2.

Getting Started with spaCy and its Linguistic Annotations

Dr. W.J.B. Mattingly

Smithsonian Data Science Lab and United States Holocaust Memorial Museum

August 2021

In this chapter, we will start working with spaCy directly. The goals of this chapter are twofold. First, it is my hope that you understand the basic spaCy syntax for creating a Doc container and how to call specific attributes of that container. Second, it is my hope that you leave this chapter with a basic understanding of the vast linguistic annotations available in spaCy. While we will not explore all attributes, we will deal with many of the most important ones, such as lemmas, parts-of-speech, and named entities. By the time you are finished with this chapter, you should have enough of a basic understanding of spaCy to begin applying it to your own texts.

2.1. Importing spaCy and Loading Data

As with all Python libraries, the first thing we need to do is import spaCy. In the last notebook, I walked you through how to install it and download the small English model. If you have followed those steps, you should be able to import it like so:

```
IMPO:tensorflow:Enabling eager execution
INFO:tensorflow:Enabling v2 tensorshape
INFO:tensorflow:Enabling resource variables
INFO:tensorflow:Enabling tensor equality
INFO:tensorflow:Enabling control flow v2
```

With spaCy imported, we can now create our nlp object. This is the standard Pythonic way to create your model in a Python script. Unless you are working with multiple models in a script, try to always name your model, nlp. It will make your script much easier to read. To do this, we will use spacy.load(). This command tells spaCy to load up a model. In order to know which model to load, it needs a string argument that corresponds to the model name. Since we will be working with the small English model, we will use "en_core_web_sm". This function can take keyword arguments to identify which parts of the model you want to load, but we will get to that later. For now, we want to import the whole thing.

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

Great! With the model loaded, let's go ahead and import our text. For this chapter, we will be working with the opening description from the Wikipedia article on the United States. In this repo, it is found in the subfolder data and is entitled wiki_us.txt.

```
with open ("data/wiki_us.txt", "r") as f:
  text = f.read()
```

Now, let's see what this text looks like. It can be a bit difficult to read in a JupyterBook, but notice the horizontal slider below. You don't need to read this in its entirety.

```
print (text)
```

E Contents

- 2.1. Importing spaCy and L Print to PDF
- 2.2. Creating a Doc Container
- 2.3. Sentence Boundary Detection (SBD)
- 2.4. Token Attributes
 - 2.4.1. Text
 - 2.4.2. Head
 - 2.4.3. Left Edge
 - 2.4.4. Right Edge
 - 2.4.5. Entity Type
 - 2.4.6. Ent IOB
 - 2.4.7. Lemma
 - 2.4.8. Morph
 - 2.4.9. Part of Speech
 - 2.4.10. Syntactic Dependency
 - 2.4.11. Language
- 2.5. Part of Speech Tagging (POS)
- 2.6. Named Entity Recognition
- 2.7. Conclusion

The United States of America (U.S.A. or USA), commonly known as the United States (U.S. or US) or America, is a country primarily located in North America. It consists of 50 states, a federal district, five major unincorporated territories, 326 Indian reservations, and some minor possessions.[j] At 3.8 million square miles (9.8 million square kilometers), it is the world's third— or fourth—largest country by total area. [d] The United States shares significant land borders with Canada to the north and Mexico to the south, as well as limited maritime borders with the Bahamas, Cuba, and Russia.[22] With a population of more than 331 million people, it is the third most populous country in the world. The national capital is Washington, D.C., and the most populous city is New York.

Paleo—Indians migrated from Siberia to the North American mainland at least 12,000 years ago, and European colonization began in the 16th century. The United States emerged from the thirteen British colonies established along the East Coast. Disputes over taxation and political representation with Great Britain led to the American Revolutionary War (1775–1783), which established independence. In the late 18th century, the U.S. began expanding across North America, gradually obtaining new territories, sometimes through war, frequently displacing Native Americans, and admitting new states; by 1848, the United States spanned the continent. Slavery was legal in the southern United States until the second half of the 19th century when the American Civil War led to its abolition. The Spanish–American War and World War I established the U.S. as a world power, a status confirmed by the outcome of World War II.

During the Cold War, the United States fought the Korean War and the Vietnam War but avoided direct military conflict with the Soviet Union. The two superpowers competed in the Space Race, culminating in the 1969 spaceflight that first landed humans on the Moon. The Soviet Union's dissolution in 1991 ended the Cold War, leaving the United States as the world's sole superpower.

