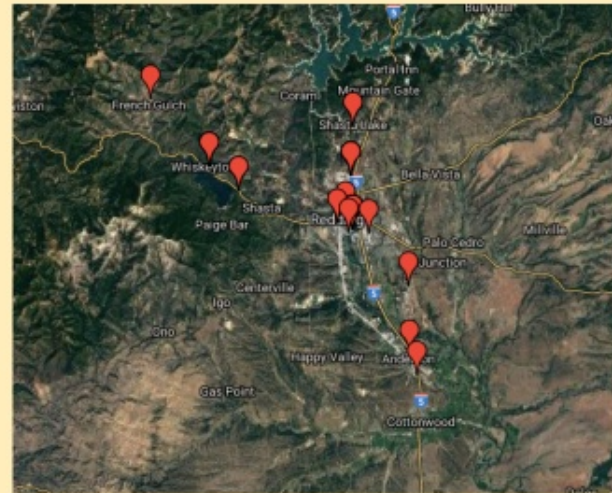
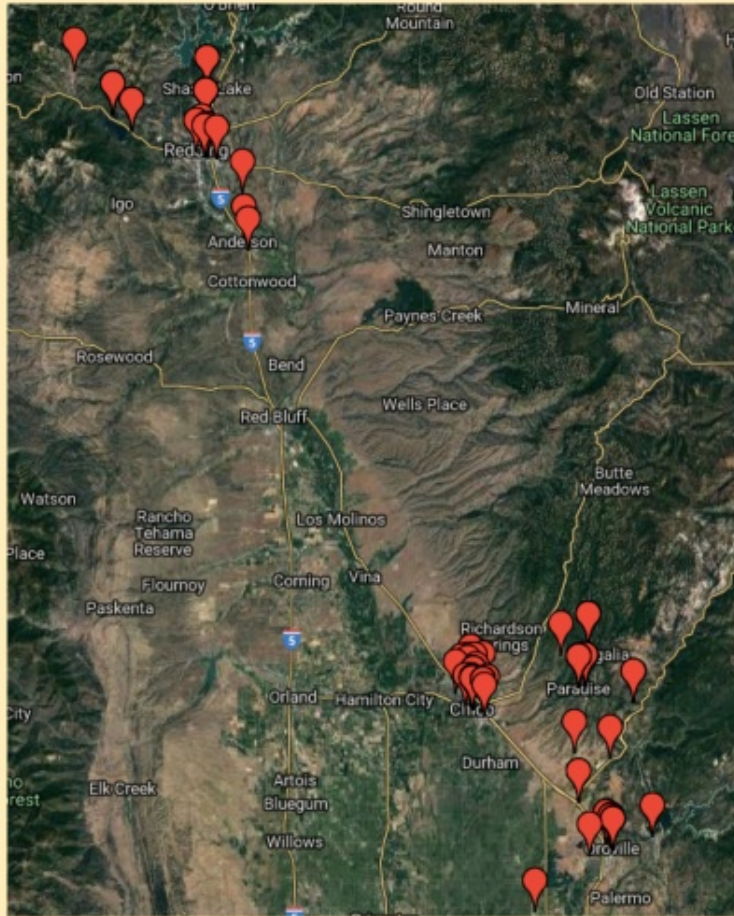
**Map  
view**

## Plotting Procedure

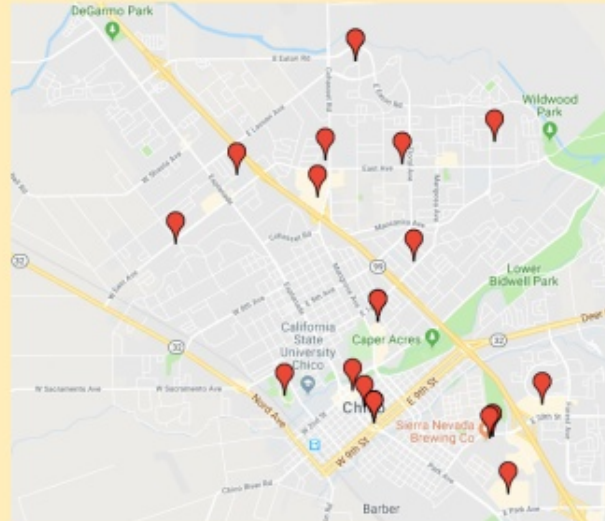
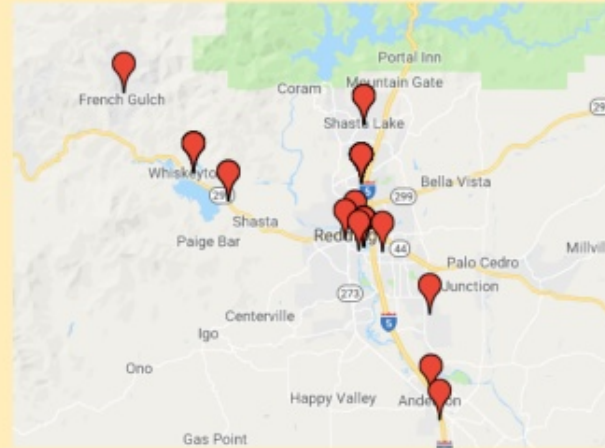
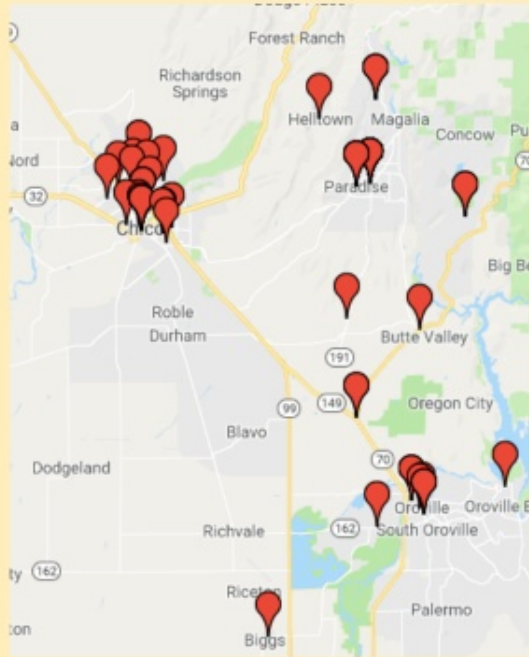
- Converted our dataframe's Latitude and Longitude series into separate lists
- Inputted these lists and our Gmaps API keys as parameters into gmplot's GoogleMapPlotter method
- Plotted help-related tweets with scatter points that spread 100 miles across Sacramento Valley (from Shasta County to Butte County)



# Satellite View



# Map View





# **Limitations & To-do**



## Limitations & to-do

- **Twitter API:** 5000 tweets per person
  - more tweets needed
- **Disaster types:** wildfire and hurricane
  - will add more: earthquake, flood, ...
- **Coordinates:** may be not accurate
  - Twitter API enterprise version
- **Plotting:** screenshots only
  - Dynamic Map API