Data Processing - Re-examine

Anthony Zhang

Viterbi School of Engineering, USC

2020

Table of Contents

Data Cleaning

Idea

- Remove [deleted], [removed], and NAs
- Remove links
- Fix encoding problems

```
>>> print(fix_encoding("(à¸‡'âŒ£')à¸‡"))
(ง'⌣')ง
```

- Strip spaces
- Only keep English posts

Results

Very clean data. Each post should only have what they actually said.

Document Embedding

Idea

- Instead of averaging sentences embeddings, use Recurrent Neural Network (RNN) like LSTM on sentences
- Extract the features in the last layer to represent document embeddings
- Don't need to fit the model, just compile it.

Another Idea

Idea

- Construct a small corpus of sentences that describes what we feel when a story sounds creepy.
- Extract all the maximum similarities to create a creepiness vector

Problem

All vectors may have different sizes.

Document Embedding

Query: A man held a knife.

Top 5 most similar sentences in corpus:

I feel chilling. (Score: 0.2185)

I feel scaried. (Score: 0.2165)

I feel creepy. (Score: 0.1618)

I feel terrifying. (Score: 0.1586)

I feel frightening. (Score: 0.1543)

Query: A woman held a knife.

Top 5 most similar sentences in corpus:

I feel chilling. (Score: 0.2274)

I feel scaried. (Score: 0.2112)

I feel terrifying. (Score: 0.1812)

I feel creepy. (Score: 0.1748)

I feel frightening. (Score: 0.1591)

Query: Apple tress can grow as tall as 20 feet.

Top 5 most similar sentences in corpus:

I feel frightening. (Score: -0.0347)

I feel terrifying. (Score: -0.0469)

I feel scaried. (Score: -0.0677)

I feel chilling. (Score: -0.0813)

I feel creepy. (Score: -0.1176)