The Joy of Character level NLP

Tal Perry - CEO LightTag
tal@lighttag.io
@thetalperry

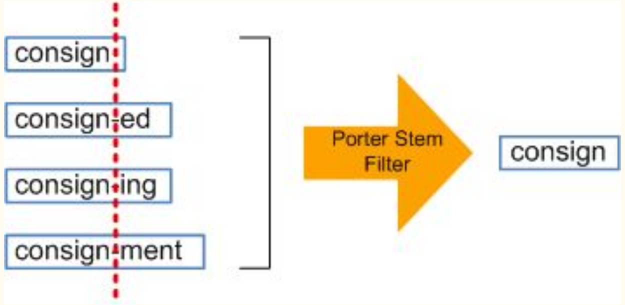
Character level NLP

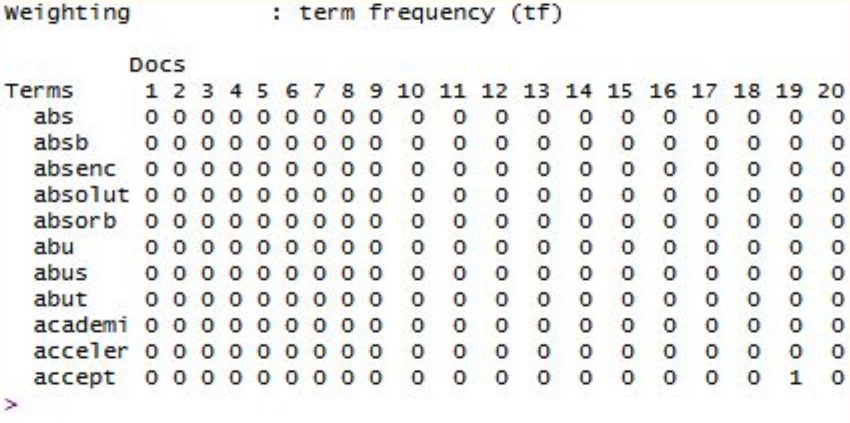
All the cool kids are doing it

- Background
- Why ?
- Benefits
- Drawbacks
- Example

Background

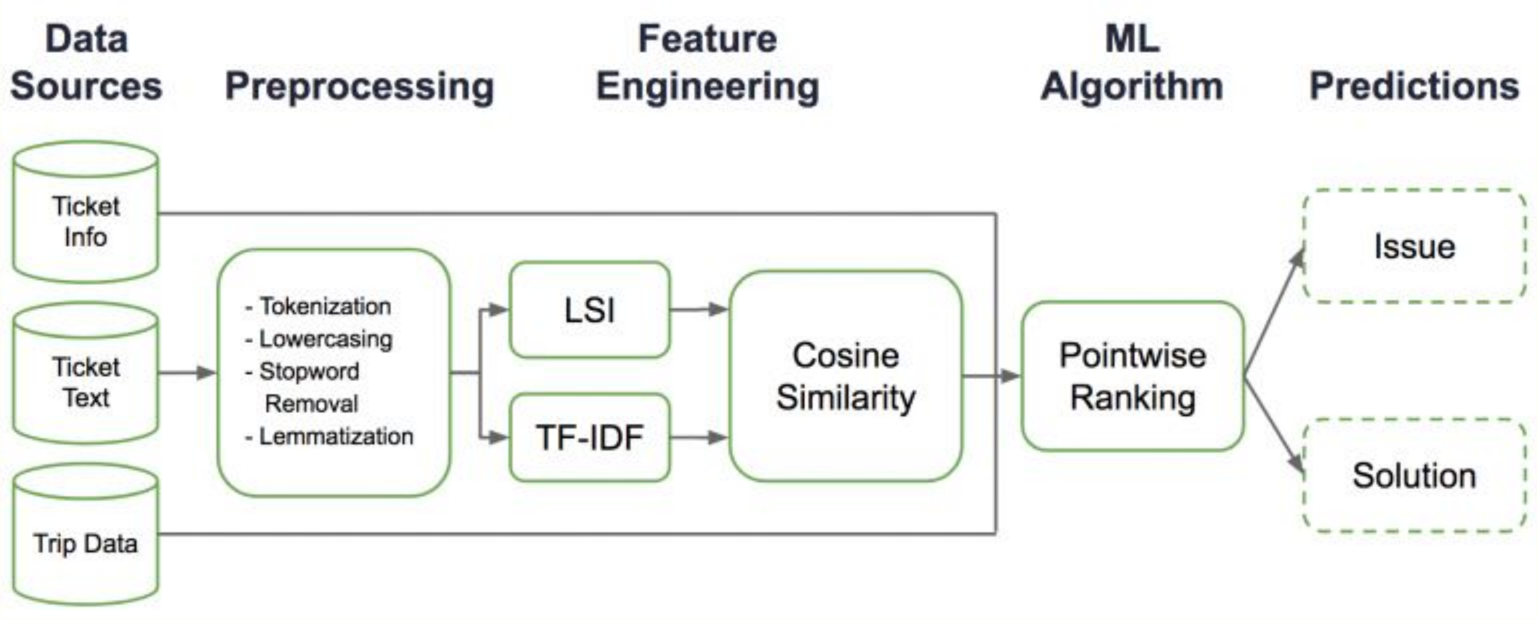
The Old *Old* Way



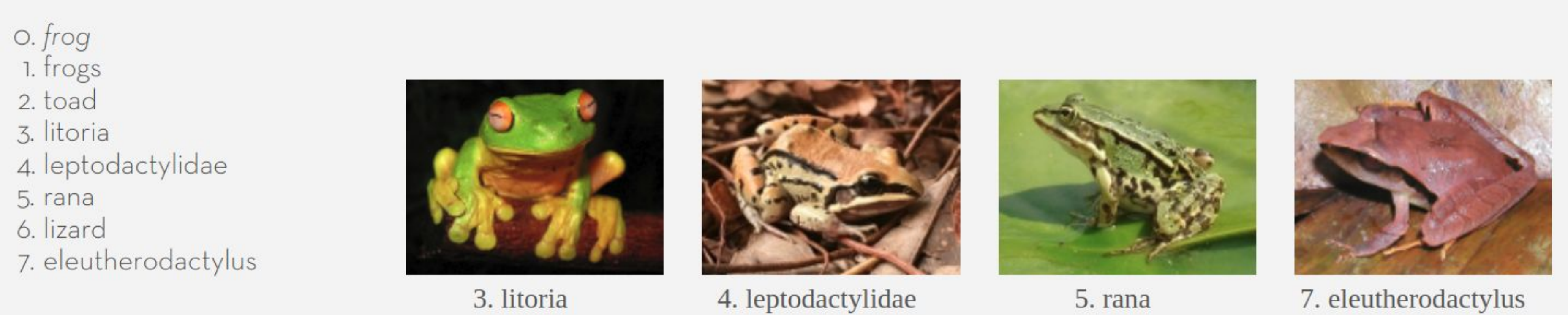


```
Weighting           : term frequency (tf)

         Docs
Terms     1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
  abs     0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
  absb    0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
  absenc  0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
