Project 3 : NLP

DSI-1128

Atigon Hongchumpol / Jan 20, 2023

Problem Statement

- Cats & Dogs photo caption contest
- The image processing team ask for help

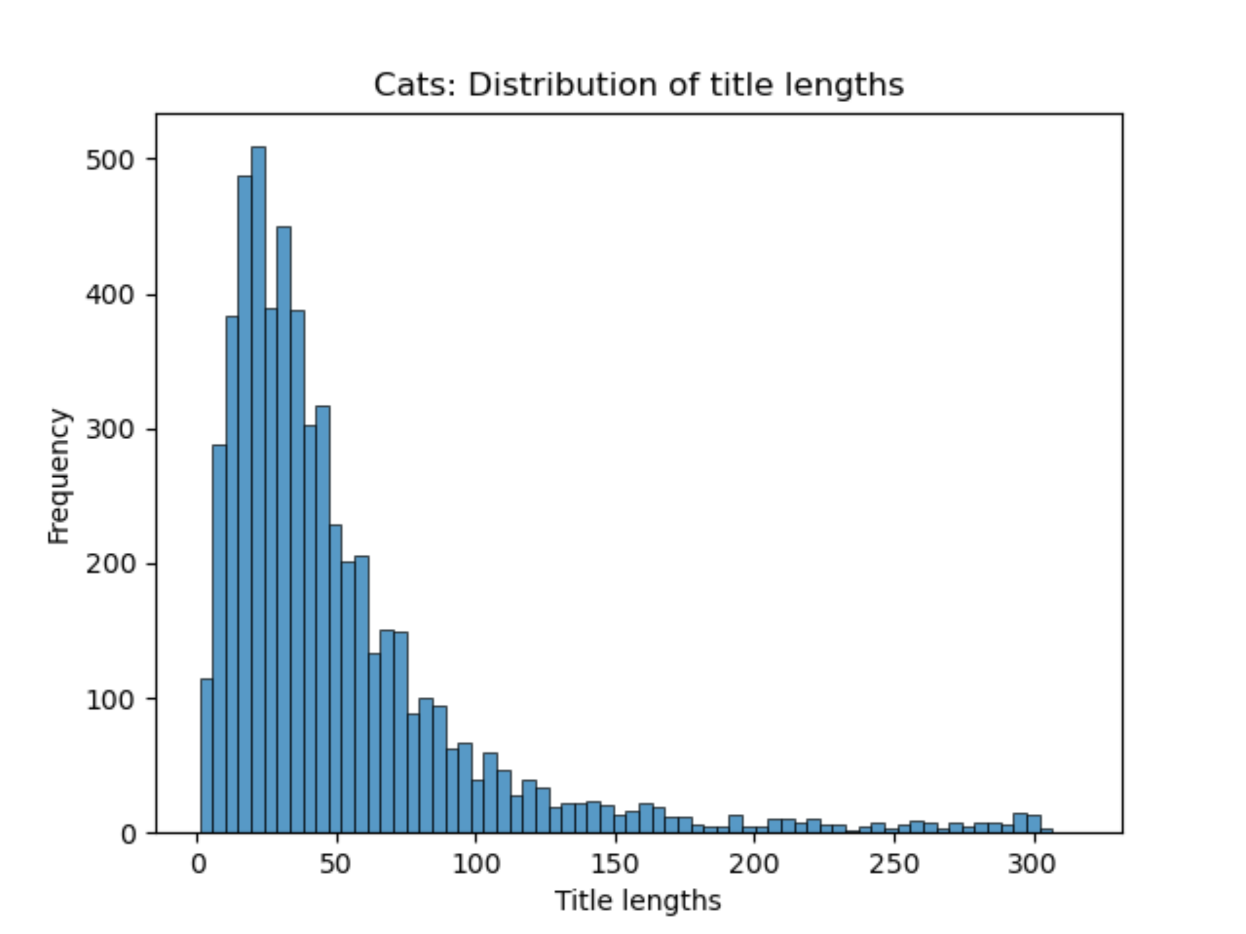
Data collection & Cleaning

