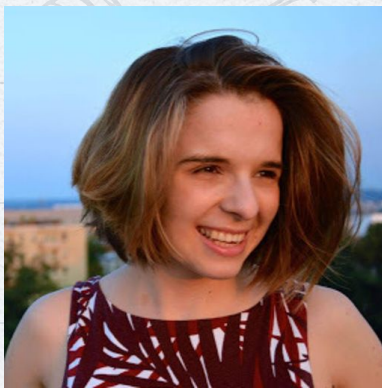# Intro to NLP with spaCy

Women Who Code DC Python

Women Who Code

Women Who Code (WWC) is a global non-profit dedicated to inspiring women to excel in technology careers. We work to support this generation in being and becoming leaders and role models in the tech industry.

- ★ We are the DC Chapter!
- ★ Support us! Volunteer or donate.
- ★ Visit our Meetup site: https://www.meetup.com/Women-Who-Code-DC/
- ★ Request to join our Slack chat website: http://bit.ly/wwcdcslack
- ★ Python Beginners: 1st Wednesday of every other month
- ★ Python Lab: 3rd Wednesday of the month

Hello!

Anastassia Kornilova
Women Who Code DC Python - Volunteer Lead
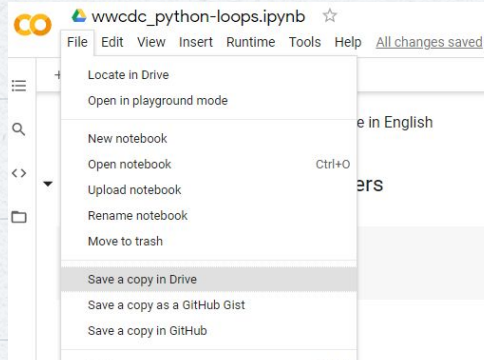Women Who Code DC Slack: @Anastassia Kornilova
Twitter: @anaphoraresolve
Website: akornilo.com

HOUSEKEEPING

- Use Twitch chat for questions.
- To get started, save a copy of the Colab notebook to your Google drive:

- In Colab:

  - Use Run button to execute code snippet

  - Use "+ Code" button to create your own code cell, copy, edit and run your version of the code snippet

Goals for this lab

- Use cases for Natural Language Processing
- Processing text with SpaCy and extracting basic properties
- Customizing SpaCy for your applications

A little about me

FN Outlook

Compared to other bills in Washington, this bill is much more likely to pass.

We calculated this based on **the strength of the bill sponsor, the language in the bill, and the network of the cosponsors.**

↑ Brian E. Blake is one of the most effective legislators in this chamber.

↑ The bill has bipartisan support.

↑ Laurie Jinkins is the Chairman on the House Committee on Rules. Marcus Riccelli is the House Majority Whip. June G. Robinson is the Vice Chair on the House Committee on Appropriations. Brian E. Blake is the Chairman on the House Committee on Rural Development, Agriculture, and Natural Resources. Larry S. Springer is the Deputy Majority Leader. Bruce Chandler is the Ranking Minority Member on the House Committee on Rural Development, Agriculture, and Natural Resources. Morgan Irwin is the House Minority Assistant Floor Leader. Laurie Jinkins is the Speaker of the House.

Click to Hide Analytics

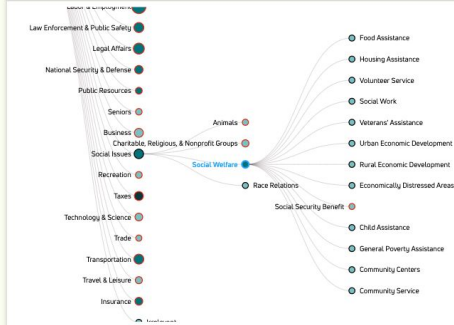| House Pre-Floor Score	House Floor Score	Senate Pre-Floor Score	Senate Floor Score
74.5%	82.0%	29.3%	67.0%

Topic Tree

Scroll, pan, and click to expand tree node to explore the hierarchy.

