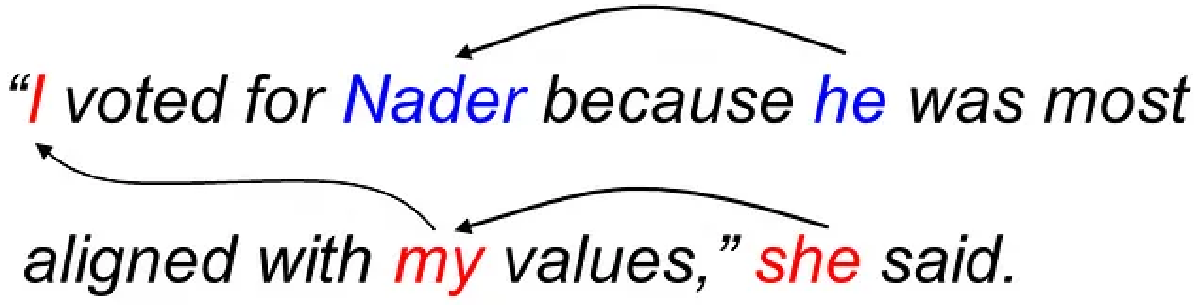
Coreference Resolution [NLP, Python]

Yash Jain · Following

Published in Dev Genius · 4 min read · Mar 15, 2022



"I voted for Nader because he was most aligned with my values," she said.

Source: https://nlp.stanford.edu/projects/coref.shtml

Coreference resolution is the task of finding all referring expressions like — (he, I, that, this…, or any subject or noun) is referred to which entity (referents like any person, thing, subject etc…)

Some types of References

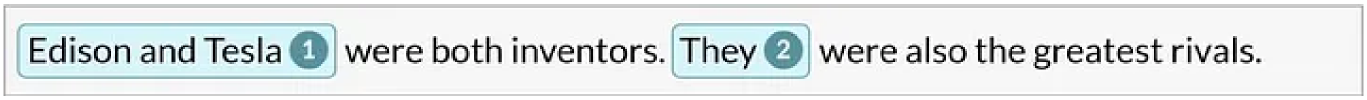
- **Anaphora** — acc. to Wikipedia → "*anaphora is the use of an expression whose interpretation depends specifically upon another (antecedent) expression*" or you can say "*when the referring expression(anaphor) is pointing backwards*"

> *Example:- **The music** was so loud that **it** couldn't be enjoyed*
>
> ***it** → here it is referring to **The music,** here "it" appeared after "The music" in sentence.*

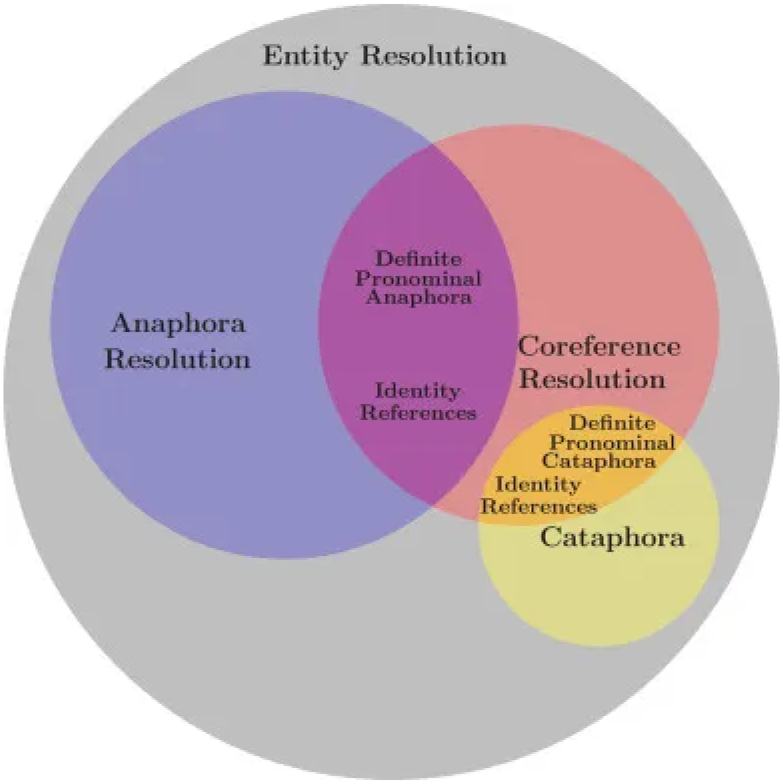
- **Cataphora** — its said to be just reverse of anaphora →"*the use of an expression that depends upon a postcedent expression*" or you can say "*when the referring expression is pointing forward*"

> *Example:- When **he** arrived home, **John** went to sleep.*
>
> ***he** is referred to **John,** and "he" came before "John" in sentence.*

Split antecedents It's an anaphoric expression where the pronoun (2) refers to more than one antecedent (1).



