Attics, guesswork and clay. Sleuthing your way into Biomedical Natural Language Processing

Seth Russell, MS

University of Colorado Anschutz Medical Campus ACCORDS Data Science Program, D2V Data Science Core

COPRH Con

Colorado Pragmatic Research in Health Conference

ACCORDS
UNIVERSITY OF COLORADO
CHILDREN'S HOSPITAL COLORADO

Colorado Clinical and Translational Sciences Institute (CCTSI)
UNIVERSITY OF COLORADO DENVER | ANSCHUTZ MEDICAL CAMPUS

https://github.com/magic-lantern/coprh-nlp-2021

Learning Objectives

- A general idea of what Natural Language Processing (NLP) is.
- A knowledge of the ethical implications of data reuse in NLP.
- Knowledge of where to get text for NLP.
- How to use and learn more about some basic NLP techniques.

https://github.com/magic-lantern/coprh-nlp-2021

WATSON WINS JEOPARDY!

Jeopardy! champion Ken Jennings points to his IBM supercomputer opponent, Watson, during a practice round for the TV game show last month. Jennings and fellow human contestant Brad Rutter competed against Watson in a three-episode tournament this week in the U... **Read More**

PHOTOGRAPH BY SETH WENIG, AP

Watson Wins Jeopardy!—6 Artificial Intelligence Milestones

IBM's Watson seemingly came from out of nowhere to win Jeopardy! But the computer is just the latest artificial intelligence sensation.

The New York Times

Meet GPT-3. It Has Learned to Code (and Blog and Argue).

The latest natural-language system generates tweets, pens poetry, summarizes emails, answers trivia questions, translates languages and even writes its own computer programs.

THE WALL STREET JOURNAL.

WSJ NEWS EXCLUSIVE | TECH

IBM Explores Sale of IBM Watson Health

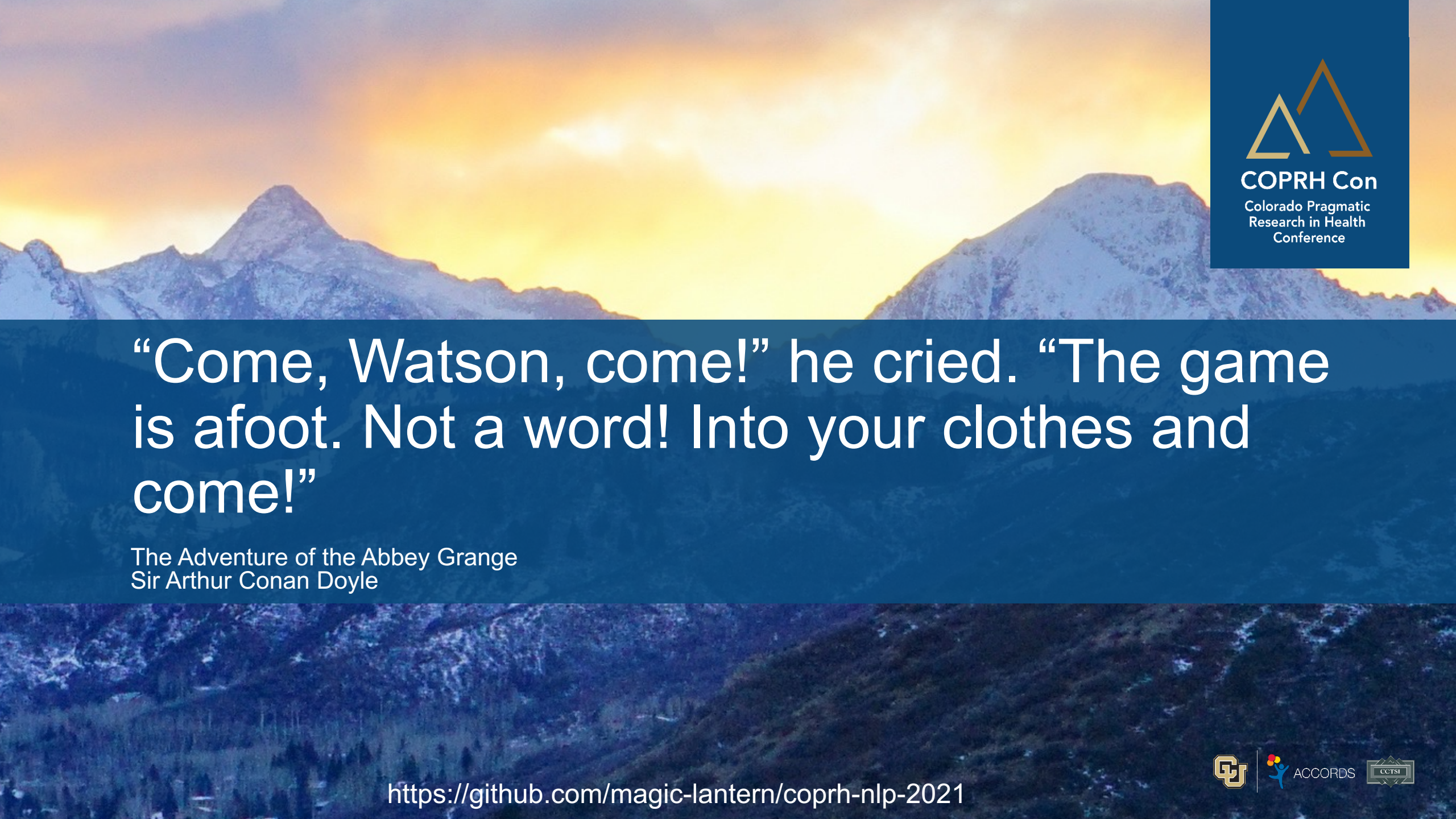
Business has roughly $1 billion in revenue and isn't profitable, sources say