  absolut 0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
  absorb  0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
  abu     0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
  abus    0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
  abut    0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
  academi 0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
  acceler 0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
  accept  0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  1  0
>
```

The Old Old Way



DL Insight 1: Word Vectors



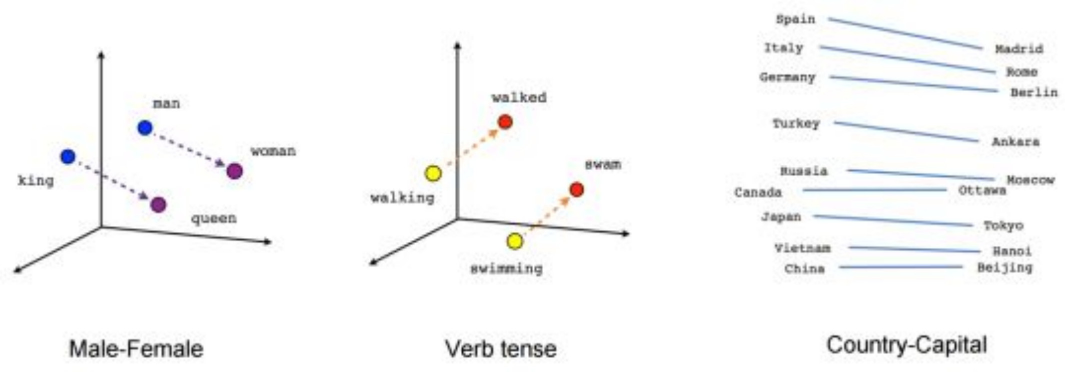
0. *frog*
1. frogs
2. toad
3. litoria
4. leptodactylidae
5. rana
6. lizard
7. eleutherodactylus