The United States is a federal republic and a representative democracy with three separate branches of government, including a bicameral legislature. It is a founding member of the United Nations, World Bank, International Monetary Fund, Organization of American States, NATO, and other international organizations. It is a permanent member of the United Nations Security Council. Considered a melting pot of cultures and ethnicities, its population has been profoundly shaped by centuries of immigration. The country ranks high in international measures of economic freedom, quality of life, education, and human rights, and has low levels of perceived corruption. However, the country has received criticism concerning inequality related to race, wealth and income, the use of capital punishment, high incarceration rates, and lack of universal health care.

The United States is a highly developed country, accounts for approximately a quarter of global GDP, and is the world's largest economy. By value, the United States is the world's largest importer and the second—largest exporter of goods. Although its population is only 4.2% of the world's total, it holds 29.4% of the total wealth in the world, the largest share held by any country. Making up more than a third of global military spending, it is the foremost military power in the world; and it is a leading political, cultural, and scientific force internationally.[23]

2.2. Creating a Doc Container

With the data loaded in, it's time to make our first Doc container. Unless you are working with multiple Doc containers, it is best practice to always call this object "doc", all lowercase. To create a doc container, we will usually just call our nlp object and pass our text to it as a single argument.

doc = nlp(text)

Great! Let's see what this looks like.

print (doc)

The United States of America (U.S.A. or USA), commonly known as the United States (U.S. or US) or America, is a country primarily located in North America. It consists of 50 states, a federal district, five major unincorporated territories, 326 Indian reservations, and some minor possessions.[j] At 3.8 million square miles (9.8 million square kilometers), it is the world's third— or fourth—largest country by total area. [d] The United States shares significant land borders with Canada to the north and Mexico to the south, as well as limited maritime borders with the Bahamas, Cuba, and Russia.[22] With a population of more than 331 million people, it is the third most populous country in the world. The national capital is Washington, D.C., and the most populous city is New York.

Paleo—Indians migrated from Siberia to the North American mainland at least 12,000 years ago, and European colonization began in the 16th century. The United States emerged from the thirteen British colonies established along the East Coast. Disputes over taxation and political representation with Great Britain led to the American Revolutionary War (1775–1783), which established independence. In the late 18th century, the U.S. began expanding across North America, gradually obtaining new territories, sometimes through war, frequently displacing Native Americans, and admitting new states; by 1848, the United States spanned the continent. Slavery was legal in the southern United States until the second half of the 19th century when the American Civil War led to its abolition. The Spanish–American War and World War I established the U.S. as a world power, a status confirmed by the outcome of World War II.

During the Cold War, the United States fought the Korean War and the Vietnam War but avoided direct military conflict with the Soviet Union. The two superpowers competed in the Space Race, culminating in the 1969 spaceflight that first landed humans on the Moon. The Soviet Union's dissolution in 1991 ended the Cold War, leaving the United States as the world's sole superpower.

The United States is a federal republic and a representative democracy with three separate branches of government, including a bicameral legislature. It is a founding member of the United Nations, World Bank, International Monetary Fund, Organization of American States, NATO, and other international organizations. It is a permanent member of the United Nations Security Council. Considered a melting pot of cultures and ethnicities, its population has been profoundly shaped by centuries of immigration. The country ranks high in international measures of economic freedom, quality of life, education, and human rights, and has low levels of perceived corruption. However, the country has received criticism concerning inequality related to race, wealth and income, the use of capital punishment, high incarceration rates, and lack of universal health care.

The United States is a highly developed country, accounts for approximately a quarter of global GDP, and is the world's largest economy. By value, the United States is the world's largest importer and the second—largest exporter of goods. Although its population is only 4.2% of the world's total, it holds 29.4% of the total wealth in the world, the largest share held by any country. Making up more than a third of global military spending, it is the foremost military power in the world; and it is a leading political, cultural, and scientific force internationally.[23]

If you are trying to spot the difference between this and the text above, good luck. You will not see a difference when printing off the doc container. But I promise you, it is quite different behind the scenes. The Doc container, unlike the text object, contains a lot of valuable metadata, or attributes, hidden behind it. To prove this, let's examine the length of the doc object and the text object.

```
print (len(doc))
print (len(text))

652
3525
```