IBM's Watson Health unit employs artificial intelligence to help hospitals, insurers and drugmakers manage their data.

PHOTO: SUZANNE KREITER/THE BOSTON GLOBE/GETTY IMAGES

By Laura Cooper and Cara Lombardo
Updated Feb. 18, 2021 8:21 pm ET

International Business Machines Corp. IBM -1.12% ▼ is exploring a potential sale of its IBM Watson Health business, according to people familiar with the matter, as the technology giant's new chief executive moves to streamline the company and become more competitive in cloud computing.


COPRH Con
Colorado Pragmatic
Research in Health
Conference
“Come, Watson, come!” he cried. “The game is afoot. Not a word! Into your clothes and come!”
The Adventure of the Abbey Grange
Sir Arthur Conan Doyle
ACCORDS
CCTSI
https://github.com/magic-lantern/coprh-nlp-2021

So, what is Natural Language Processing ?

- Document Retrieval
- Information Extraction
- Knowledge Representation
- Word/Concept/Abbreviation disambiguation
- Automated reasoning
- Classification
- Sentiment Analysis

https://github.com/magic-lantern/coprh-nlp-2021

Google

smithsonian hours

All | News | Maps | Shopping | Images | More | Settings | Tools

About 15,800,000 results (1.07 seconds)

Rating | Hours | Your past visits

Smithsonian Institution Offices

4.5 ★★★★½ (277) · Research institute

600 Maryland Ave SW · In Capital Gallery

8:30AM–5:30PM

Vast museum, zoo & research complex

Smithsonian National Air and Space Museum

4.7 ★★★★½ (35,509) · National museum

600 Independence Ave SW

Temporarily closed

Exhibit-filled trip across the universe

COPRH Con
Colorado Pragmatic Research in Health Conference

ACCORDS

CCTSI

Technology

YouTube says it's getting better at taking down videos that break its rules. They still number in the millions.

The Google-owned site is blocking millions of videos that contain hate speech and disinformation, but researchers say there's more it could do

A demonstrator listens to a speech during a Day of Solidarity event on Saturday at the DeKalb County Court House in Auburn, Ind. (Jon Cherry/Getty Images)

By **Gerrit De Vynck**

April 6, 2021 at 9:55 a.m. MDT

YouTube released data on Tuesday arguing that it is getting better at spotting and removing videos that break its rules against disinformation, hate speech and other banned content.

I consider that a man's brain originally is like a little empty attic, and you have to stock it with such furniture as you choose.

A fool takes in all the lumber of every sort that he comes across, so that the knowledge which might be useful to him gets crowded out, or at best is jumbled up with a lot of other things so that he has a difficulty in laying his hands upon it. Now the skilful workman is very careful indeed as to what he takes into his brain-attic. He will have nothing but the tools which may help him in doing his work, but of these he has a large assortment, and all in the most perfect order. It is a mistake to think that that little room has elastic walls and can distend to any extent. Depend upon it there comes a time when for every addition of knowledge you forget something that you knew before. It is of the highest importance, therefore, not to have useless facts elbowing out the useful ones.

A Study In Scarlet
Sir Arthur Conan Doyle

Some common NLP techniques

- Regular expressions
- Syntactical Analysis
- Stemming
- Lemmatization
- Stop word removal
- Word to vector representations
- Deep Learning & Language Models