3. litoria

4. leptodactylidae

5. rana

7. eleutherodactylus

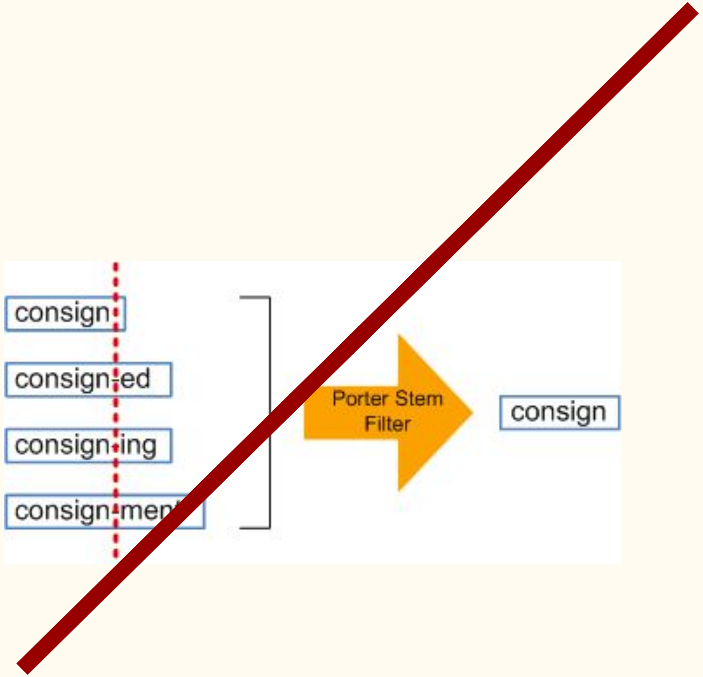


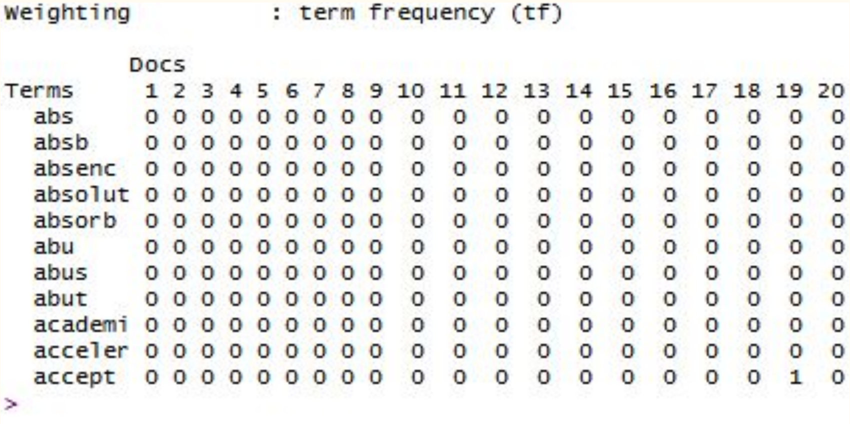
Male-Female

Verb tense

Country-Capital

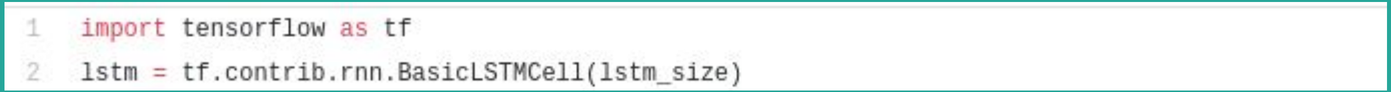
The Old *Old* Way





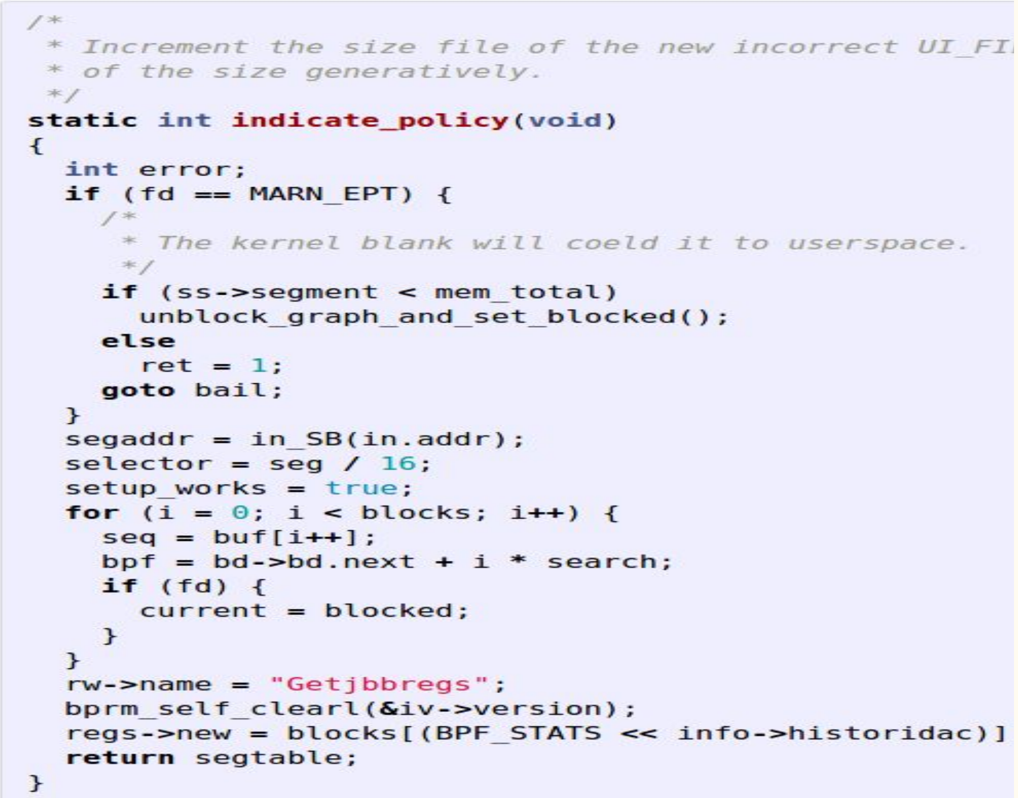
```
Weighting            : term frequency (tf)

          Docs
Terms      1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20
  abs      0  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0   0   0
  absb     0  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0   0   0
  absenc   0  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0   0   0
  absolut  0  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0   0   0
  absorb   0  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0   0   0
  abu      0  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0   0   0
  abus     0  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0   0   0
  abut     0  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0   0   0
  academi  0  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0   0   0
  acceler  0  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0   0   0
  accept   0  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0   1   0
>
```

Deep Learning made it easy to work with long sequences

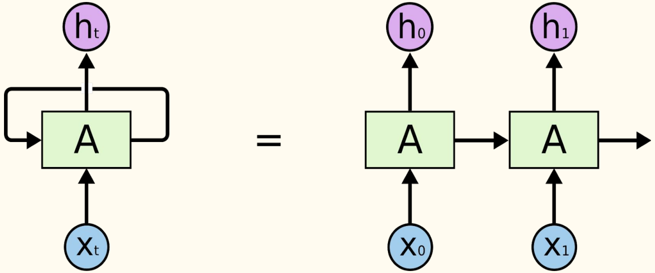


```python
import tensorflow as tf
lstm = tf.contrib.rnn.BasicLSTMCell(lstm_size)
```

