



# Custom Name Entity Model with spaCy

Domain: WTA Tournament Press Conferences

Van

[goovan@gmail.com](mailto:goovan@gmail.com)

General Assembly

# Table of Contents

01

Overview

02

Why This

03

Data

04

Framework

05

Features

06

Difficulties

07

Next

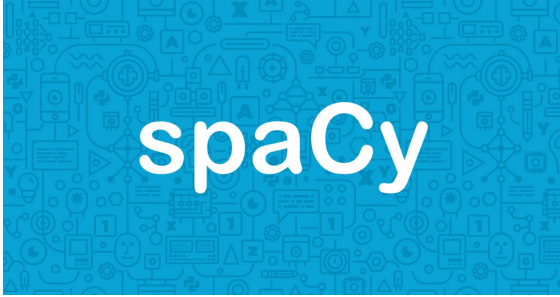
08

References



# 01

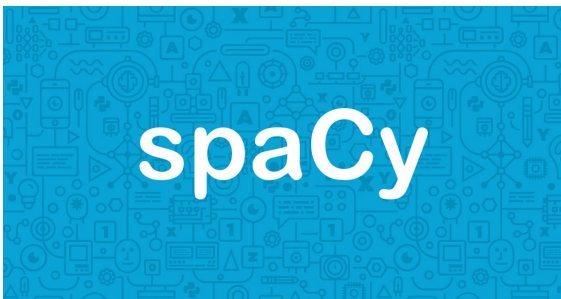
## Overview



## Core:

A custom name entity model





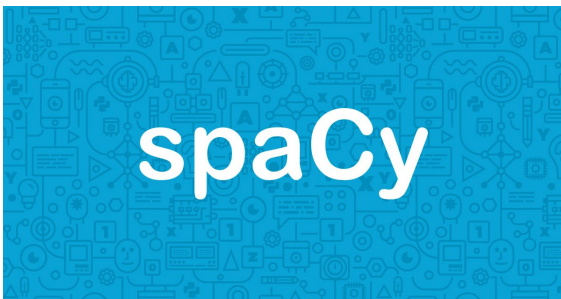
## Core:

A custom name entity model

## Domain:

WTA tournament press conferences





## Core:

A custom name entity model

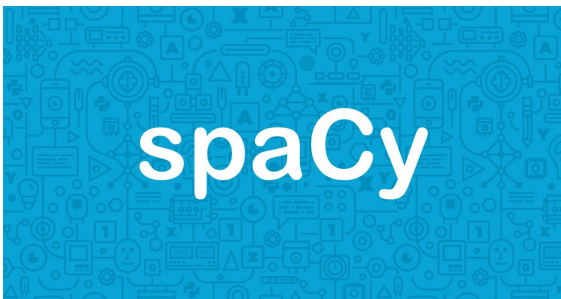
## Domain:

WTA tournament press conferences

## Tool:

spaCy





## Core:

A custom name entity model

## Domain:

WTA tournament press conferences

## Tool:

spaCy

## Target:

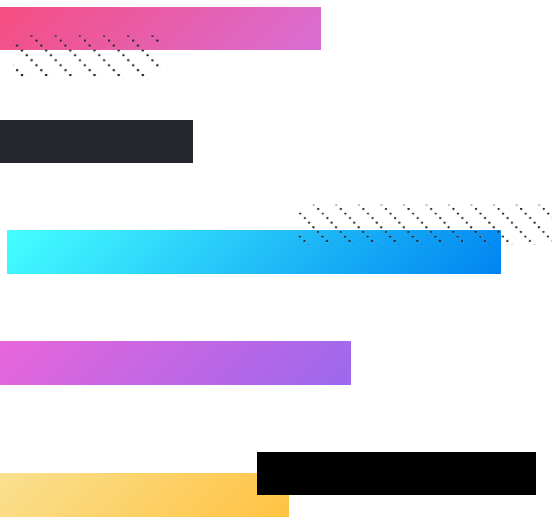
English/Chinese interpreters

# 02

Why this







# Problem 1

Going through everything  
without skipping

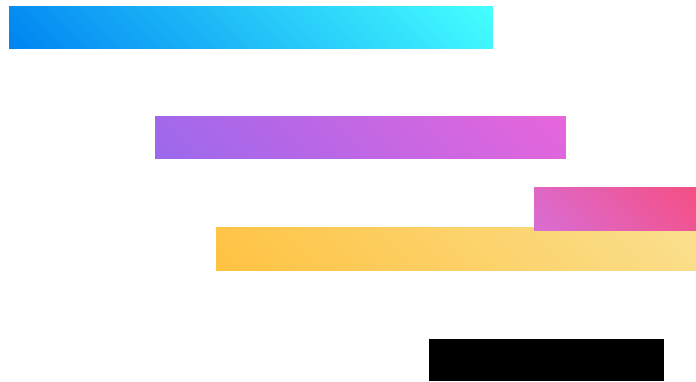


# Problem 1

Going through everything  
without skipping

## Solution

Reduce the amount to read



# Problem 2

Google things in China



# Problem 2

Google things in China

Solution

Not to google at all



In fact, the Chinese NORP market has the three CARDINAL most influential names of the retail and tech space – Alibaba GPE , Baidu ORG , and Tencent PERSON (collectively touted as BAT ORG ), and is betting big in the global AI GPE in retail industry space . The three CARDINAL giants which are claimed to have a cut-throat competition with the U.S. GPE (in terms of resources and capital) are positioning themselves to become the ‘future AI PERSON platforms’. The trio is also expanding in other Asian NORP countries and investing heavily in the U.S. GPE based AI GPE startups to leverage the power of AI GPE . Backed by such powerful initiatives and presence of these conglomerates, the market in APAC AI is forecast to be the fastest-growing one CARDINAL , with an anticipated CAGR PERSON of 45% PERCENT over 2018 - 2024 DATE .

To further elaborate on the geographical trends, North America LOC has procured more than 50% PERCENT of the global share in 2017 DATE and has been leading the regional landscape of AI GPE in the retail market. The U.S. GPE has a significant credit in the regional trends with over 65% PERCENT of investments (including M&As, private equity, and venture capital) in artificial intelligence technology. Additionally, the region is a huge hub for startups in tandem with the presence of tech titans, such as Google ORG , IBM ORG , and Microsoft ORG .