Edison and Tesla 1 were both inventors. They 2 were also the greatest rivals.



Source: ScienceDirect.com

We'll focus on **Coreference Resolution** - it is the task of determining whether two or more mentions *corefer (meaning do they refer to same entity…)*

A coreferent expression is only anaphoric if its interpretation depends on a previous expression in the text (i.e., its antecedent)

> *Example of named mentions instead of pronouns:-*
> *1.* ***International Business Machines*** *sought patent compensation from Amazon;* ***IBM*** *had previously sued other companies.*
> *Well you can see IBM referred to International Business Machines…. these type of references are also there..*
>
> ***2. Barack Obama*** *traveled to ….* ***Obama*** *…*
> *So we can see "obama" and "Barack Obama" are referred to same person.*

Coreference resolution thus comprises two tasks (although they are often performed jointly): (1) identifying the mentions, and (2) clustering them into coreference chains/discourse entities

Let's see another example

> *"Victoria Chen, CFO of Megabucks Banking, saw her pay jump to $2.3 million, as the 38-year-old became the company's president. It is widely known that she came to Megabucks from rival Lotsabucks"*

Let's make cluster of above example:

1. Victoria Chen, her, the 38-year-old, She
2. Megabucks Banking, the company, Megabucks
3. her pay
4. Lotsabucks

An entity that has only a single mention in a text (like Lotsabucks and her pay) is called a **singleton**.

Applications:

- Text Summarization
- Machine Translation
- Information Extraction
- Chatbots / Question-Answer system

> Let’s take a real life example of question answer engine:-

Content provided to QnA engine →
“*Joseph Robinette Biden Jr*. is an American politician who is the 46th and current president of the United States. A member of the Democratic Party, he served as the 47th vice president from 2009 to 2017 under Barack Obama and represented Delaware in the United States Senate from 1973 to 2009.”

Now if we ask question → “Who served as the 47th vice president from 2009 to 2017?”

In QnA engine **without reference resolution** it will give us → “**he**” but that is not the answer we want right? We know here that “**he**” is referred to **Joseph Robinette Biden.** and we want his name as an answer right..? That’s where coreference resolution comes in. It let’s us knows that he was referred to Joseph Robinette Biden so that we work something out and replace “he” or any other mentions, behind the scene. So that we get appropriate answer.

Implementation

spaCy-huggingface(NeuralCoref) coreference resolution

NeuralCoref is a pipeline extension for spaCy 2.1+ which annotates and resolves coreference clusters using a neural network. NeuralCoref is production-ready, integrated in spaCy’s NLP pipeline and extensible to new training datasets.

some more attributes other than `coref_clusters`, `coref_resolved` are there that you can checkout on github here.

To train neural coreference model you can checkout blog here.

Allennlp coreference resolution

```
 pip install allennlp
```

```
pip install allennlp-models
```

If you notice above output cluster formed are [0 to 3 and 26 to 26] and [34 to 34 and 56 to 56]these are the indexes that are given to tokens. See below which are these

```
print(prediction['document'])
```

Output:

```
['Joseph', 'Robinette', 'Biden', 'Jr.', 'is', 'an', 'American',
'politician', 'who', 'is', 'the', '46th', 'andcurrent', 'president',
'of', 'the', 'United', 'States', '.', 'A', 'member', 'of', 'the',
'Democratic', 'Party', ',', 'he', 'served', 'as', 'the', '47th',
'vice', 'president', 'from', '2009', 'to', '2017', 'under', 'Barack',
'Obama', 'andrepresented', 'Delaware', 'in', 'the', 'United',
'States', 'Senate', 'from', '1973', 'to', '2009', '.']
```

and if you notice index 0–3 which represents `Joseph Robinette Biden Jr.` (0,1,2,3) and index 26 (`‘he’`) these are 1 cluster. Another one is `2009` with itself so we can ignore that.

Read this blog about how to make and effective coreference resolution model.
It also explains that **Allennlp seems to find much more clusters than Huggingface neuralcoref**, and Huggingface have some problems while detecting cataphora, allennlp detects cataphora in sentence but it replaces with its first mention in cluster because it considers first mentions as its head. In this beautifully explained blog Martha has explained how to overcome that issue and make some improvements. You can checkout blog **here** and code **here**… These libraries with improvement will still give us good but not perfect result. There is still research going on, on how to make better coreference resolution model.