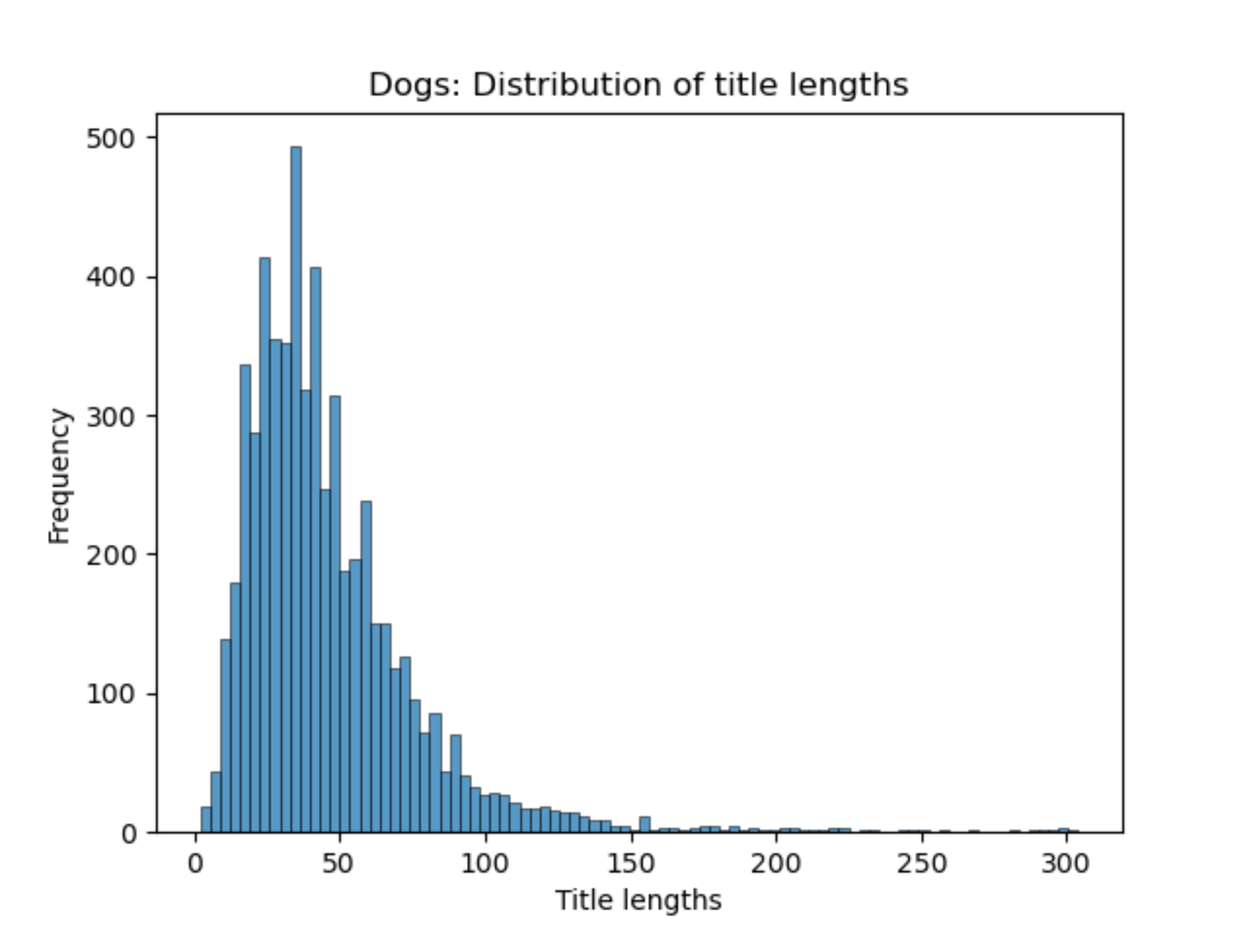
- Reddit / Pushshift API
- r/cats, r/dogs
- 5,000 + 5,000 = 10,000 posts

- Column ‘title’ = X / column ‘subreddit’ = y
- Remove duplicates title
- Remove incorrect category