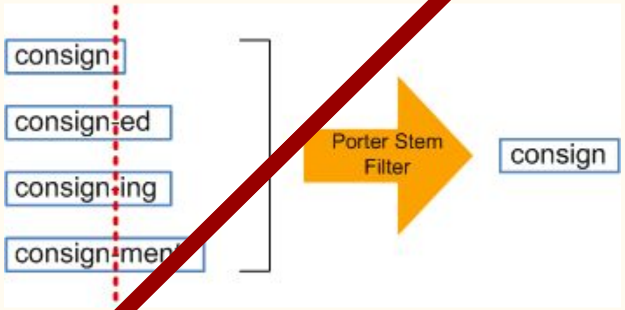
RNNs

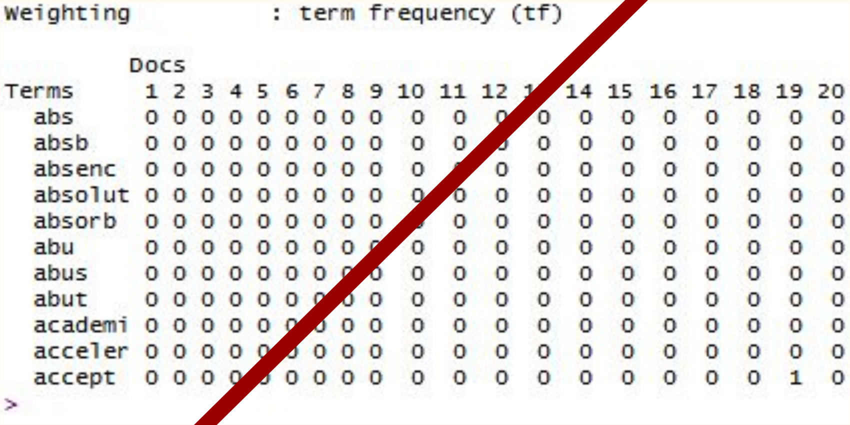


```
/*
 * Increment the size file of the new incorrect UI_FI
 * of the size generatively.
 */
static int indicate_policy(void)
{
  int error;
  if (fd == MARN_EPT) {
    /*
     * The kernel blank will coeld it to userspace.
     */
    if (ss->segment < mem_total)
      unblock_graph_and_set_blocked();
    else
      ret = 1;
    goto bail;
  }
  segaddr = in_SB(in.addr);
  selector = seg / 16;
  setup_works = true;
  for (i = 0; i < blocks; i++) {
    seq = buf[i++];
    bpf = bd->bd.next + i * search;
    if (fd) {
      current = blocked;
    }
  }
  rw->name = "Getjbbregs";
  bprm_self_clearl(&iv->version);
  regs->new = blocks[(BPF_STATS << info->historidac)]
  return segtable;
}
```



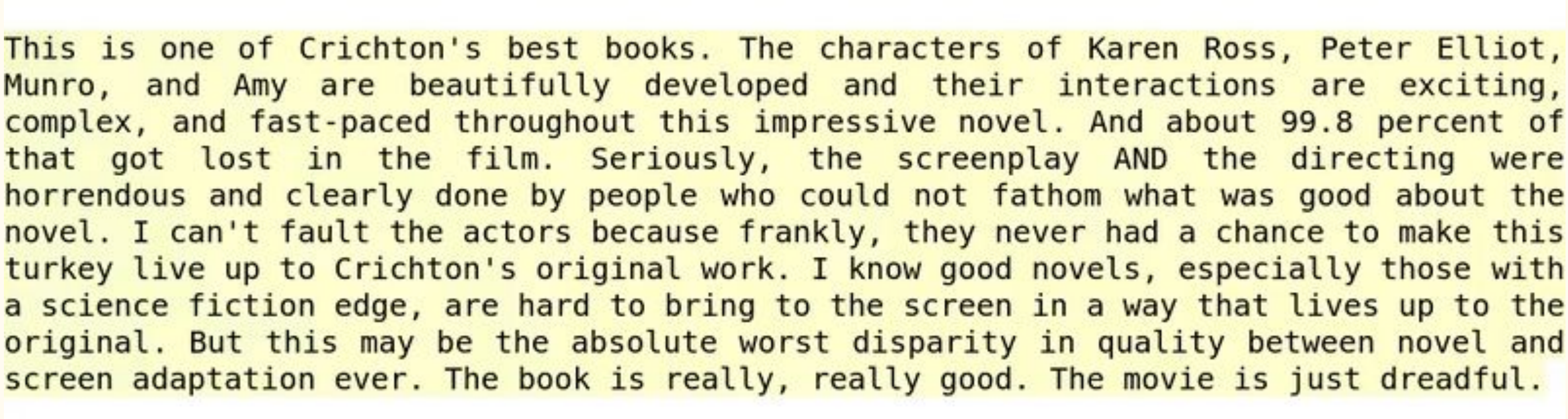
The Old *Old* Way



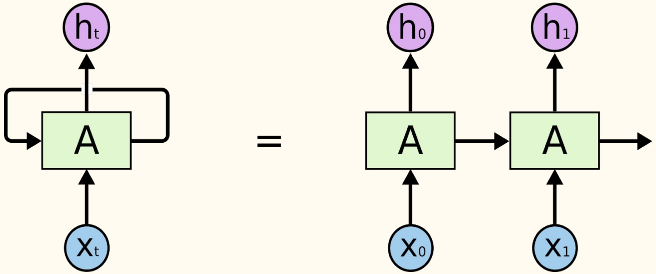


```
Weighting          : term frequency (tf)

         Docs
Terms     1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
  abs     0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
  absb    0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
  absenc  0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
  absolut 0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
  absorb  0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
  abu     0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
  abus    0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
  abut    0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
  academi 0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
  acceler 0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
  accept  0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  1  0
>
```