Law Enforcement & Public Safety
Legal Affairs
National Security & Defense
Public Resources
Seniors
Business
Charitable, Religious, & Nonprofit Groups
Social Issues
Recreation
Taxes
Technology & Science
Trade
Transportation
Travel & Leisure
Insurance

Animals
Social Welfare
Race Relations
Social Security Benefit

Food Assistance
Housing Assistance
Volunteer Service
Social Work
Veterans' Assistance
Urban Economic Development
Rural Economic Development
Economically Distressed Areas
Child Assistance
General Poverty Assistance
Community Centers
Community Service

Current Selection ⌄

Topic: Social Welfare (29-3)

Total bills: 93,375
Average score: 5.00
No child score: 2.19

Current Level ⌄

You are on **level 2** of the hierarchy.
Highest average score: **Social Welfare (5.00)**

Highest total bills: **Charitable, Religious, & Nonprofit Groups (98,020)**

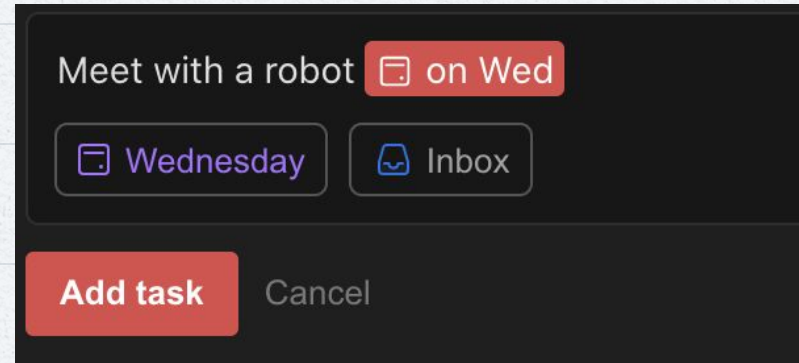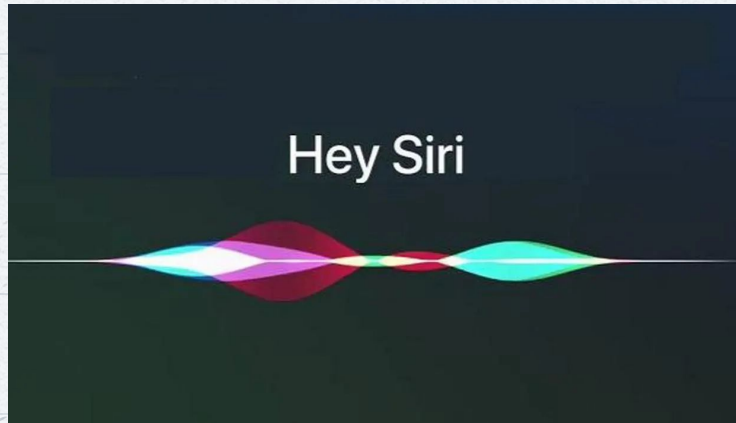## Key Terms

```
[
  {
    "key": "prescription drugs",
    "doc_count": 196,
    "score": 0.0005841005805186974,
    "bg_count": 2519
  },
```
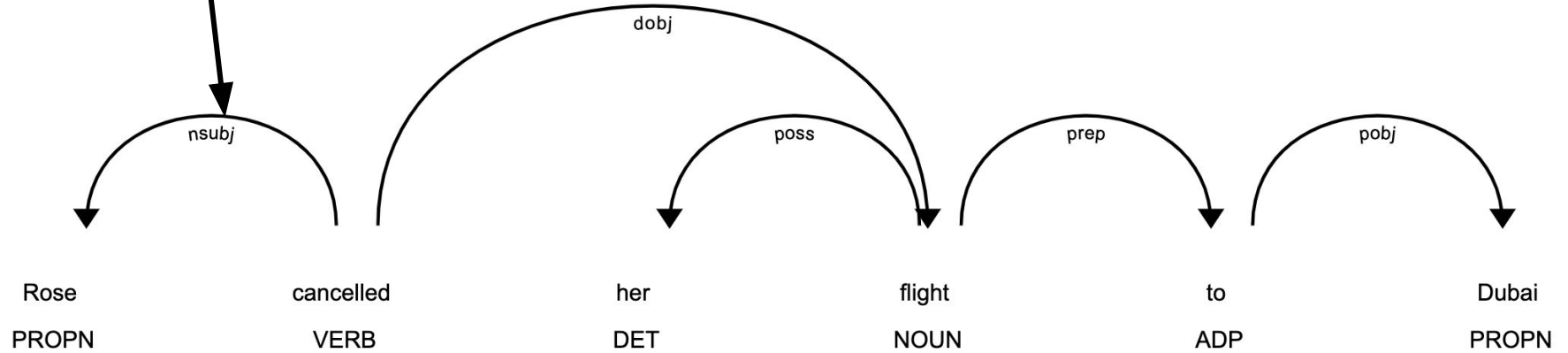
A Quick Overview