EDA

Title lengths distribution





EDA

Most 5 common words using stemming

- CATS

1. cat
2. like
3. love
4. just
5. kitten

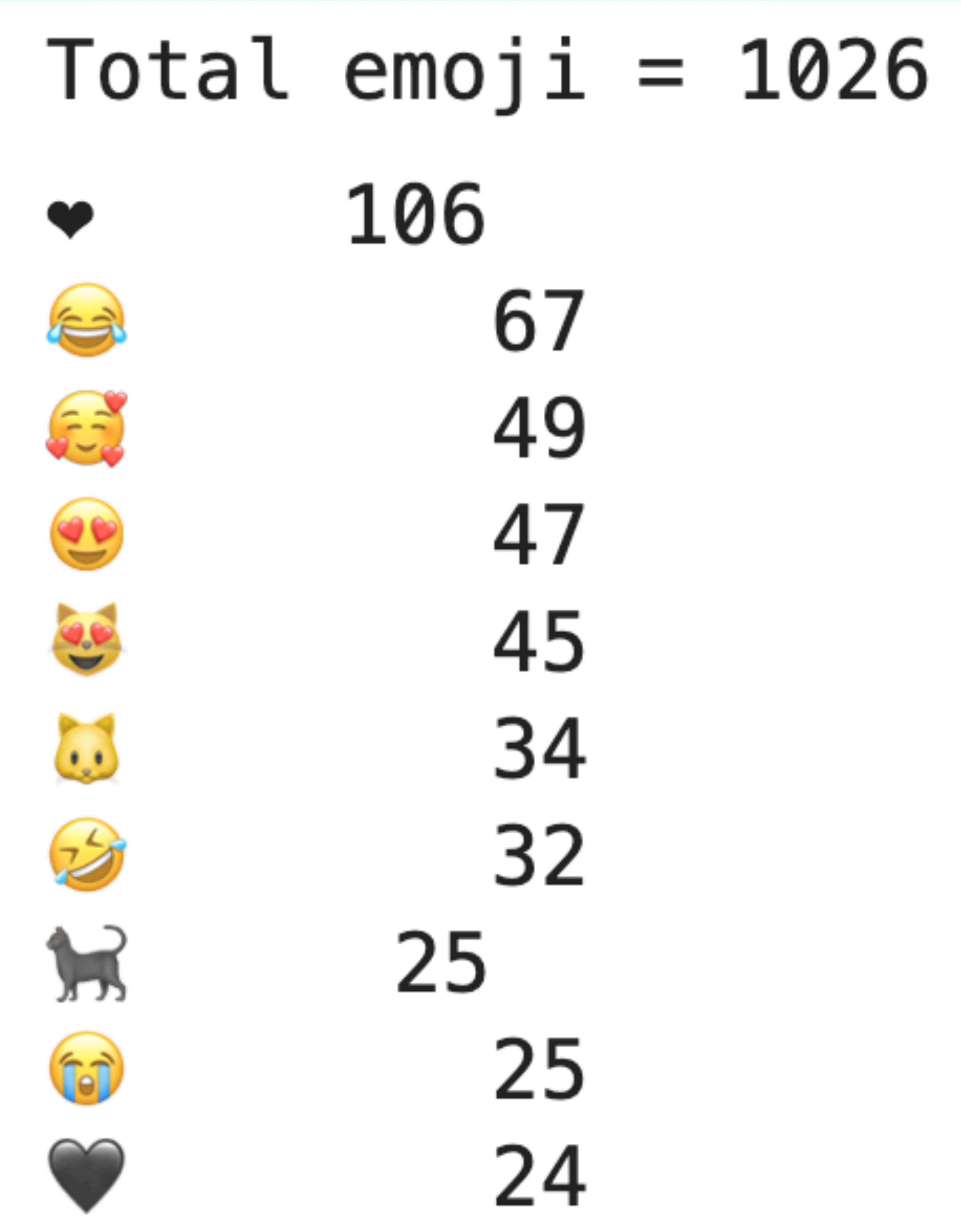
- DOGS

1. dog
2. **help**
3. puppi
4. **advic**
5. **need**

EDA

Top 10 common emoji

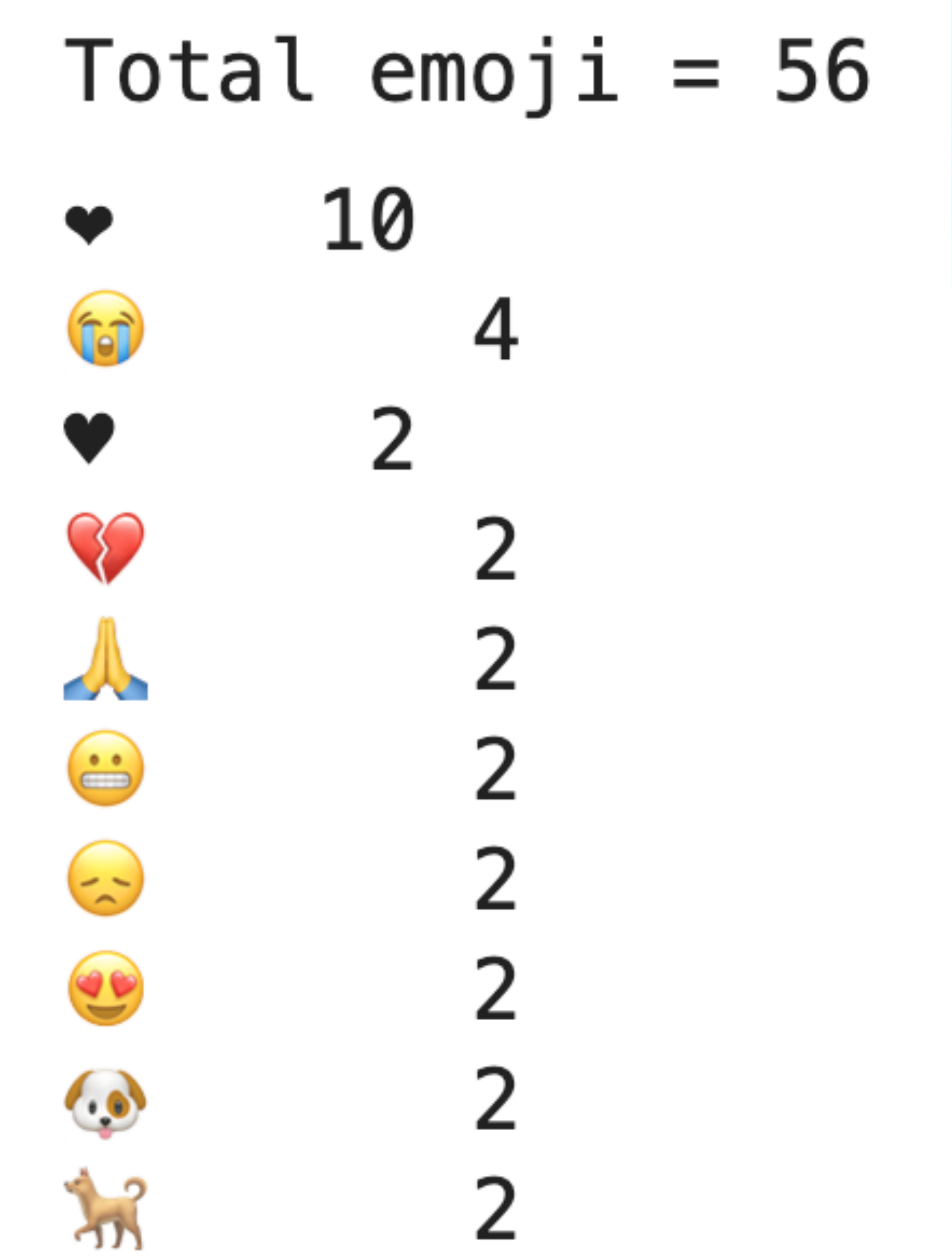
- CATS



```
Total emoji = 1026

❤      106
😂        67
🥰        49
😍        47
😻        45
🐱        34
🤣        32
🐈      25
😭        25
🖤        24
```

- DOGS



```
Total emoji = 56

❤     10
😭       4
♥     2
💔       2
🙏       2
😬       2
😞       2
😍       2
🐶       2
🐕       2
```

EDA

Sentimental Analysis using compound score

- CATS

0.12

- DOGS

0.02

Cats > Dogs
6 times!!!

Modeling

- Baseline accuracy 50%
- (CVEC, TF-IDF, STEM) + (RF, LR, KNN, NB, ADA)
- STACK best 3

Evaluation

	Train score	Test score	Recall	F-1
CVEC + RF	0.99	0.90	0.96	0.90
TFIDF + LR	0.96	0.90	0.95	0.91
CVEC + NB	0.97	0.90	0.88	0.90
CVEC+ADA	0.86	0.85	**0.98**	0.87
STACK (RF+LR+NB)	0.98	**0.93**	0.95	**0.93**

Summary

- Will deliver
  - Stack model (best test and f1 score)
  - Best recall model

Recommendation

- Bigger dataset
- New word remove technique
- Use another source / combine

Thank you