More RNNs

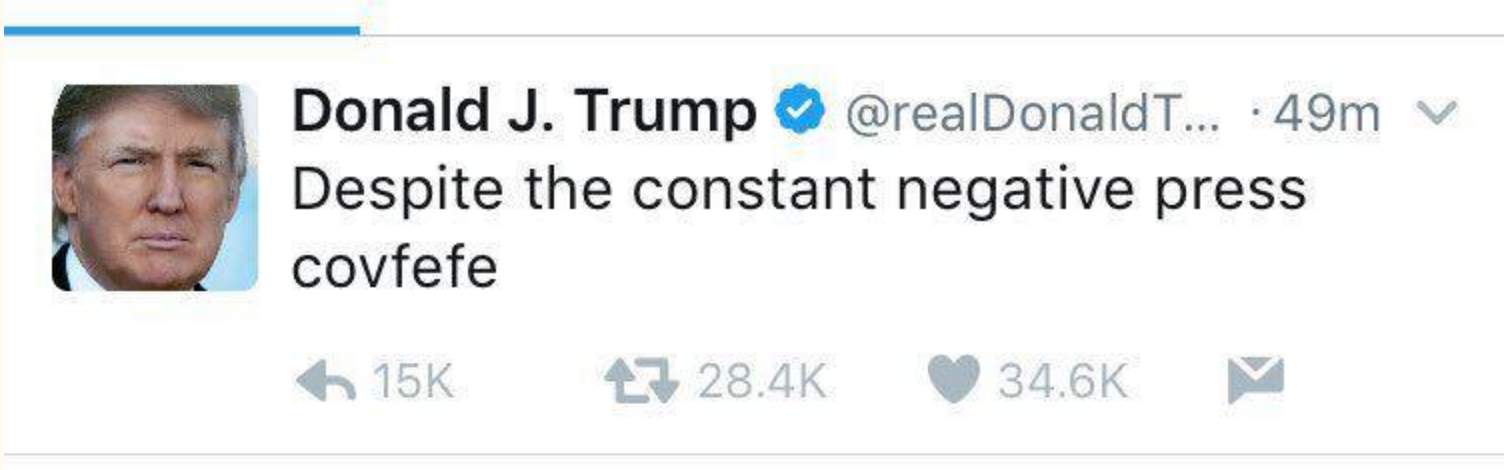


This is one of Crichton's best books. The characters of Karen Ross, Peter Elliot, Munro, and Amy are beautifully developed and their interactions are exciting, complex, and fast-paced throughout this impressive novel. And about 99.8 percent of that got lost in the film. Seriously, the screenplay AND the directing were horrendous and clearly done by people who could not fathom what was good about the novel. I can't fault the actors because frankly, they never had a chance to make this turkey live up to Crichton's original work. I know good novels, especially those with a science fiction edge, are hard to bring to the screen in a way that lives up to the original. But this may be the absolute worst disparity in quality between novel and screen adaptation ever. The book is really, really good. The movie is just dreadful.



Why Characters ?

But there are many “words”



Donald J. Trump @realDonaldT... · 49m

Despite the constant negative press covfefe

15K 28.4K 34.6K

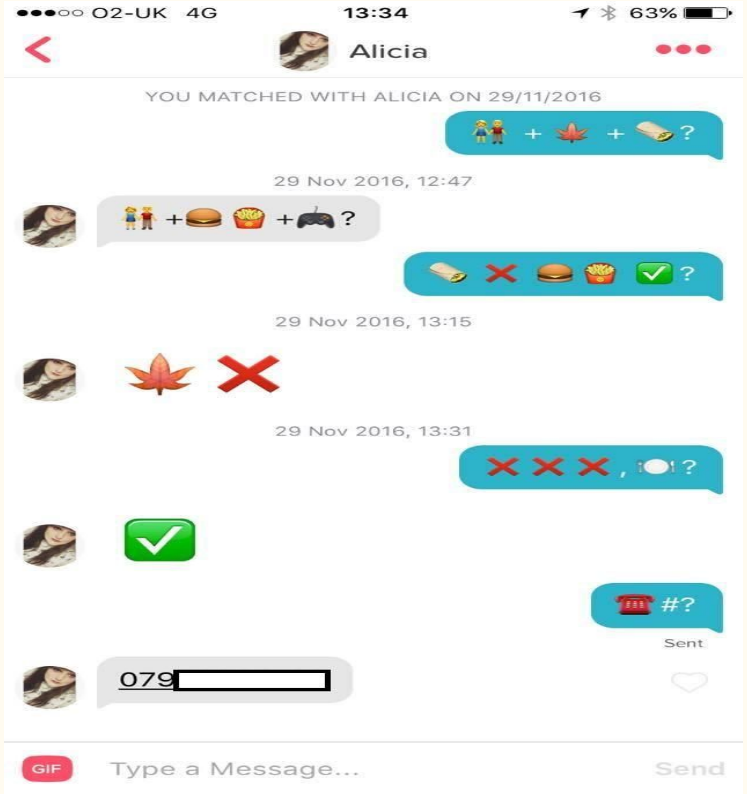
With many meanings

What has four letters, never has five letters but sometimes has nine letters.

With many meanings

What has four letters, never has five letters but sometimes has nine letters?

WTF ?



YOU MATCHED WITH ALICIA ON 29/11/2016

👫 + 🍁 + 🌯 ?

29 Nov 2016, 12:47

👫 + 🍔 🍟 + 🎮 ?

🌯 ❌ 🍔 🍟 ✅ ?

29 Nov 2016, 13:15

🍁 ❌

29 Nov 2016, 13:31

❌ ❌ ❌ , 🍽 ?

✅

☎ #?