What is Natural Language Processing

- Computers dealing with human languages (usually text)
- Often involves Machine Learning (ML) methods

Some Specific Tasks

Dependency Parsing

Part of Speech
Tagging

9

Find Good Keywords – Only Pick Noun Phrases

This is a **sentence** about **vegetarian burritos** . Yet, the next one is about **white elephants** . These **phrases** are **good hashtags** .

> 1   *Be it enacted by the Senate and House of Representa-*
> 2   *tives of the United States of America in Congress assembled,*
> 3   **SECTION 1. SHORT TITLE.**
> 4   This Act may be cited as the "Coronavirus Aid, Re-
> 5   lief, and Economic Security Act" or the "CARES Act".

- clinical nurse specialist
- coronavirus-related distribution
- covered small business
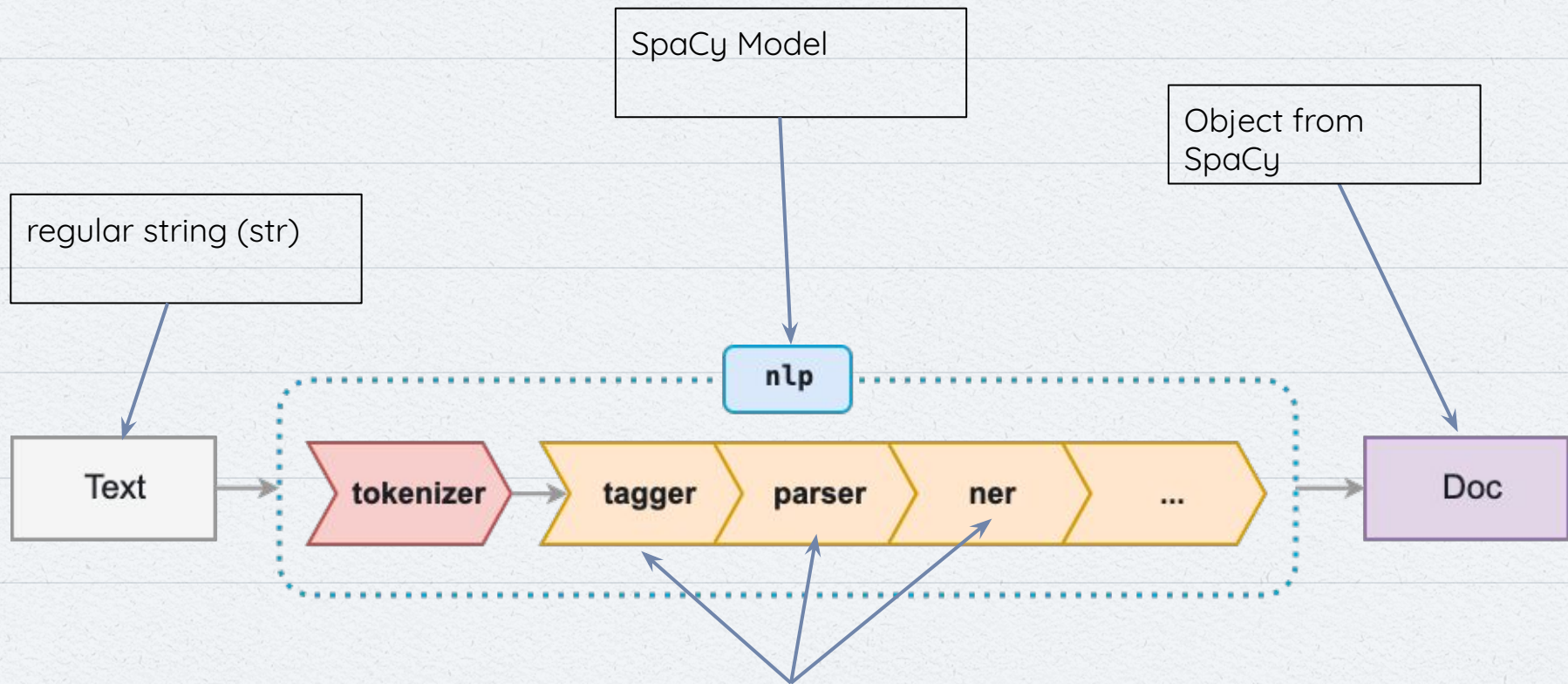- antimicrobial drug

