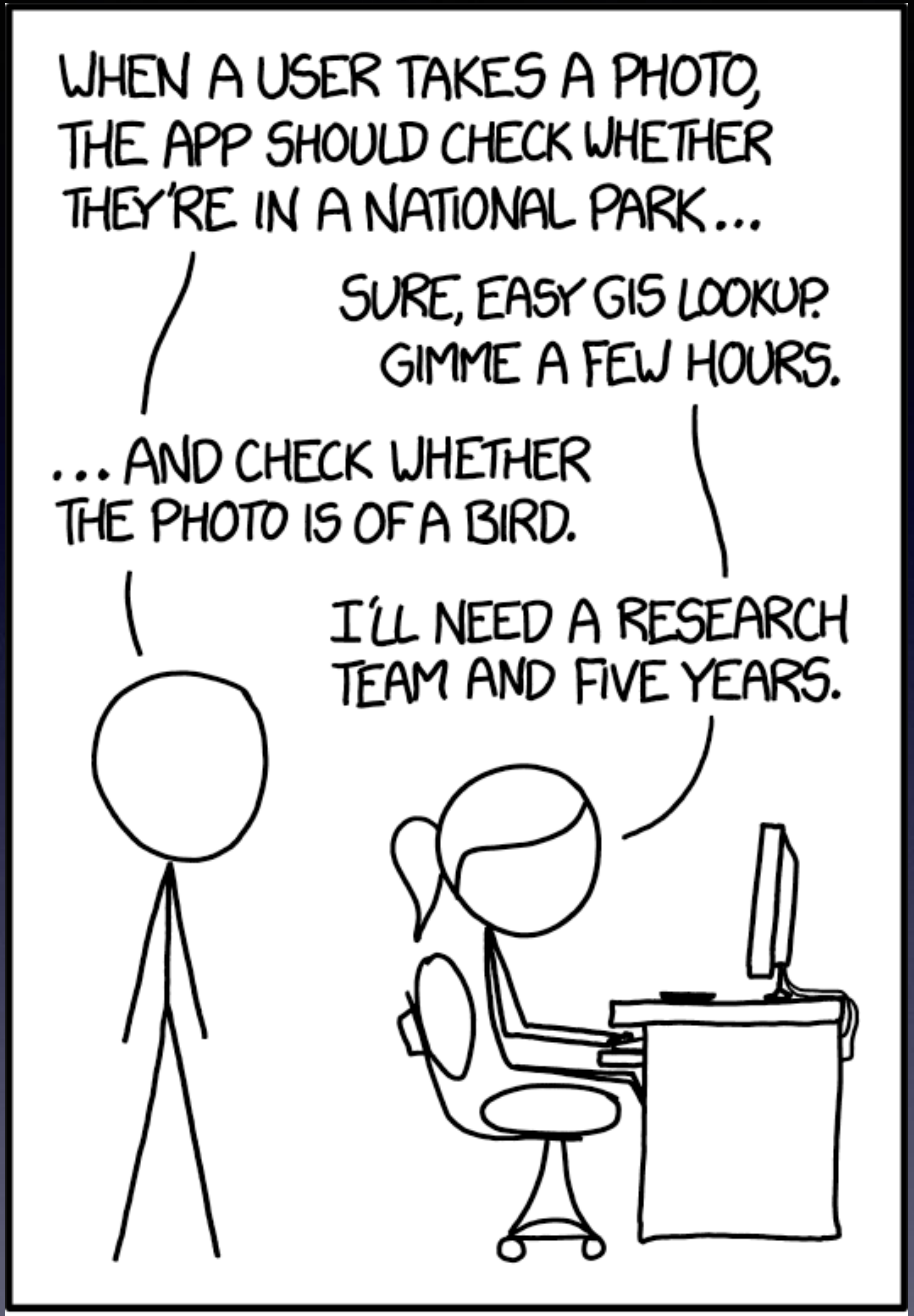
Deep Learning + Natural Language Processing

Brendan Herger, hergertarian.com

Slides: https://goo.gl/rvcLon


WHEN A USER TAKES A PHOTO, THE APP SHOULD CHECK WHETHER THEY'RE IN A NATIONAL PARK...
SURE, EASY GIS LOOKUP. GIMME A FEW HOURS.
...AND CHECK WHETHER THE PHOTO IS OF A BIRD.
I'LL NEED A RESEARCH TEAM AND FIVE YEARS.

IN CS, IT CAN BE HARD TO EXPLAIN THE DIFFERENCE BETWEEN THE EASY AND THE VIRTUALLY IMPOSSIBLE.

Intro

Word Models

Letter Models

Case Study: Spoilers

Recap

Intro

Intro

What is NLP?