https://github.com/magic-lantern/coprh-nlp-2021

Regular Expressions

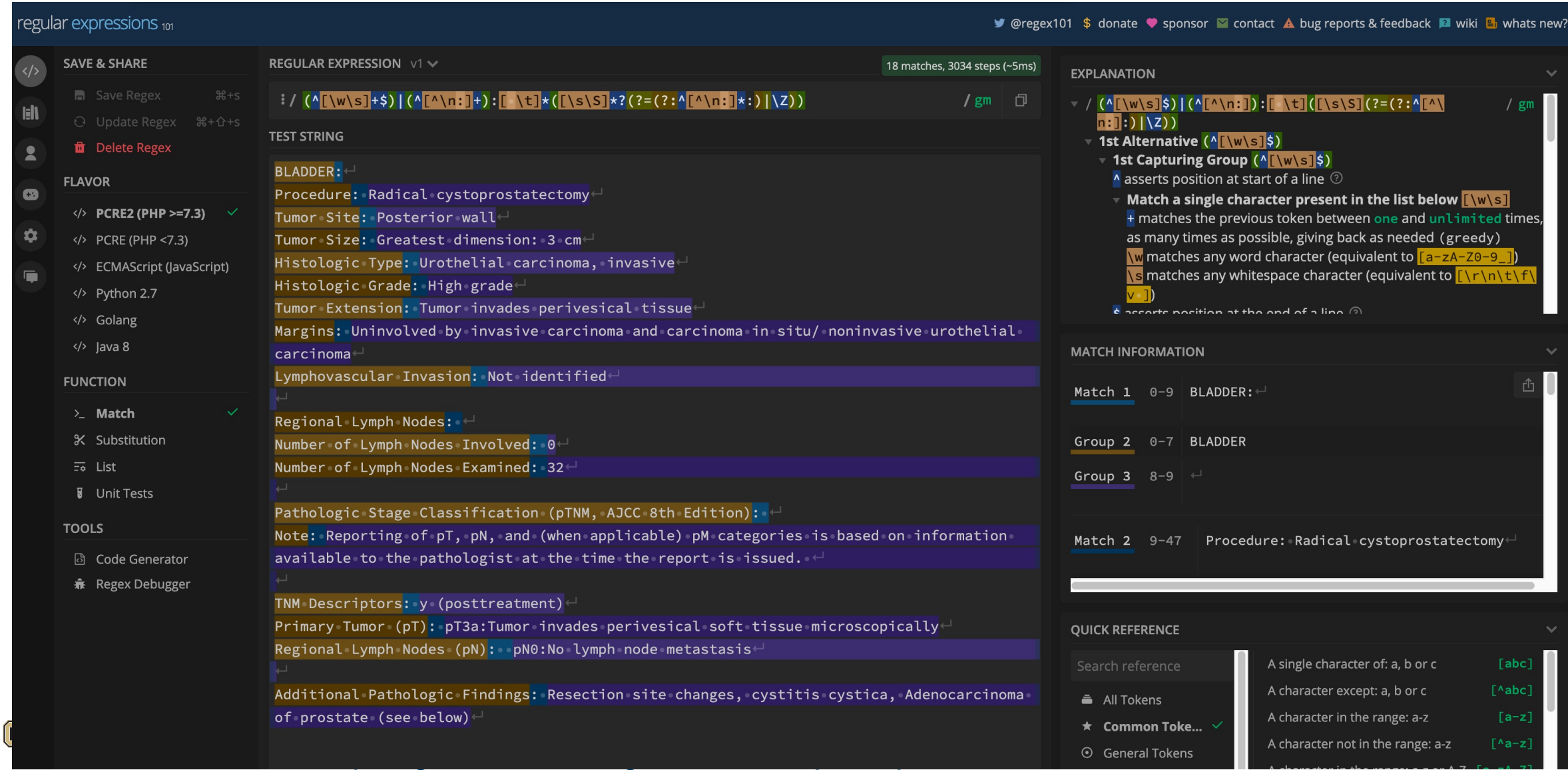
```
import re

get_subsections_pat = re.compile(r'(^[\w\s]+$)|(^[^\n:]+):[ \t]*([\s\S]*?(?=(?:^[^\n:]*:)|\Z))', flags=re.M)
# patterns for matching concepts
dimension_pat = re.compile(r'^[\d.xX\s]+(cm)?$')
def no_spec_char(input):
    return re.sub(r'[-(),:]', '', input.lower()).strip()

def no_spec_char_space(input):
    return re.sub(r'\s+', ' ', re.sub(r'[-(),:]', ' ', input.lower())).strip()

def no_paren(input):
    return re.sub(r'\([^)]*\)', '', input.lower()).strip()
# end pattern matching stuff

def get_sections(instr):
    return {m.group(1).strip() : m.group(3).strip() for m in get_sections_pat.finditer(instr)}
```

