

named entity recogniton

March 14, 2022

1 Finding Entities in Multiple Sclerosis Research

This isn't as much to develop more for Gregory, it's to see what I can do with Spacy and Named Entity Recognition (NER). This is a Jupyter notebook because I want to try giving it a proper use and because it will make it easier to achieve two goals.

1. Show others what my thought process was.
2. Make it easier to ask questions to people who know more than me.
3. Discover what is the best NER model to analyse Multiple Sclerosis (MS) articles

1.1 Data sources

<https://api.gregory-ms.com/articles/all>

1.2 Initilize modules and get data

```
[ ]: import os
import scispacy
import spacy
import pandas as pd
import requests
from spacy import displacy
```

```
[ ]: url = 'https://api.gregory-ms.com/articles/all'

urlData = requests.get(url).content

df = pd.read_json(urlData)

print(df)
```

	article_id	title \
0	1138	The Relationship Between Walking Speed and the...
1	1139	Microglial changes associated with meningeal i...
2	1201	Association of neurogranin gene expression wit...
3	843	Depression in multiple sclerosis: Is one appro...
4	1145	An engineered neurovascular unit for modeling ...
...
7960	12696	Does the Serum Expression Level of High-Mobili...

7961	14071	The microbiota restrains neurodegenerative mic...
7962	14074	Autologous treatment for ALS with implication ...
7963	14538	Timed Up & Go (TUG) With Cognitive and Man...
7964	14804	Chromatin accessibility and transcriptome inte...

		summary \
0		<div><p style=x3D;quot;color: #4aa5...
1		<div><p style=x3D;quot;color: #4aa5...
2		<div><p style=x3D;quot;color: #4aa5...
3		<div><p style=x3D;quot;color: #4aa5...
4		<div><p style=x3D;quot;color: #4aa5...
...		...
7960		CONCLUSION: The serum level of HMGB1 could pre...
7961		The gut microbiota can affect neurologic disea...
7962		Amyotrophic lateral sclerosis (ALS) is charact...
7963		Timed Up & Go (TUG) With Cognitive and Man...
7964		Psoriasis is a chronic and hyperproliferative ...

		link \
0		https://pubmed.ncbi.nlm.nih.gov/33847188/?utm_...
1		https://pubmed.ncbi.nlm.nih.gov/33846617/?utm_...
2		https://pubmed.ncbi.nlm.nih.gov/33860070/?utm_...
3		https://pubmed.ncbi.nlm.nih.gov/33780807/?utm_...
4		https://pubmed.ncbi.nlm.nih.gov/33849004/?utm_...
...		...
7960		https://pubmed.ncbi.nlm.nih.gov/35263889/?utm_...
7961		https://microbiomejournal.biomedcentral.com/ar...
7962		https://translationalneurodegeneration.biomedc...
7963		https://www.apta.org/patient-care/evidence-bas...
7964		https://clinicalepigeneticsjournal.biomedcentr...

	published_date	relevant	discovery_date \
0	2021-01-06T00:00:00.000Z	NaN	2021-04-13T21:19:40.000Z
1	2021-01-05T00:00:00.000Z	NaN	2021-04-13T21:19:40.000Z
2	2021-04-08T23:00:00.000Z	NaN	2021-04-16T19:19:38.000Z
3	2021-01-06T00:00:00.000Z	NaN	2021-03-30T05:28:24.000Z
4	2021-05-04T23:00:00.000Z	NaN	2021-04-14T06:19:39.000Z
...
7960	2022-03-10T11:00:00.000Z	NaN	2022-03-10T12:42:30.855Z
7961	2022-03-11T00:00:00.000Z	NaN	2022-03-11T08:17:48.236Z
7962	2022-03-11T00:00:00.000Z	NaN	2022-03-11T08:17:48.488Z
7963	2022-03-11T14:20:18.802Z	NaN	2022-03-11T14:17:49.099Z
7964	2022-03-11T00:00:00.000Z	NaN	2022-03-11T18:17:48.562Z

	ml_prediction_gnb	ml_prediction_lr \
0	0.0	0.0
1	0.0	0.0
2	0.0	0.0

3	0.0	0.0
4	0.0	0.0
...
7960	0.0	0.0
7961	0.0	0.0
7962	0.0	0.0
7963	0.0	0.0
7964	0.0	0.0