# Domain

## WTA Tournament Press Conferences



Sloane Stephens, 2017 US Open Champion

# Categories





# Categories

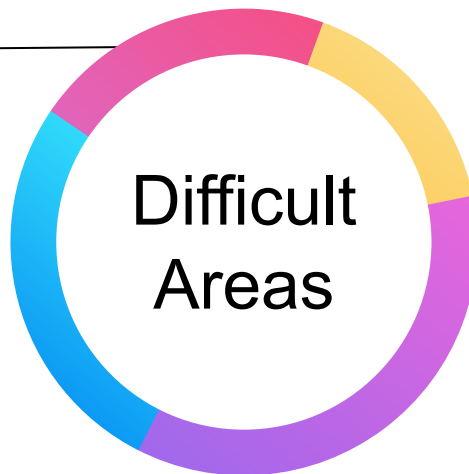


## Tournament titles

Grand Slams

Fed Cup

Location-based ones



# Categories



## Tournament titles

Grand Slams

Fed Cup

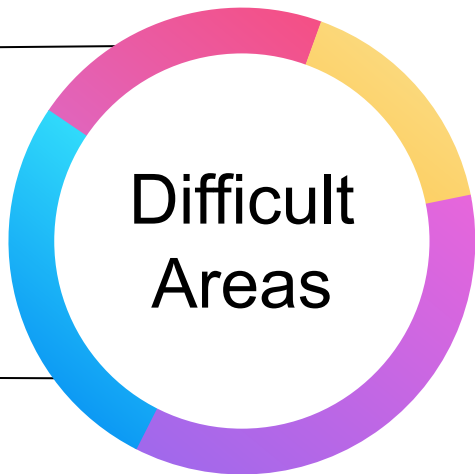
Location-based ones

## Tennis game terms

Serving sets

Break points

The eagle eye



# Categories

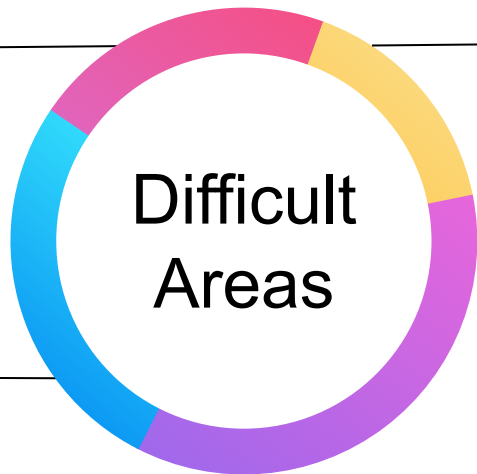


## Tournament titles

- Grand Slams
- Fed Cup
- Location-based ones

## Tennis game terms

- Serving sets
- Break points
- The eagle eye



## Tennis players

- Names
- Styles
- Rankings
- Career achievements

# Categories

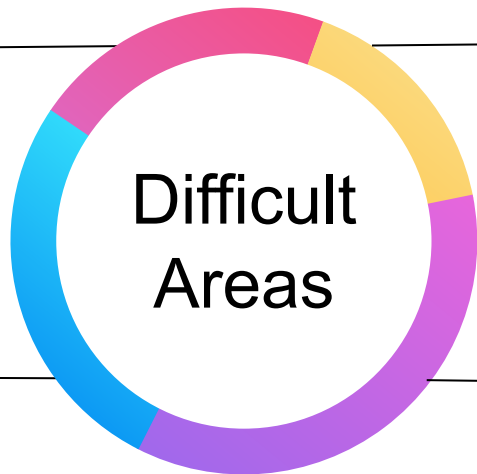


## Tournament titles

Grand Slams  
Fed Cup  
Location-based ones

## Tennis game terms

Serving sets  
Break points  
The eagle eye



## Tennis players

Names  
Styles  
Rankings  
Career achievements

## Generic words

that mean something different in this context