COPRH Con Colorado Pragmatic Research in Health Conference

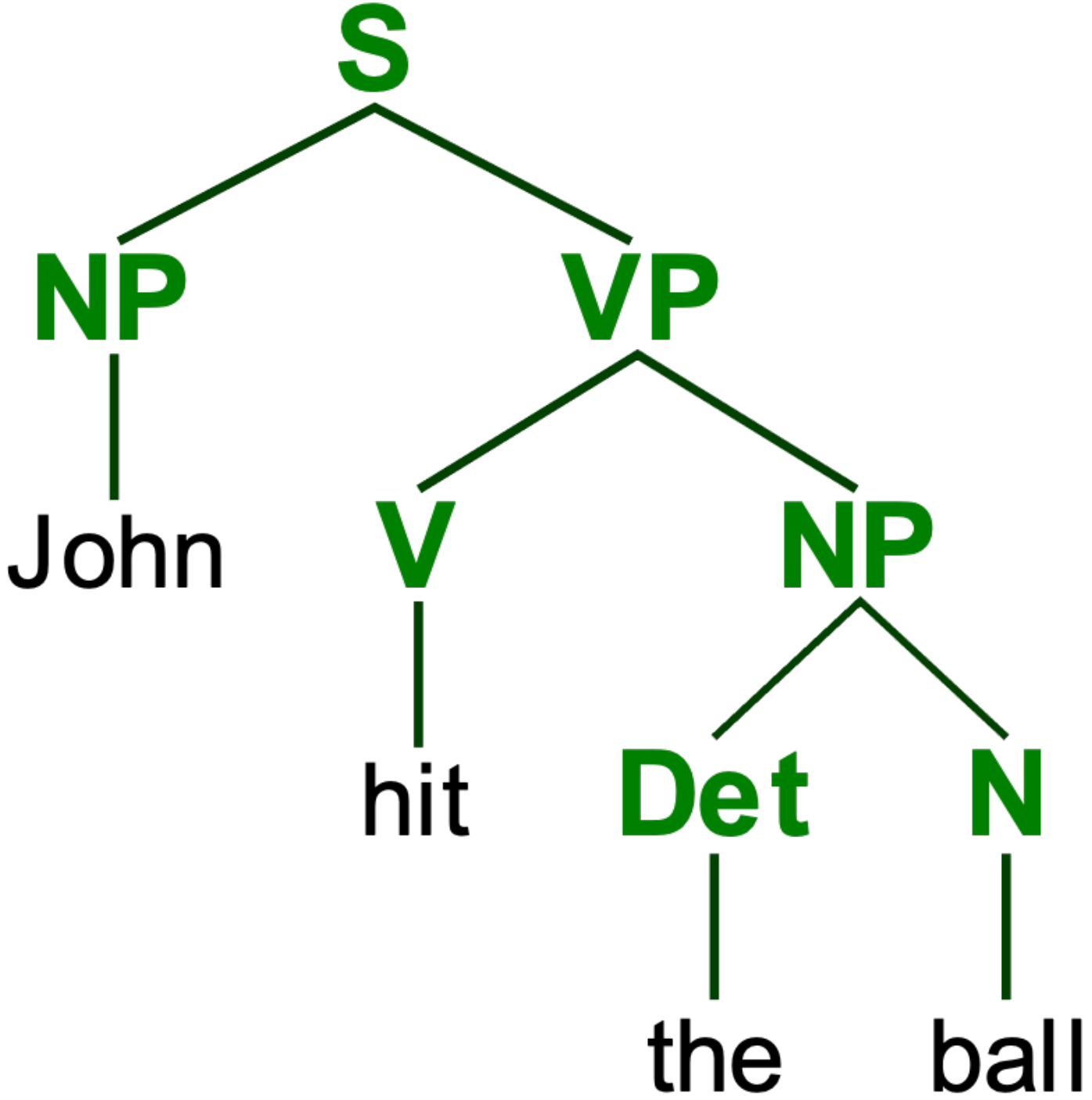
ACCORDS

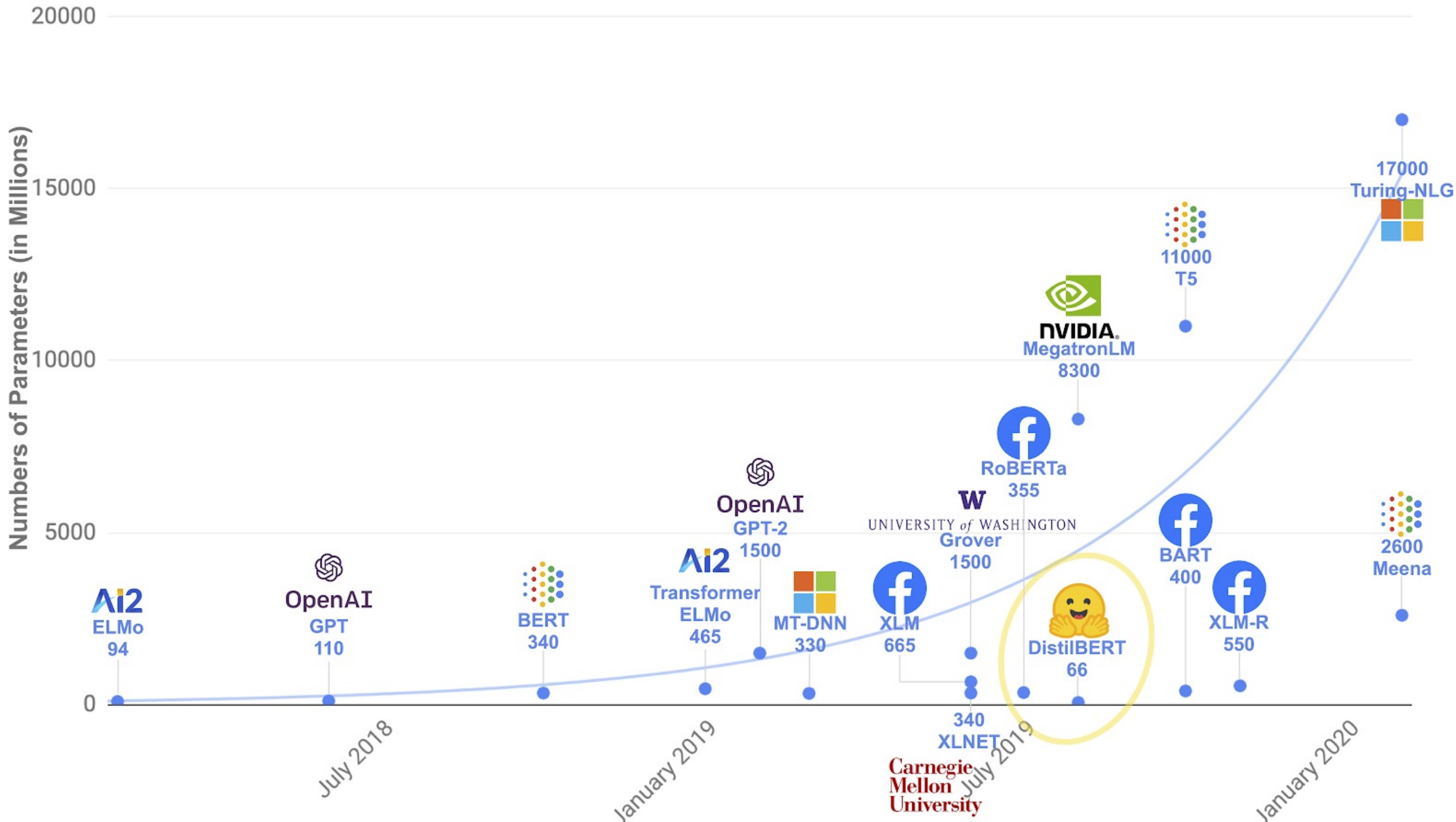
https://github.com/magic-lantern/coprh-nlp-2021