	noun_phrases	source \
0	[The Relationship, Walking Speed, the Energeti...	PubMed
1	[Microglial changes, meningeal inflammation, m...	PubMed
2	[Association, neurogranin gene expression, Alz...	PubMed
3	[Depression, multiple sclerosis, one approach,...	PubMed
4	[An engineered neurovascular unit, neuroinflam...	PubMed
...
7960	[the Serum Expression Level, High-Mobility Gro...	PubMed
7961	[The microbiota, a model, amyotrophic lateral ...	BioMedCentral
7962	[Autologous treatment, ALS, implication, broad...	BioMedCentral
7963	[Timed, Go, TUG, Cognitive and Manual Tasks, M...	APTA
7964	[Chromatin accessibility, transcriptome integr...	BioMedCentral

	source__link	source__language \
0	https://pubmed.ncbi.nlm.nih.gov/rss/search/10g...	en
1	https://pubmed.ncbi.nlm.nih.gov/rss/search/10g...	en
2	https://pubmed.ncbi.nlm.nih.gov/rss/search/10g...	en
3	https://pubmed.ncbi.nlm.nih.gov/rss/search/10g...	en
4	https://pubmed.ncbi.nlm.nih.gov/rss/search/10g...	en
...
7960	https://pubmed.ncbi.nlm.nih.gov/rss/search/10g...	en
7961	https://www.biomedcentral.com/search?searchTyp...	en
7962	https://www.biomedcentral.com/search?searchTyp...	en
7963	https://www.apta.org/search?Q=%22Multiple+Scle...	en
7964	https://www.biomedcentral.com/search?searchTyp...	en

	sources__subject
0	Multiple Sclerosis
1	Multiple Sclerosis
2	Multiple Sclerosis
3	Multiple Sclerosis
4	Multiple Sclerosis
...	...
7960	Multiple Sclerosis
7961	Multiple Sclerosis
7962	Multiple Sclerosis
7963	Multiple Sclerosis
7964	Multiple Sclerosis

[7965 rows x 14 columns]

```
[ ]: summary = df.loc[0, 'summary']  
print(summary)
```

<div><p style="color: #4aa564;">Neurorehabil Neural Repair. 2021 Apr 13:15459683211005028. doi: 10.1177/15459683211005028. Online ahead of print.</p><p>ABSTRACT<p><math>\text{xmlns:xlink="http://www.w3.org/1999/xhtml" xmlns:mml="http://www.w3.org/1998/Math/MathML" xmlns:p1="http://pubmed.gov/pub-one">BACKGROUND: Persons with multiple sclerosis (pwMS) experience walking impairments, characterized by decreased walking speeds. In healthy subjects, the self-selected walking speed is the energetically most optimal. In pwMS, the energetically most optimal walking speed remains underexposed. Therefore, this review aimed to determine the relationship between walking speed and energetic cost of walking (Cw) in pwMS, compared with healthy subjects, thereby assessing the walking speed with the lowest energetic cost. As it is unclear whether the Cw in pwMS differs between overground and treadmill walking, as reported in healthy subjects, a second review aim was to compare both conditions.</p><p><math>\text{xmlns:xlink="http://www.w3.org/1999/xhtml" xmlns:mml="http://www.w3.org/1998/Math/MathML" xmlns:p1="http://pubmed.gov/pub-one">METHOD: PubMed and Web of Science were systematically searched. Studies assessing pwMS, reporting walking speed (converted to meters per second), and reporting oxygen consumption were included. Study quality was assessed with a modified National Heart, Lung and Blood Institute checklist. The relationship between Cw and walking speed was calculated with a second-order polynomial function and compared between groups and conditions.</p><p><math>\text{xmlns:xlink="http://www.w3.org/1999/xhtml" xmlns:mml="http://www.w3.org/1998/Math/MathML" xmlns:p1="http://pubmed.gov/pub-one">RESULTS: Twenty-nine studies were included (n = 1535 pwMS) of which 8 included healthy subjects (n = 179 healthy subjects). PwMS showed a similar energetically most optimal walking speed of 1.44 m/s with a Cw of 0.16, compared with 0.14 mL O₂/kg/m in healthy subjects. The most optimal walking speed in treadmill was 1.48 m/s, compared with 1.28 m/s in overground walking with a similar Cw.</p><p><math>\text{xmlns:xlink="http://www.w3.org/1999/xhtml" xmlns:mml="http://www.w3.org/1998/Math/MathML" xmlns:p1="http://pubmed.gov/pub-one">CONCLUSION: Overall, the Cw is elevated in pwMS but with a similar energetically most optimal walking speed, compared with healthy subjects. Treadmill walking showed a similar most optimal Cw but a higher speed, compared

with overground walking.<#x2F;p><p style#x3D;"color: lightgray">PMID:<a href#x3D;"https:#x2F;#x2F;pubmed.ncbi.nlm.nih.gov#x2F;33847188#x2F;?utm_source#x3D;Other&utm_medium#x3D;rss&utm_campaign#x3D;pubmed-2&utm_content#x3D;10guX6I3SqrBUeeLKSTD6FCRM44ewnrN2MKKTQLLPMHB4xNsZU&fc#x3D;20210216052009&ff#x3D;20210413171936&v#x3D;2.14.3">33847188<#x2F;a> | DOI:<a href#x3D;https:#x2F;#x2F;doi.org#x2F;10.1177#x2F;15459683211005028&10.1177#x2F;15459683211005028<#x2F;a><#x2F;p><#x2F;div>