Hmm... What's going on here? Same text, but different length. Why does this occur? To answer that, let's explore it more deeply and try and print off each item in each object.

```
for token in text[:10]:
    print (token)

T
h
e
U
n
i
t
e
d
```

As we would expect. We have printed off each character, including white spaces. Let's try and do the same with the Doc container.

```
for token in doc[:10]:
    print (token)
The
United
States
of
America
(
U.S.A.
or
USA
)
```

And now we see the magical difference. While on the surface it may seem that the Doc container's length is dependent on the quantity of words, look more closely. You should notice that the open and close parentheses are also considered an item in the container. These are all known as tokens. **Tokens** are a fundamental building block of spaCy or any NLP framework. They can be words or punctuation marks. Tokens are something that has syntactic purpose in a sentence and is self-contained. A good example of this is the contraction "don't" in English. When tokenized, or the process of converting the text into tokens, we will have two tokens. "do" and "n't" because the contraction represents two words, "do" and "not".

On the surface, this may not seem exceptional. But it is. You may be thinking to yourself that you could easily use the split method in Python to split by whitespace and have the same result. But you'd be wrong. Let's see why.

```
for token in text.split()[:10]:
    print (token)
The
United
States
of
America
(U.S.A.
or
USA),
commonly
known
```

Notice that the parentheses are not removed or handled individually. To see this more clearly, let's print off all tokens from index 5 to 8 in both the text and doc objects.

```
words = text.split()[:10]

i=5
for token in doc[i:8]:
    print (f"SpaCy Token {i}:\n{token}\nWord Split {i}:\n{words[i]}\n\n")
    i=i+1

SpaCy Token 5:
    (
    Word Split 5:
    (U.S.A.

SpaCy Token 6:
    U.S.A.
Word Split 6:
    or

SpaCy Token 7:
    or
    Word Split 7:
    USA),
```

We can see clearly now how the spaCy Doc container does much more with its tokenization than a simple split method. We could, surely, write complex rules for a language to achieve the same results, but why bother? SpaCy does it exceptionally well for all languages. In my entire time using spaCy, I have never seen the tokenizer make a mistake. I am sure that mistakes may occur, but these are probably rare exceptions.

Let's see what else this Doc Container holds.

2.3. Sentence Boundary Detection (SBD)

In NLP, sentence boundary detection, or SBD, is the identification of sentences in a text. Again, this may seem fairly easy to do with rules. One could use split("."), but in English we use the period to also denote abbreviation. You could, again, write rules to look for periods not proceeded by a lowercase word, but again, I ask the question, "why bother?". We can use spaCy and in seconds have all sentences fully separated through SBD.

To access the sentences in the Doc container, we can use the attribute sents, like so:

for sent in doc.sents:
 print (sent)

The United States of America (U.S.A. or USA), commonly known as the United States (U.S. or US) or America, is a country primarily located in North America. It consists of 50 states, a federal district, five major unincorporated territories, 326 Indian reservations, and some minor possessions.[j]

At 3.8 million square miles (9.8 million square kilometers), it is the world's third-or fourth-largest country by total area.[d] The United States shares significant land borders with Canada to the north and Mexico to the south, as well as limited maritime borders with the Bahamas, Cuba, and Russia.[22] With a population of more than 331 million people, it is the third most populous country in the world.

The national capital is Washington, D.C., and the most populous city is New York.

Paleo-Indians migrated from Siberia to the North American mainland at least 12,000 years ago, and European colonization began in the 16th century.

The United States emerged from the thirteen British colonies established along the East Coast.

Disputes over taxation and political representation with Great Britain led to the American Revolutionary War (1775–1783), which established independence. In the late 18th century, the U.S. began expanding across North America, gradually obtaining new territories, sometimes through war, frequently displacing Native Americans, and admitting new states; by 1848, the United States spanned the continent.

Slavery was legal in the southern United States until the second half of the 19th century when the American Civil War led to its abolition.

The Spanish–American War and World War I established the U.S. as a world power, a status confirmed by the outcome of World War II.

During the Cold War, the United States fought the Korean War and the Vietnam War but avoided direct military conflict with the Soviet Union.

The two superpowers competed in the Space Race, culminating in the 1969 spaceflight that first landed humans on the Moon.

The Soviet Union's dissolution in 1991 ended the Cold War, leaving the United States as the world's sole superpower.

The United States is a federal republic and a representative democracy with three separate branches of government, including a bicameral legislature.