Sent

079[redacted]

Infinite

Vocabulary

- Expanding Vocabulary
- Many Professional Terms
- Lots of numbers / codes
- User generated data

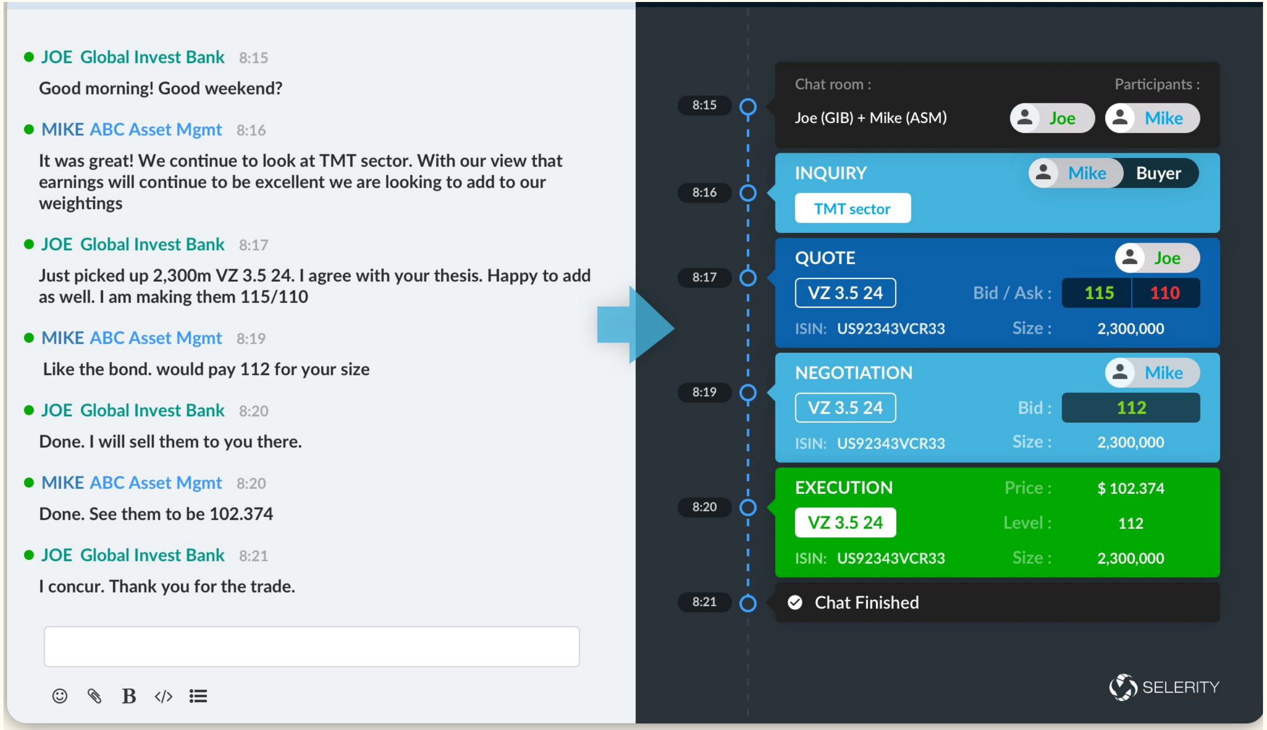
Infinite

Vocabulary

10/30/04: Cont. C&T. Drld cmt plug f/5693' to 5707'. Performed LOT to 960 psi = 15.6 ppg. Drld to 5875'.

Infinite

Vocabulary



● JOE Global Invest Bank 8:15

Good morning! Good weekend?

● MIKE ABC Asset Mgmt 8:16

It was great! We continue to look at TMT sector. With our view that earnings will continue to be excellent we are looking to add to our weightings

● JOE Global Invest Bank 8:17

Just picked up 2,300m VZ 3.5 24. I agree with your thesis. Happy to add as well. I am making them 115/110

● MIKE ABC Asset Mgmt 8:19

Like the bond. would pay 112 for your size

● JOE Global Invest Bank 8:20

Done. I will sell them to you there.

● MIKE ABC Asset Mgmt 8:20

Done. See them to be 102.374

● JOE Global Invest Bank 8:21

I concur. Thank you for the trade.

8:15 — Chat room : Joe (GIB) + Mike (ASM) — Participants : Joe, Mike

8:16 — INQUIRY — Mike Buyer — TMT sector

8:17 — QUOTE — Joe — VZ 3.5 24 — Bid / Ask : 115 110 — ISIN: US92343VCR33 — Size : 2,300,000

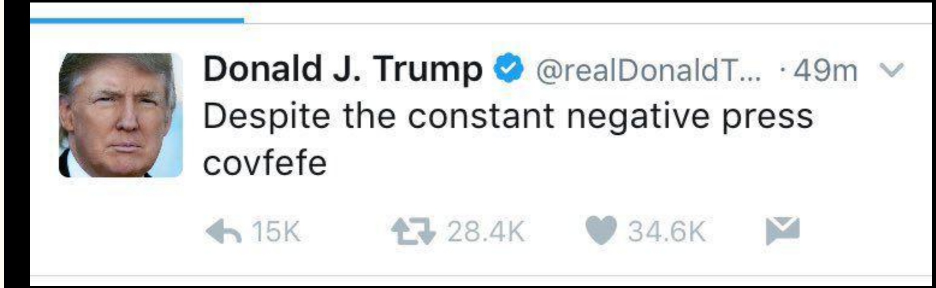
8:19 — NEGOTIATION — Mike — VZ 3.5 24 — Bid : 112 — ISIN: US92343VCR33 — Size : 2,300,000

8:20 — EXECUTION — VZ 3.5 24 — Price : $ 102.374 — Level : 112 — ISIN: US92343VCR33 — Size : 2,300,000

