Language Hackday

Summer of Hacks 2019

Code of Conduct

https://github.com/summer-of-hacks/soh19website/blob/master/CODEOFCONDUCT.md

Adam Leskis - leskis@gmail.com / @aleskis (slack) / @BaronVonLeskis (twitter)

Rich Douglas - rich.douglas.evans@gmail.com / @rich (slack) / @richdevans (twitter)

Organiser Account - @Ox_SOH (twitter...DM’s open!)

Sponsors

Health and Safety

Fire exits

Toilets are in the toilet place

Plan for the Day

- 10:10 - Welcome and housekeeping
- 10:25 - Team up, pair up, or go it alone!
- 12:30 - Break for lunch
- 1:30 - Recap of current projects/progress
- 3:30 - Present what you did (...if you wanna)

Three possible “tracks”

1) No programming experience, no NLP experience
   - Install and get familiar with Python
   - Do some basic stuff
2) Programming experience, no NLP experience
   - Investigate specific libraries for NLP
   - Apply NLP techniques to do something cool
3) Programming experience, NLP experience
   - Work on some super cool stuff!
   - Adam and Rich can *probably* answer questions you have about your own projects

Track 1 - no programming exp, no NLP exp

WEB SCRAPE

1) find text source (probably website here)

2) make a request (using requests/beautifulsoup/scrapy/etc?)

3) get only the text

4) save to file

Track 1 - no programming exp, no NLP exp

STRING PROCESSING

1) find text source (website/file/etc?)

2) load into memory

3) replace one word with a different word

4) save to file

Track 1 - no programming exp, no NLP exp

CLEAN TEXT

1) find text source (website/file/etc?)

2) convert to utf-8

3) remove all non-alphas

4) trim all extra spaces

5) save to file

Track 1 - no programming exp, no NLP exp

COUNT WORDS

1) find text source (website/file/etc?)

2) split words into a list

3) count total words (tokens)

4) find most common words (frequency distribution)

5) turn words into a set

6) count total unique words (types)

Sample tutorials

https://www.nltk.org/book/ch01.html

NLTK (basic tutorial) - Python

https://nlpforhackers.io/complete-guide-to-spacy/

Spacy (intermediate tutorial) - Python

https://stanfordnlp.github.io/CoreNLP/tutorials.html

Stanford CoreNLP (advanced tutorial) - Java

https://github.com/kbarry91/Eliza-ChatBot

Chatbot adventure game

Track 2 - programming exp, no NLP exp

MAKE RHYMES

1) turn utf-8 into IPA

2) get IPA vowels and compare them

3) find all the single syllable rhyme words in a text (on a website? Forum post?)

4) POS tag these rhymes and split into NP, VP, AJ

5) based on the part-of-speech, create templates a la madlibs to input correct form

(eg, All the time I like to _____, I like to _____ out in the _____...VP, VP, NP)

--stretch goal: tweet this rhyme from the command line

Track 2 - programming exp, no NLP exp

HOW ACADEMIC IS IT?

1) get a sample text (either from disk or from the internet)

2) use a library like awlify to check how many of the words in each sentence come from the Academic Word List

3) count which percentage of the words in the text are “academic” (good intro in the nltk book)

4) use these counts to compare texts from several different genres (fiction, news, arXiv papers) and see if you can guess which percentage of each is “academic” words.

--stretch goal: rank the trending tweets of the day in order of “academicness”

Possible activities

https://github.com/lpmi-13/simple-NLP-pos

Practice doing some simple POS tagging

https://github.com/lpmi-13/simple-NLP-stats

Practice doing some simple frequency counts

https://github.com/lpmi-13/rhyme-check

Practice working with rhymes

https://github.com/lpmi-13/tinderflow-python

Visualise how a machine learning model uses NLP features to make predictions

Track 3 - programming exp, NLP exp

You probably have an idea of something you'd like to work on, or are already working on stuff.

Feel free to chat to Adam and/or Rich if you have any questions or want feedback.

Workshops

We'll decide on one of these to work through at 2:30PM if anyone's interested... join if you like!

turning text into NLP objects

finding parts of speech

sentiment analysis

Let’s get to languaging!

If you would like to be in a team but haven't found one yet, let Rich or Adam know and we can help you find one.

Break for lunch

Eat some food here, or outside if you like

We'll reconvene here at 1:30PM

...and we're back

Workshop at 2:30

Show and Tell at 3:30

Presentations

- Showing and telling