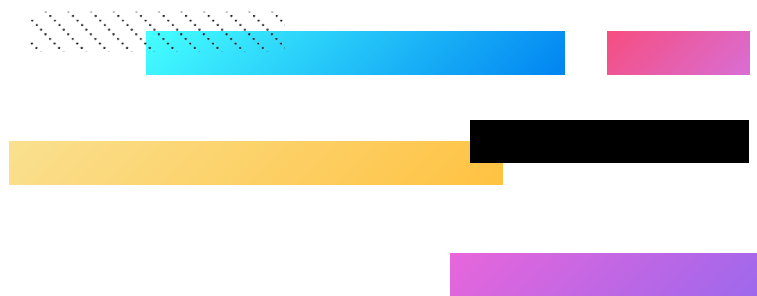
03

Data



# Data

Word  
documents



# Data

Word  
documents



Cleaned



# Data



Word  
documents



Cleaned



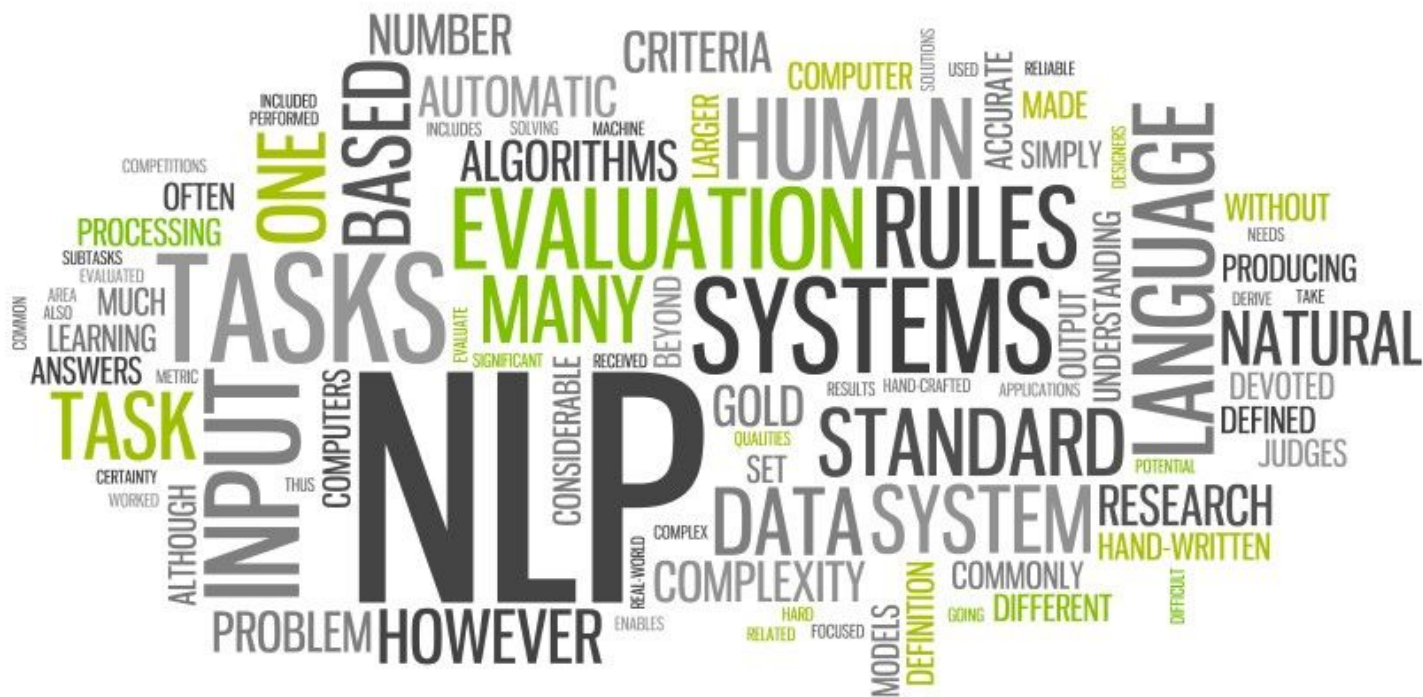
Split





# Data





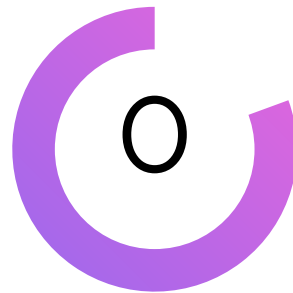


# Labels



11%

210



89%

1701


# 04

## Framework



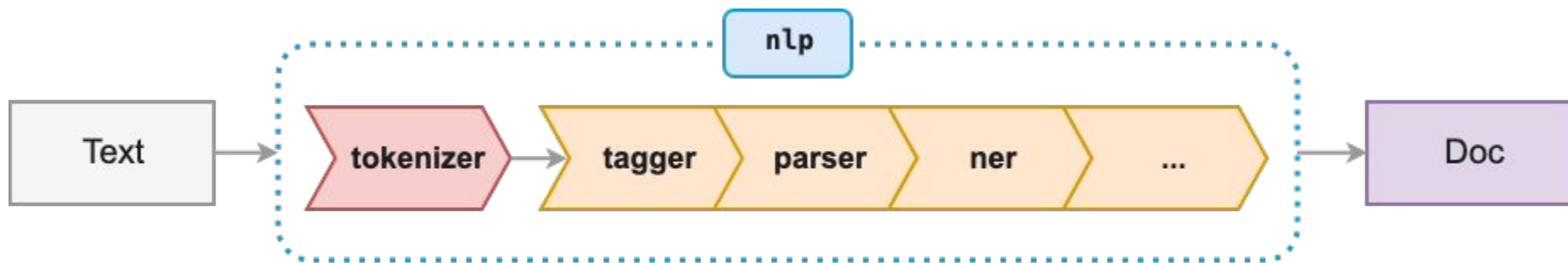


# spaCy



An open-source software  
library for advanced natural  
language processing

# Pipeline



# 05

## Features





# spaCy rule-based matching





#### *#type 1. tournament titles*

```
fedcup_pattern = [{'IS_TITLE': True}, {'LOWER': 'cup'}]  
grand_slam_pattern = [{'LOWER': 'grand'}, {'LOWER': {'IN': ['slam', 'slams']}}]  
tournament_pattern = [{'IS_ALPHA': True, 'POS': 'PROPN'}, {'LOWER': 'open'}]
```

#### *#type 2. player names*

```
first_name_pattern = [{'POS': 'PROPN', 'IS_TITLE': True}]  
full_name_pattern = [{'IS_TITLE': True}, {'IS_TITLE': True}]
```