Regular Expressions

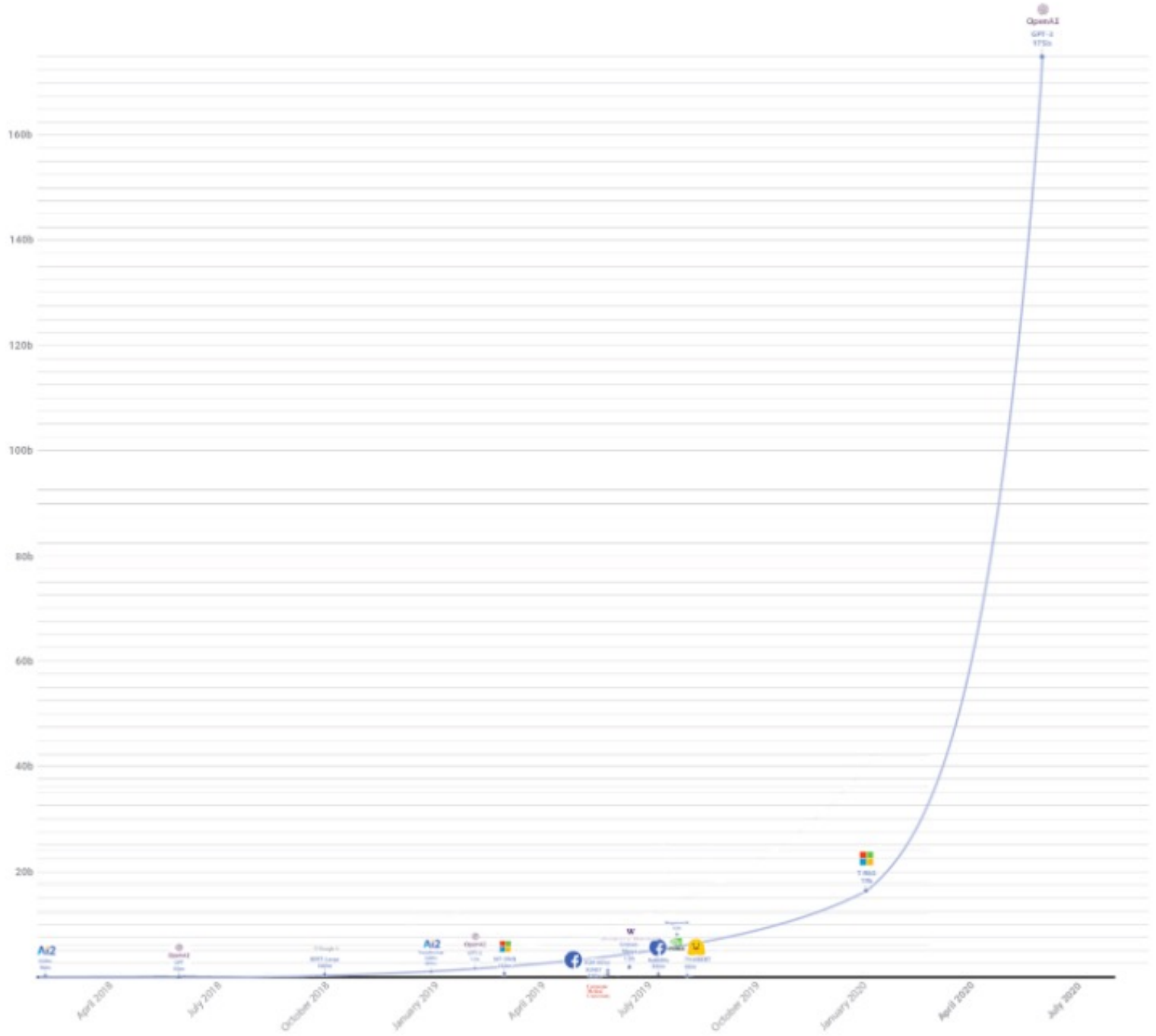


COPRH Con
Colorado Pragmatic Research in Health Conference





Source https://research.aimultiple.com/gpt/



COPRH Con
Colorado Pragmatic Research in Health Conference

ACCORDS

CCTSI

Language Modeling – Self Supervised

Given the sentence: John hit the ball

- _____ hit the ball
- John _____ the ball
- John hit _____ ball
- John hit the _____

- _THE_ hit the ball
- John _BE_ the ball
- John hit _A_ ball
- John hit the _OF_

- _YOU_ hit the ball
- John _ATE_ the ball
- John hit _MY_ ball
- John hit the _CAR_