- Natural language processing is the area of ML / AI focused on human languages (e.g. English or Mandarin)

Why DL + NLP?

- DL can handle high-dimensionality data with complex, non-linear relationships
- DL can map words into real-valued vectors

Why now?

- Fundamentals borrowed from computer vision
- Increasing amounts of text data

Word Models

Tokenization / preprocessing

- **Tokenize:** Convert one long string into ‘words’
- **Lemmatize:** Normalize words (e.g. running -> run, cats ->cat)
- **Pad:** Convert input into fixed length, by truncating or padding

Tokenization / preprocessing

“Running can’t be fun”

- **Tokenize:** [running, can, 't, be, fun]
- **Lemmatize:** [run, can, not, be, fun]
- **Pad:** [run, can, not, be, fun, ⍰, ⍰, ⍰, …]

Architectures

- **Embedding:** Converts tokens to numerical vectors (Word2Vec)
  - Unseen words replaced w/ ‘UNK’
- **Convolutional:** Similar to computer vision
- **RNN:** Able to ‘read’ document, one word at a time. Basic RNN, LSTM or GRU, generally bi-directional.
- **Output:** Whatever output layer(s) you want

Frameworks

Deep learning

- torchtext: PyTorch’s NLP data loaders
- keras: Treats text as 1D time series

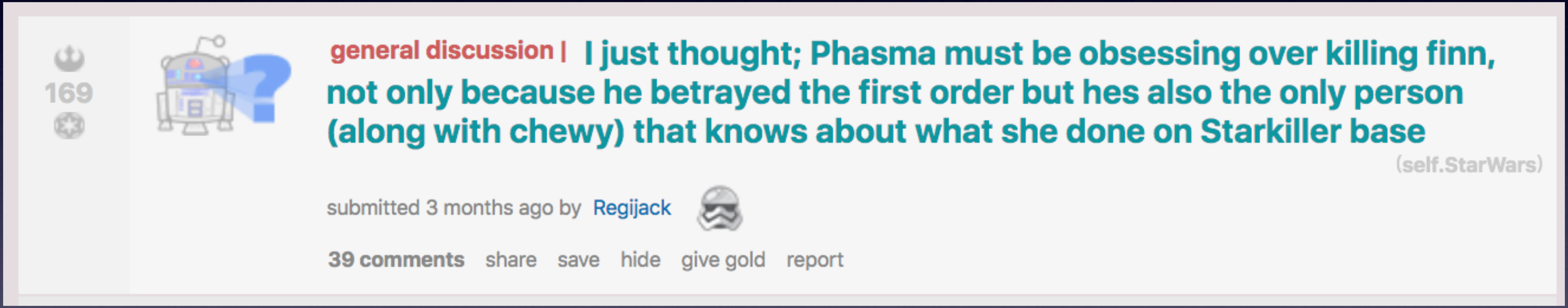
NLP

- spacy: Common framework for lemmatization, part of speech extraction
- nltk, CoreNLP (java, python interace), OpenNLP (java)

Word Models

Tokenization / preprocessing
Architectures
Frameworks

Letter Models



general discussion | I just thought; Phasma must be obsessing over killing finn, not only because he betrayed the first order but hes also the only person (along with chewy) that knows about what she done on Starkiller base (self.StarWars)

submitted 3 months ago by Regijack

https://www.reddit.com/r/StarWars/comments/7hhgis/i_just_thought_phasma_must_be_obsessing_over/

Letter Models

- **Preprocessing:** Longer padded sequences, fixed vocabulary (check your encoding!)

- **Architectures:** More / larger convolutions, due to larger sequence length. Otherwise the same

- **Frameworks:** Same

Case Study: Spoilers

Letter Models

- **Data:** Pre-labelled, textual reddit posts
- **Preprocessing:**
    - Converted text into lower case, removed non-standard characters
    - Added start and end markers
    - Padded / truncated to 2000 characters
- **Architectures:** CNN w/ Bi-directional LSTM
- **Frameworks:** Keras

https://github.com/bjherger/spoilers_model

Letter Models

Input

- How old is Chewy?

Converted text into lower case, removed non-standard characters

- how old is chewy?

Padded / truncated to 2000 characters

- [h, o, w, , o, l, d, , i, s, , c, h, e, w, y, ?, ⍰, ⍰, ⍰, …]

https://github.com/bjherger/spoilers_model

Recap

Word models
Letter models
Case Study: Spoilers

Resources

- NLP Whitepaper, by Yoav Goldberg
- Deep Learning, book by Ian Goodfellow and Yoshua Bengio and Aaron Courville
- Introduction to LSTMs, by Christopher Olah
- LSTM / GRU intro & comparison

Thanks!

Brendan Herger, hergertarian.com
Slides: https://goo.gl/rvcLon