#### *#type 3. tennis terms*

```
double_fault_pattern = [{'LOWER': 'double'}, {'LEMMA': 'fault'}]  
rally_pattern = [{'LEMMA': 'rally'}]  
set_pattern = [{'POS': 'ADJ'}, {'LOWER': 'set'}]  
dropshot_pattern = [{'LOWER': 'dropshot'}]  
serve_pattern = [{'LEMMA': 'serve'}]  
timeout_pattern = [{'LOWER': 'medical'}, {'LOWER': 'timeout'}]  
break_pattern = [{'LEMMA': 'break'}]  
round_robin_pattern = [{'LOWER': 'round'}, {'LEMMA': 'robin'}]  
ace_pattern = [{'LEMMA': {'IN': ['ace', 'volley', 'dropshot']}}]  
winner_pattern = [{'LEMMA': 'winner'}]  
break_point_pattern = [{'LOWER': 'break'}, {'LEMMA': 'point'}]  
three_setter_pattern = [{'POS': 'NUM'}, {'IS_PUNCT': True}, {'LEMMA': 'set'}]
```

#### *#type 4. contextual words*


```
aggressive_pattern = [{'LEMMA': 'aggressive'}]
```



[(" At the US Open I had problem on my foot, but now I'm much better", {'entities': [(8, 15, '网球比赛名称')]}),

('Earlier this year you play Shenzhen Open and a lost to a Chinese player', {'entities': [(27, 40, '网球比赛名称')]}),

("I think for a long time, maybe back 10, 15 years ago when Hingis was winning Grand Slams at 16, 17, if you're not at least top 20, top 10, you don't have a chance", {'entities': [(77, 88, '网球比赛名称')]}))]

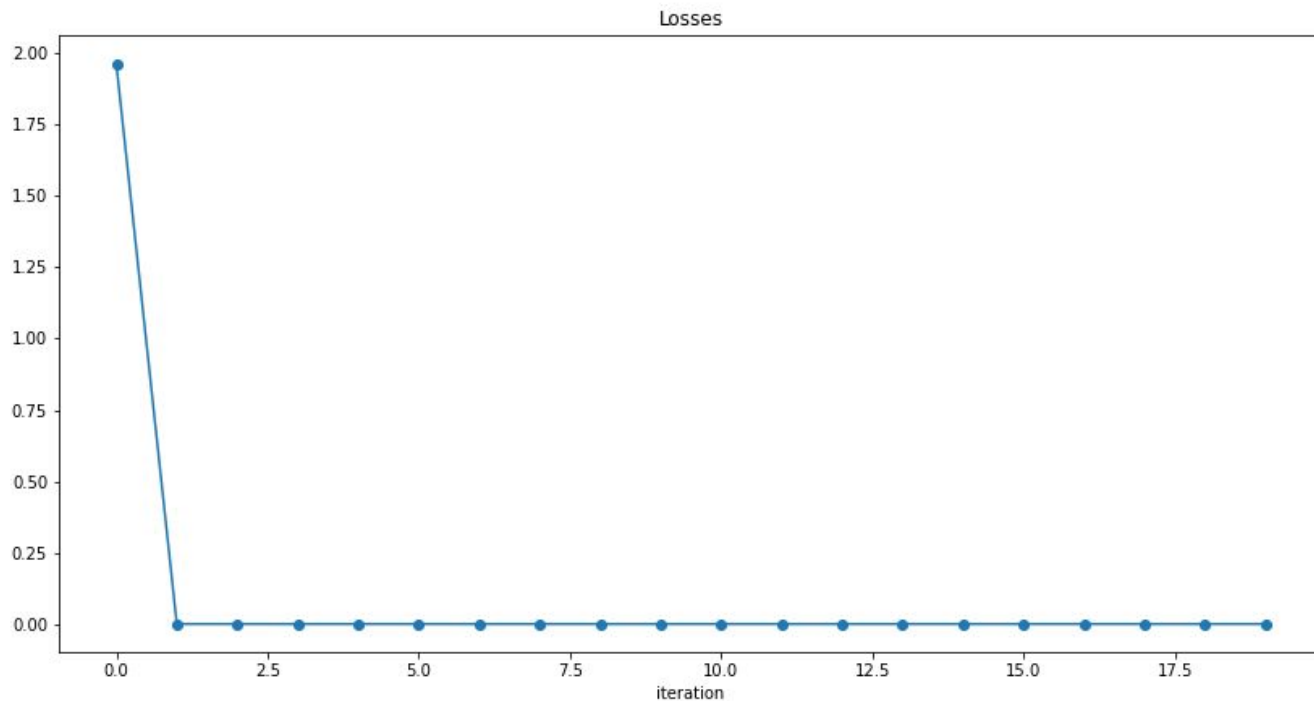


[(" At the **US Open** I had problem on my foot, but now I'm much better", {'entities': [(8, 15, '网球比赛名称')]}),