Named Entity Recognition

Anastassia Kornilova PERSON gave a talk on NERs on Thursday DATE at the WWC ORG conference from her home in Washington GPE , DC

Would you like to tag @Anastassia?

Set a reminder for Thursday?

Text Classification (Not Covered in Lab)

- Sentiment: Positive/Negative
    - Predict if client likes product
- Topic: Categorize new data

What is spaCy

- A library with algorithms for doing NLP Tasks
- Contains pre-built models for multiple languages
- Contains framework for expanding SpaCy to new tasks and languages.

SpaCy Model

Object from SpaCy

regular string (str)

Text

tokenizer → tagger → parser → ner → ...

nlp

Doc

Trained with Machine Learning + Rules created by linguists

```
nlp = spacy.load("en_core_web_sm")

text = "Sentence goes here"

doc = nlp(text)
```

en _core_web_ sm

Determines tokenization rules +
indicates data used is in English

en `_core_web_` sm

Describes dataset: Data from the
internet: wikipedia + random sites

en _core_web_ [sm]

Model size: bigger is more accurate but also takes up more space

Example Time!

We will be working through the example
posted in the Meetup and in chat
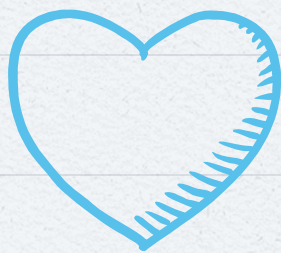
Resources <3

★ Our WWCodeDC Python Github:

https://github.com/womenwhocodedc/python-community

★ Inspiration for this lab:

- Transforming Code into Beautiful Idiomatic Python, Raymond Hettinger
  https://www.youtube.com/watch?reload=9&v=anrOzOapJ2E
- Mental Game of Python, Raymond Hettinger https://www.youtube.com/watch?v=Uwuv05aZ6ug

★ Slack: Sign up! womenwhocodedc.slack.com

THANKS!

Any questions?

You can find me on Women Who Code DC Slack

@ElvitaLobo

Credits

Special thanks to all the people who made and released these awesome resources for free:

- ✘ Presentation template by <u>SlidesCarnival</u>
- ✘ Photographs by <u>Unsplash</u>