Summary includes html, so we need to clean the data

```
[ ]: import html
summary = html.unescape(summary)

from bs4 import BeautifulSoup
soup = BeautifulSoup(summary, features="html.parser")
for script in soup(["script", "style"]):
    script.extract() # rip it out
summary = soup.get_text()
print(summary)
```

Neurorehabil Neural Repair. 2021 Apr 13:15459683211005028. doi: 10.1177/15459683211005028. Online ahead of print.ABSTRACTBACKGROUND: Persons with multiple sclerosis (pwMS) experience walking impairments, characterized by decreased walking speeds. In healthy subjects, the self-selected walking speed is the energetically most optimal. In pwMS, the energetically most optimal walking speed remains underexposed. Therefore, this review aimed to determine the relationship between walking speed and energetic cost of walking (Cw) in pwMS, compared with healthy subjects, thereby assessing the walking speed with the lowest energetic cost. As it is unclear whether the Cw in pwMS differs between overground and treadmill walking, as reported in healthy subjects, a second review aim was to compare both conditions.METHOD: PubMed and Web of Science were systematically searched. Studies assessing pwMS, reporting walking speed (converted to meters per second), and reporting oxygen consumption were included. Study quality was assessed with a modified National Heart, Lung and Blood Institute checklist. The relationship between Cw and walking speed was calculated with a second-order polynomial function and compared between groups and conditions.RESULTS: Twenty-nine studies were included (n = 1535 pwMS) of which 8 included healthy subjects (n = 179 healthy subjects). PwMS showed a similar energetically most optimal walking speed of 1.44 m/s with a Cw of 0.16, compared with 0.14 mL O₂/kg/m in healthy subjects. The most optimal walking speed in treadmill was 1.48 m/s, compared with 1.28 m/s in overground walking with a similar Cw.CONCLUSION: Overall, the Cw is elevated in pwMS but with a similar energetically most optimal walking speed, compared with healthy subjects. Treadmill walking showed a similar most optimal Cw but a higher speed, compared with overground walking.PMID:33847188 | DOI:10.1177/15459683211005028

Let's look at the output of 'en_core_sci_md' as a NER, and we'll see that it identifies entities, but does not show what they are.

```
[ ]: nlp = spacy.load('en_core_sci_md')
      doc = nlp(summary)
      displacy_image = displacy.render(doc, style = 'ent' )
```

<IPython.core.display.HTML object>

Same thing, this time with `en_ner_jnlpba_md` as a NER model, and we don't see any entities at all.

There are four models available for NER on science articles that we are going to test. These all come from [SciSpacy](#). 1. `en_ner_craft_md` 2. `en_ner_bc5cdr_md` 3. `en_ner_bionlp13cg_md` 4. `en_ner_jnlpba_md`

```
[ ]: nlp_cr = spacy.load('en_ner_craft_md')
      nlp_bc = spacy.load('en_ner_bc5cdr_md')
      nlp_bi = spacy.load('en_ner_bionlp13cg_md')
      nlp_jn = spacy.load('en_ner_jnlpba_md')
```

```
[ ]: doc = nlp_jn(summary)
      displacy_image = displacy.render(doc, style = 'ent')
```

```
/Users/brunoamaral/Labs/gregory/env/lib/python3.7/site-
packages/spacy/displacy/__init__.py:200: UserWarning: [W006] No entities to
visualize found in Doc object. If this is surprising to you, make sure the Doc
was processed using a model that supports named entity recognition, and check
the `doc.ents` property manually if necessary.
  warnings.warn(Warnings.W006)
```

<IPython.core.display.HTML object>

```
[ ]: doc = nlp_cr(summary)
      displacy_image = displacy.render(doc, style = 'ent')
```

<IPython.core.display.HTML object>

```
[ ]: doc = nlp_bc(summary)
      displacy_image = displacy.render(doc, style = 'ent')
```

<IPython.core.display.HTML object>

```
[ ]: doc = nlp_bi(summary)
      displacy_image = displacy.render(doc, style = 'ent')
```

<IPython.core.display.HTML object>

2 Conclusion so far

I don't think any of these will work