Coreference Resolution Python NLP Naturallanguageprocessing Anaphora

More from the list: "NLP"

Curated by Himanshu Birla

Jon Gi... in Towards Data ...

Characteristics of Word Embeddings

✦ · 11 min read · Sep 4, 2021

Jon Gi... in Towards Data ...

The Word2vec Hyperparameters

✦ · 6 min read · Sep 3, 2021

Jon Gi... in

The Word2ve

✦ · 15 min rea

View list

Written by Yash Jain

Following

94 Followers · Writer for Dev Genius

Data Scientist/ Data Engineer at IBM | Alumnus of @niituniversity | Natural Language Processing | Pronouns: He, Him, His

More from Yash Jain and Dev Genius

Yash Jain

Spell check and correction[NLP, Python]

In Natural Language Processing it's important that spelling errors should be as less as…

4 min read · Feb 19, 2022

21 1

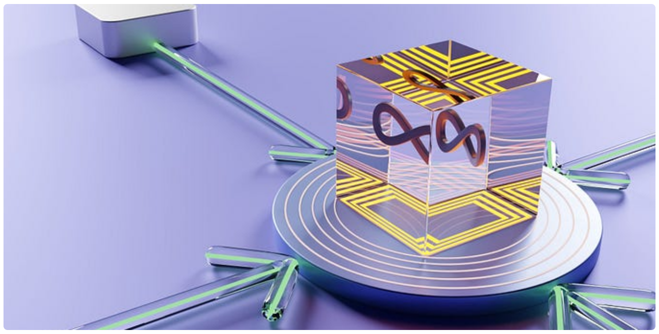
Devansh- Machine Learning Made Si… in Dev Gen…

Amazon Prime Video reduced costs by 90% by ditching…

This article is *not* sponsored by the World Monolith Supremacy Association

6 min read · May 10

1.3K 41



Priyansh Khodiyar in Dev Genius

Python for DevOps — A Definitive Guide

Python Python Python! Apple might have to add the pronunciation of the word Python in…

9 min read · Jul 9

265 6

Yash Jain

Stopwords [NLP, Python]

Stop words are common words in any language that occur with a high frequency b…

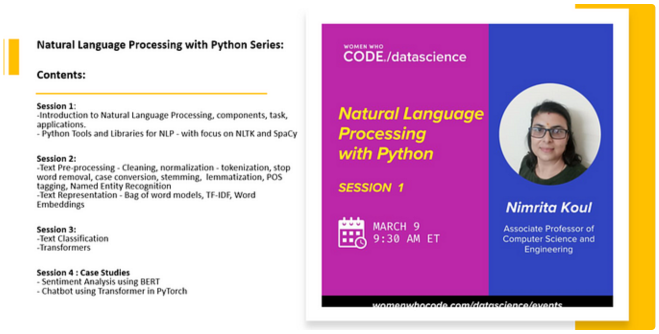
6 min read · Feb 23, 2022

8

See all from Yash Jain

See all from Dev Genius

Recommended from Medium



Natural Language Processing with Python Series:

Contents:

Session 1:
-Introduction to Natural Language Processing, components, task, applications.
- Python Tools and Libraries for NLP - with focus on NLTK and SpaCy

Session 2:
-Text Pre-processing - Cleaning, normalization - tokenization, stop word removal, case conversion, stemming, lemmatization, POS tagging, Named Entity Recognition
-Text Representation - Bag of word models, TF-IDF, Word Embeddings

Session 3:
-Text Classification
-Transformers

Session 4 : Case Studies
- Sentiment Analysis using BERT
- Chatbot using Transformer in PyTorch

WOMEN WHO CODE./datascience

Natural Language Processing with Python

SESSION 1

MARCH 9
9:30 AM ET

Nimrita Koul

Associate Professor of Computer Science and Engineering

womenwhocode.com/datascience/events

Nimrita Koul

NLP with Python Part 2 NLTK

This is the second article in the series of articles on Natural Language Processing…

5 min read · Apr 5

2

Build	Stable (2.0.0)	Preview (Nightly)		
	Linux	Mac	Windows	
	Conda	Pip	LibTorch	Source
e	Python	C++ / Java		
e Platform	CUDA 11.7	CUDA 11.8	~~ROCm 5.4.2~~	CPU
Command:	pip3 install torch torchvision torchaudio --index-url https://download.p orch.org/whl/cu117			

```
pip3 install torch torchvision torchaudio --index-url https://download.p
orch.org/whl/cu117
```