1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.

Semi-supervised Learning Step

Model:

Dataset:

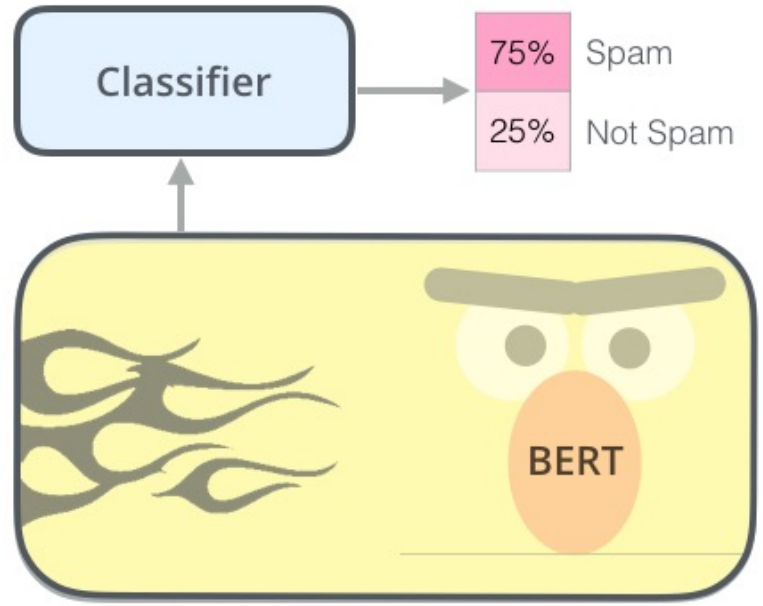
Objective:

Predict the masked word (language modeling)

2 - Supervised training on a specific task with a labeled dataset.

Supervised Learning Step

Model: (pre-trained in step #1)



Dataset:

Email message	Class
Buy these pills	Spam
Win cash prizes	Spam
Dear Mr. Atreides, please find attached…	Not Spam

Language Model based concept matching

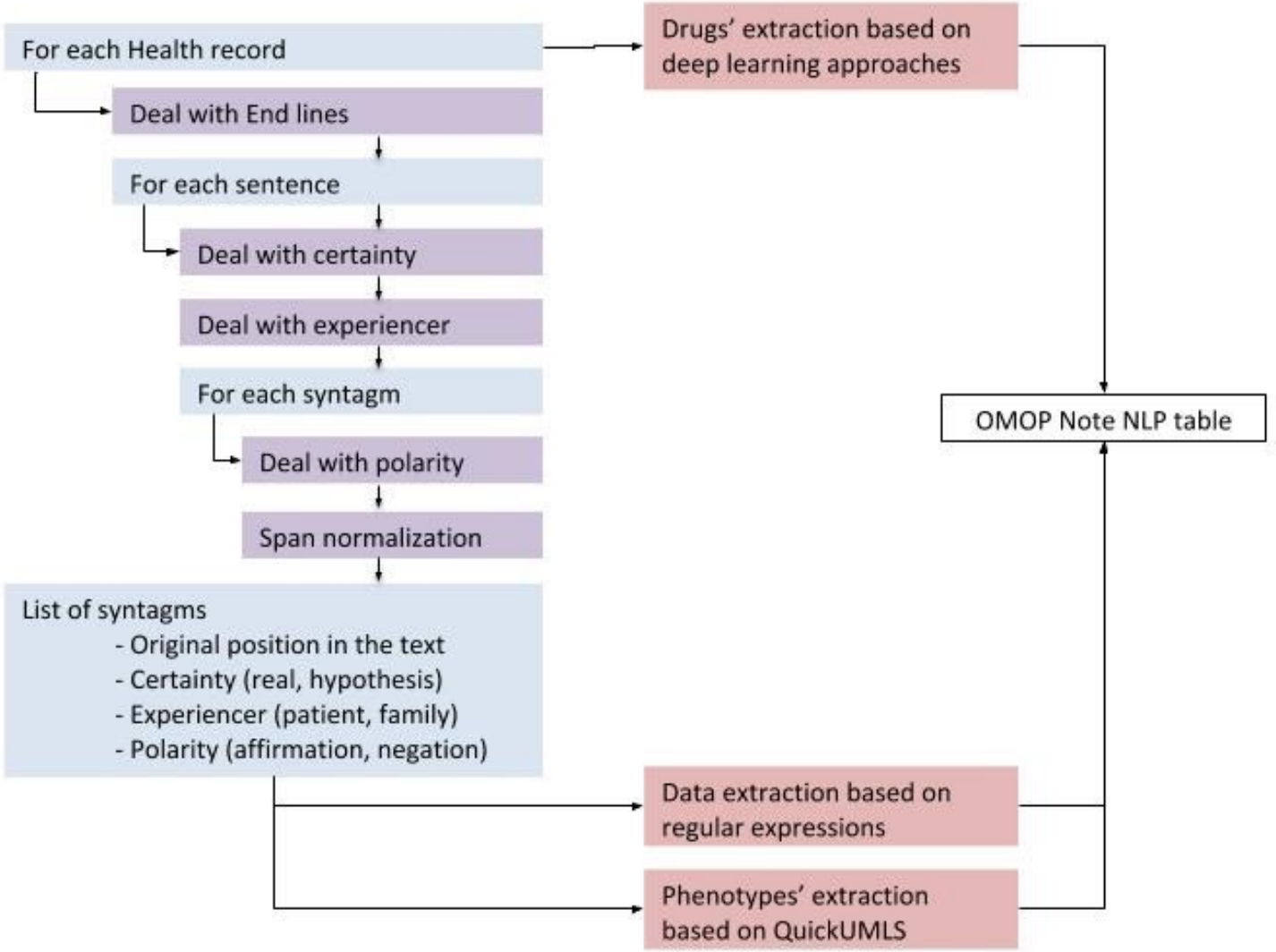
Given the phrase: “Number of Lymph Nodes Involved/ Examined”