8:21 — ✓ Chat Finished

Infinite

Vocabulary



Donald J. Trumpⓥ @realDonaldT... · 49m

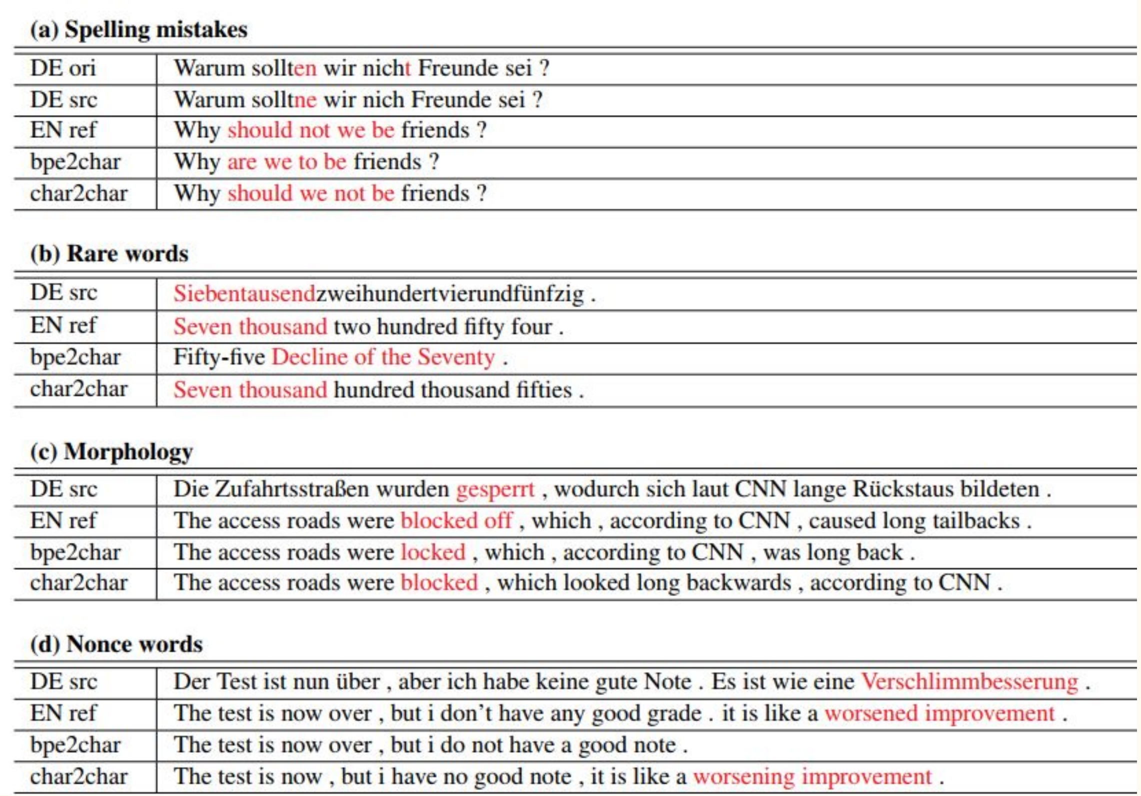
Despite the constant negative press covfefe

Operational

Simplicity

- No need to carry a dictionary
- Less preprocessing
- Less room for error

In Action



(a) Spelling mistakes

DE ori	Warum sollten wir nicht Freunde sei ?
DE src	Warum solltne wir nich Freunde sei ?
EN ref	Why should not we be friends ?
bpe2char	Why are we to be friends ?
char2char	Why should we not be friends ?

(b) Rare words

DE src	Siebentausendzweihundertvierundfünfzig .
EN ref	Seven thousand two hundred fifty four .
bpe2char	Fifty-five Decline of the Seventy .
char2char	Seven thousand hundred thousand fifties .

(c) Morphology

DE src	Die Zufahrtsstraßen wurden gesperrt , wodurch sich laut CNN lange Rückstaus bildeten .
EN ref	The access roads were blocked off , which , according to CNN , caused long tailbacks .
bpe2char	The access roads were locked , which , according to CNN , was long back .
char2char	The access roads were blocked , which looked long backwards , according to CNN .

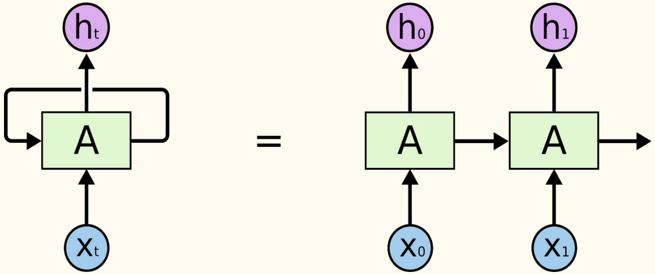
(d) Nonce words

DE src	Der Test ist nun über , aber ich habe keine gute Note . Es ist wie eine Verschlimmbesserung .
EN ref	The test is now over , but i don't have any good grade . it is like a worsened improvement .
bpe2char	The test is now over , but i do not have a good note .
char2char	The test is now , but i have no good note , it is like a worsening improvement .