Bluetick Consultants

The Future of Automatic Speech Recognition — Whisper

Introduction

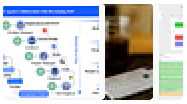
4 min read · Jun 2

1

Lists

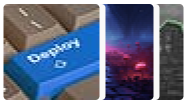
Coding & Development

11 stories · 200 saves



Natural Language Processing

669 stories · 283 saves

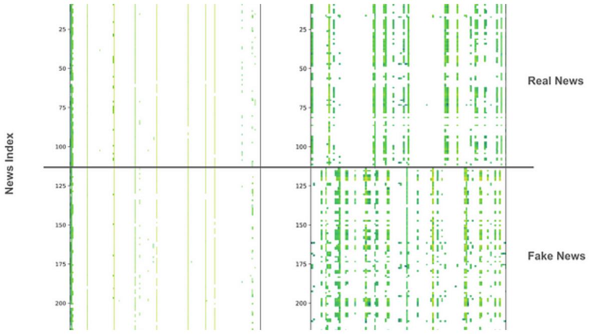


Predictive Modeling w/ Python

20 stories · 452 saves

Practical Guides to Machine Learning

10 stories · 519 saves



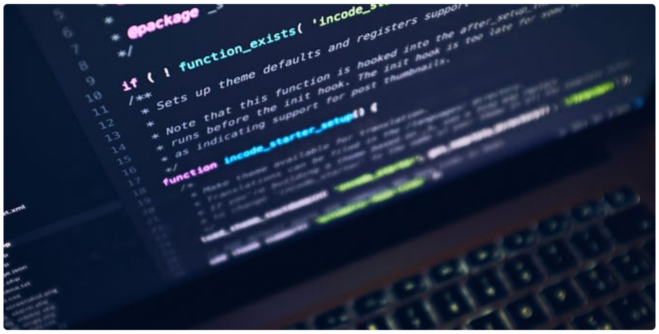
Sherry Wu in Stanford CS224W GraphML Tutorials

Spread No More: Twitter Fake News Detection with GNN

By Li Tian, Sherry Wu, Yifei Zheng as part of the Stanford CS224W course project.

11 min read · May 16

1



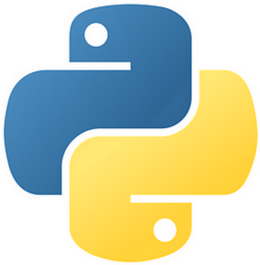
Zahrizhal Ali

Crafting Your Custom Text-to-Speech Model.

Welcome to a wild journey where code and comedy collide! In this blog, we're going to...

11 min read · Jun 2

200 2



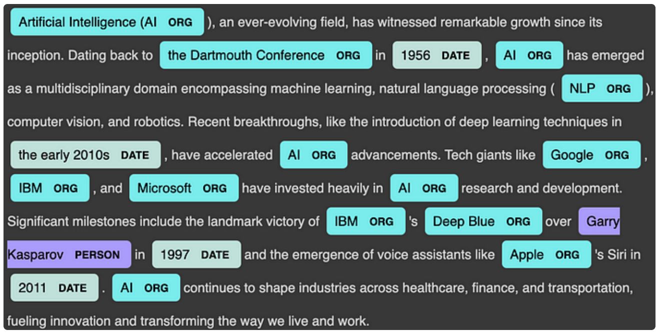
Informula in Dev Genius

How to Build a Word Collocation Network Graph via Python (Colab...

Previously on How to Build a Word Collocation Network Graph via Tableau & R? We...

2 min read · May 18

24



Artificial Intelligence (AI ORG), an ever-evolving field, has witnessed remarkable growth since its inception. Dating back to the Dartmouth Conference ORG in 1956 DATE , AI ORG has emerged as a multidisciplinary domain encompassing machine learning, natural language processing (NLP ORG), computer vision, and robotics. Recent breakthroughs, like the introduction of deep learning techniques in the early 2010s DATE , have accelerated AI ORG advancements. Tech giants like Google ORG , IBM ORG , and Microsoft ORG have invested heavily in AI ORG research and development. Significant milestones include the landmark victory of IBM ORG 's Deep Blue ORG over Garry Kasparov PERSON in 1997 DATE and the emergence of voice assistants like Apple ORG 's Siri in 2011 DATE . AI ORG continues to shape industries across healthcare, finance, and transportation, fueling innovation and transforming the way we live and work.

Sahithi Reddy

Named Entity Recognition using Python and Spacy.

NER, short for Named Entity Recognition, is an NLP method that detects and categorizes...

4 min read · Jun 11

24

See more recommendations