top F1 results: tensor([0.9105, 0.9044, 0.7661])
top F1 matches:

- 'number of lymph nodes involved’
- 'number of lymph nodes examined’
- 'no lymph nodes submitted or found'

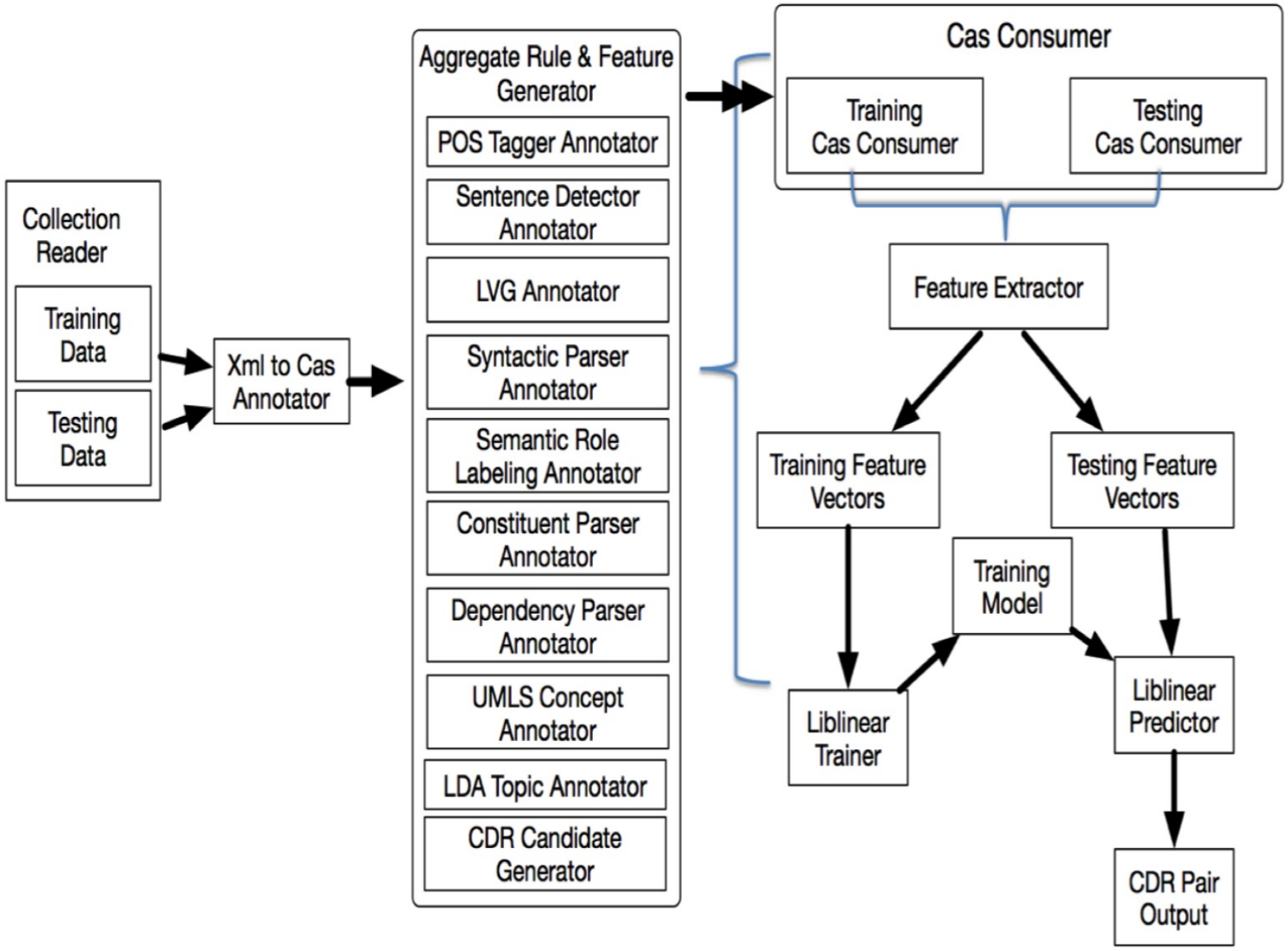
https://github.com/magic-lantern/coprh-nlp-2021