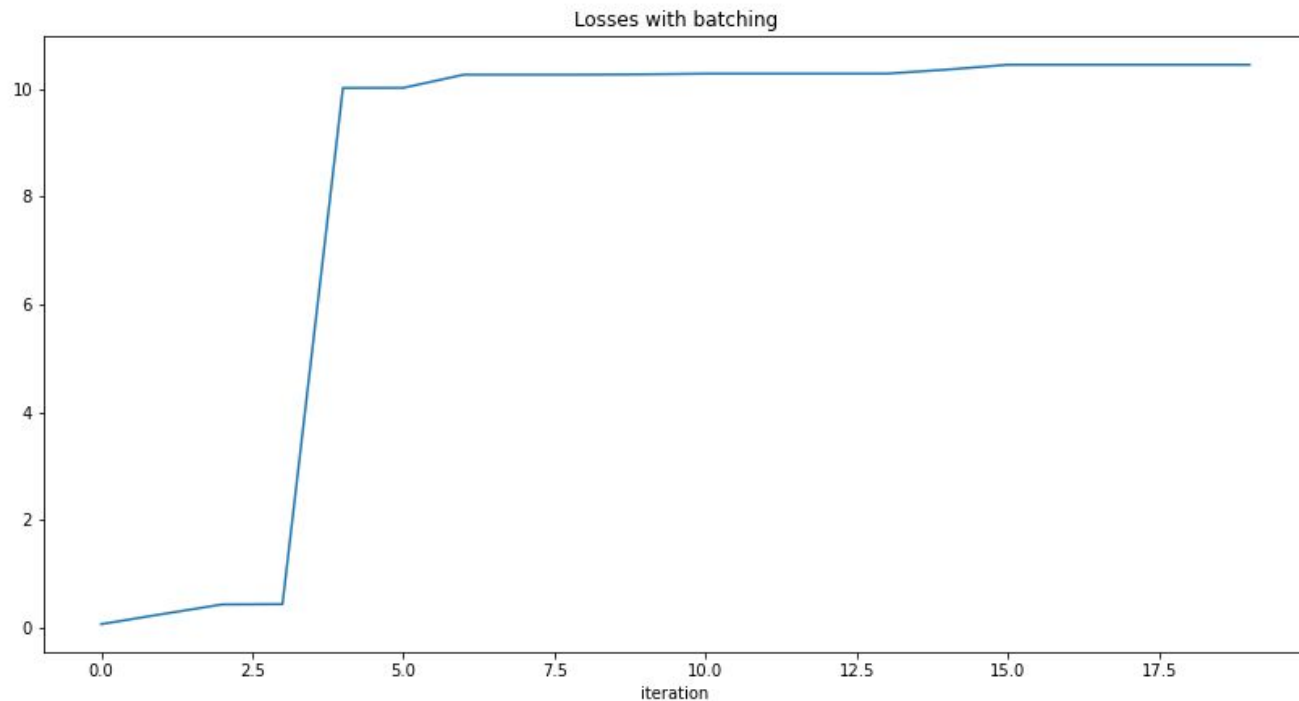
('Earlier this year you play **Shenzhen Open** and a lost to a Chinese player', {'entities': [(27, 40, '网球比赛名称')]}),

("I think for a long time, maybe back 10, 15 years ago when Hingis was winning **Grand Slams** at 16, 17, if you're not at least top 20, top 10, you don't have a chance", {'entities': [(77, 88, '网球比赛名称')]}))]

# Training



# Training



# 06 Difficulties



# Difficulties



## Creating matches

1. How patterns work
2. What you want
3. How many to create

# Difficulties



## Creating matches

1. How patterns work
2. What you want
3. How many to create

## Working with spaCy

1. Version
2. Doc





07

Next



- Include more labels





- Include more labels
- Evaluate



- 
- Include more labels
  - Evaluate
  - Streamlit app
- 

08

References





Natural Language Processing in Python [YouTube]

Intro to NLP with spaCy [YouTube]

spaCy documentation



# Thanks!

**CREDITS:** This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik**

Van  
[goovan@gmail.com](mailto:goovan@gmail.com)  
General Assembly