Drawbacks

Long Sequences

Mean more computation



- RNNs are O(n) in sequence length
- The average english word has ~5 letters
- So 5x more expensive

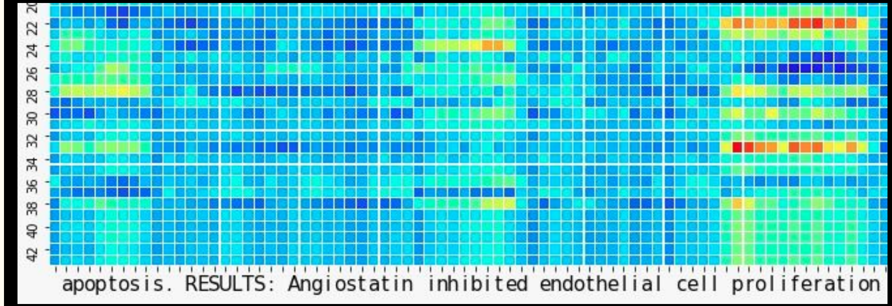
Meaningless

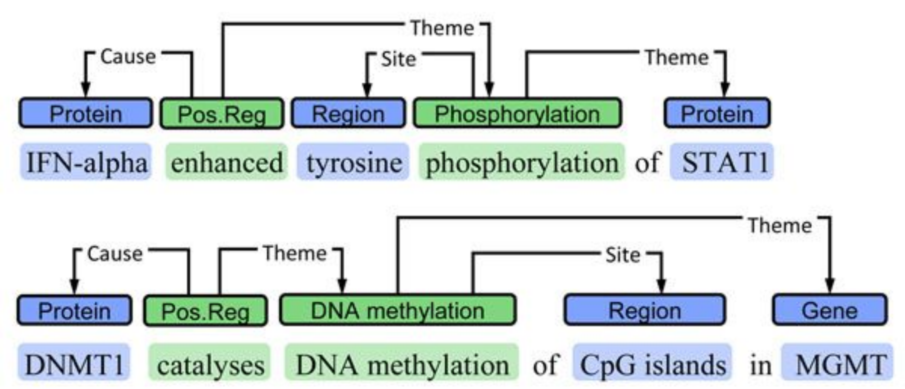
Semantic Unit

- Characters have no semantic meaning
- Its harder to form relations between a bunch of meaningless things

Meaningless

Semantic Unit

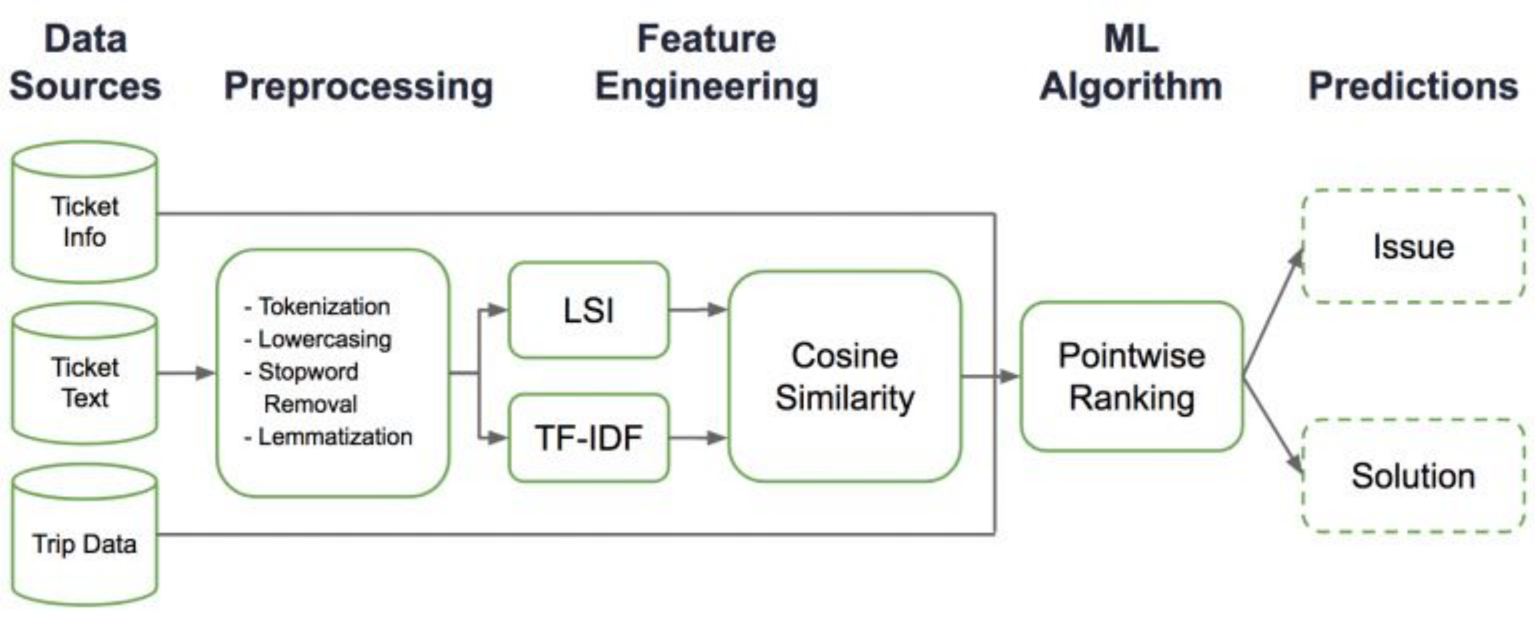




Example

https://github.com/LightTag/BibSample

Conclusion



Data Sources
Preprocessing
Feature Engineering
ML Algorithm
Predictions
Ticket Info
Ticket Text
Trip Data
- Tokenization
- Lowercasing
- Stopword Removal
- Lemmatization
LSI
TF-IDF
Cosine Similarity
Pointwise Ranking
Issue
Solution

The End

Anime44.Com
King of Deep Learning
HOW YOU FEEL WHEN YOU FINISH
THE DEEP LEARNING SPECIALIZATION IN COURSERA

Questions ?