Practical NLP - Pipelines



https://doi.org/10.2196/20773

Practical NLP - Pipelines



Resolution of Chemical Disease Relations with Diverse Features and Rules

“We are coming now rather into the region of guesswork,” said Dr. Mortimer.

Holmes' reply: “Say, rather, into the region where we balance probabilities and choose the most likely. It is the scientific use of the imagination, but we have always some material basis on which to start our speculation.”

The Hound of the Baskervilles
Sir Arthur Conan Doyle

COPRH Con
Colorado Pragmatic Research in Health Conference

ACCORDS

https://github.com/magic-lantern/coprh-nlp-2021

Ethics in NLP

ACM Code of Ethics and Professional Conduct

Association for Computing Machinery, 2018

Computing professionals' actions change the world. To act responsibly, they should reflect upon the wider impacts of their work, consistently supporting the public good. The ACM Code of Ethics and Professional Conduct ("the Code") expresses the conscience of the profession...

1. GENERAL ETHICAL PRINCIPLES.

A computing professional should...

1. Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing
2. Avoid harm.
3. Be honest and trustworthy.
4. Be fair and take action not to discriminate.
5. Respect the work required to produce new ideas, inventions, creative works, and computing artifacts.
6. Respect privacy.
7. Honor confidentiality.

Continued at https://www.acm.org/code-of-ethics


COPRH Con
Colorado Pragmatic
Research in Health
Conference
“Data! data! data!” he cried impatiently.
“I can’t make bricks without clay.”
The Adventure of the Copper Beeches
Sir Arthur Conan Doyle
https://github.com/magic-lantern/coprh-nlp-2021
ACCORDS
CCTSI

Sources of text for NLP

- CU Anschutz Health Data Compass - With IRB Approval can access many clinical notes from UC Health and Children's Hospital of Colorado https://www.healthdatacompass.org/
- PhysioNet (MIMIC et al) https://physionet.org/about/database/

Other Text Sources

- Kaggle https://www.kaggle.com/datasets
- Hugging Face https://huggingface.co/datasets
- Amazon product reviews and ratings http://jmcauley.ucsd.edu/data/amazon/
- Twitter https://developer.twitter.com/en/products/twitter-api

It is a capital mistake to theorise before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.

A Scandal in Bohemia
Sir Arthur Conan Doyle

COPRH Con
Colorado Pragmatic Research in Health Conference

ACCORDS

Resources - Tutorials

- "Clinical Natural Language Processing" Laura K. Wiley, PhD Asst Prof @ CU Anschutz https://www.coursera.org/learn/clinical-natural-language-processing
- "Natural Language Processing Specialization" https://www.deeplearning.ai/program/natural-language-processing-specialization/
- "A Code-First Introduction to Natural Language Processing" https://www.fast.ai/2019/07/08/fastai-nlp/
- https://towardsdatascience.com/introduction-to-clinical-natural-language-processing-predicting-hospital-readmission-with-1736d52bc709
- Medical Transcription Classification: https://www.kaggle.com/ritheshsreenivasan/clinical-text-classification

Resources - Books

- "Introduction to Information Retrieval" https://nlp.stanford.edu/IR-book/
- "Speech and Language Processing" https://web.stanford.edu/~jurafsky/slp3/
- "Deep Learning for Coders with Fastai and PyTorch" https://book.fast.ai

https://github.com/magic-lantern/coprh-nlp-2021

Resources - NLP Libraries

- https://spacy.io/ - Python
- http://www.nltk.org/ - Python
- https://opennlp.apache.org/ - Java

https://github.com/magic-lantern/coprh-nlp-2021

Resources – NLP Researchers/Teachers at CU

- James Martin, Ph.D. @ CU Boulder
- Larry Hunter, Ph.D. @ CU Anschutz
- Laura K. Wiley, PhD @ CU Anschutz

https://github.com/magic-lantern/coprh-nlp-2021

Questions?

https://github.com/magic-lantern/coprh-nlp-2021

ACCORDS