It is a founding member of the United Nations, World Bank, International Monetary Fund, Organization of American States, NATO, and other international organizations. It is a permanent member of the United Nations Security Council.

Considered a melting pot of cultures and ethnicities, its population has been profoundly shaped by centuries of immigration.

The country ranks high in international measures of economic freedom, quality of life, education, and human rights, and has low levels of perceived corruption. However, the country has received criticism concerning inequality related to race, wealth and income, the use of capital punishment, high incarceration rates, and lack of universal health care.

The United States is a highly developed country, accounts for approximately a quarter of global GDP, and is the world's largest economy.

By value, the United States is the world's largest importer and the second-largest exporter of goods.

Although its population is only 4.2% of the world's total, it holds 29.4% of the total wealth in the world, the largest share held by any country.

Making up more than a third of global military spending, it is the foremost military power in the world; and it is a leading political, cultural, and scientific force internationally.[23]

Let's move forward with just one of these sentences. Let's try and grab index 0 in this attribute.

sentence1 = doc.sents[0]
print (sentence1)

Uh oh! We got an error. That is because the sents attribute is a generator. In python, we can usually iterate over generators by converting them into a list. So, let's do that.

```
sentence1 = list(doc.sents)[0]
print (sentence1)
```

The United States of America (U.S.A. or USA), commonly known as the United States (U.S. or US) or America, is a country primarily located in North America.

Now we have the first sentence. Now that we have a smaller text, let's explore spaCy's other building block, the token.

2.4. Token Attributes

The token object contains a lot of different attributes that are VITAL do performing NLP in spaCy. We will be working with a few of them, such as:

- .text
- .head
- .left_edge
- .right_edge
- .ent_type_
- .iob_
- .lemma_
- .morph
- .pos_
- .dep_
- .lang_

I will briefly describe these here and show you how to grab each one and what they look like. We will be exploring each of these attributes more deeply in this chapter and future chapters. To demonstrate each of these attributes, we will use one token, "States" which is part of a sequence of tokens that make up "The United States of America"

```
token2 = sentence1[2]
print (token2)
States
```

2.4.1. Text

Verbatim text content. -spaCy docs

```
token2.text
'States'
```

2.4.2. Head

The syntactic parent, or "governor", of this token. -spaCy docs

```
token2.head
is
```

This tells to which word it is governed by, in this case, the primary verb, "is", as it is part of the noun subject.

2.4.3. Left Edge

The leftmost token of this token's syntactic descendants. -spaCy docs

```
token2.left_edge

The
```

If part of a sequence of tokens that are collectively meaningful, known as **multi-word tokens**, this will tell us where the multi-word token begins.

2.4.4. Right Edge

The rightmost token of this token's syntactic descendants. -spaCy docs

```
token2.right_edge

America
```

This will tell us where the multi-word token ends.

2.4.5. Entity Type

Named entity type. -spaCy docs

```
token2.ent_type

384
```

Note the absence of the _ at the end of the attribute. This will return an integer that corresponds to an entity type, where as _ will give you the string equivalent., as in below.

```
token2.ent_type_
'GPE'
```

We will learn all about types of entities in our chapter on named entity recognition, or NER. For now, simply understand that GPE is geopolitical entity and is correct.

2.4.6. Ent IOB

IOB code of named entity tag. "B" means the token begins an entity, "I" means it is inside an entity, "O" means it is outside an entity, and "" means no entity tag is set.

```
token2.ent_iob_
'I'
```

IOB is a method of annotating a text. In this case, we see "I" because states is inside an entity, that is to say that it is part of the United States of America.

2.4.7. Lemma

Base form of the token, with no inflectional suffixes. -spaCy docs

```
token2.lemma_
```

```
'States'

sentence1[12].lemma_

'know'
```

2.4.8. Morph

Morphological analysis -spaCy docs

```
sentence1[12].morph

Aspect=Perf|Tense=Past|VerbForm=Part
```

2.4.9. Part of Speech

Coarse-grained part-of-speech from the Universal POS tag set.-spaCy docs

```
token2.pos_
'PROPN'
```

2.4.10. Syntactic Dependency

Syntactic dependency relation. -spaCy docs

```
token2.dep_
'nsubj'
```

2.4.11. Language

Language of the parent document's vocabulary. -spaCy docs

```
token2.lang_
'en'
```

2.5. Part of Speech Tagging (POS)

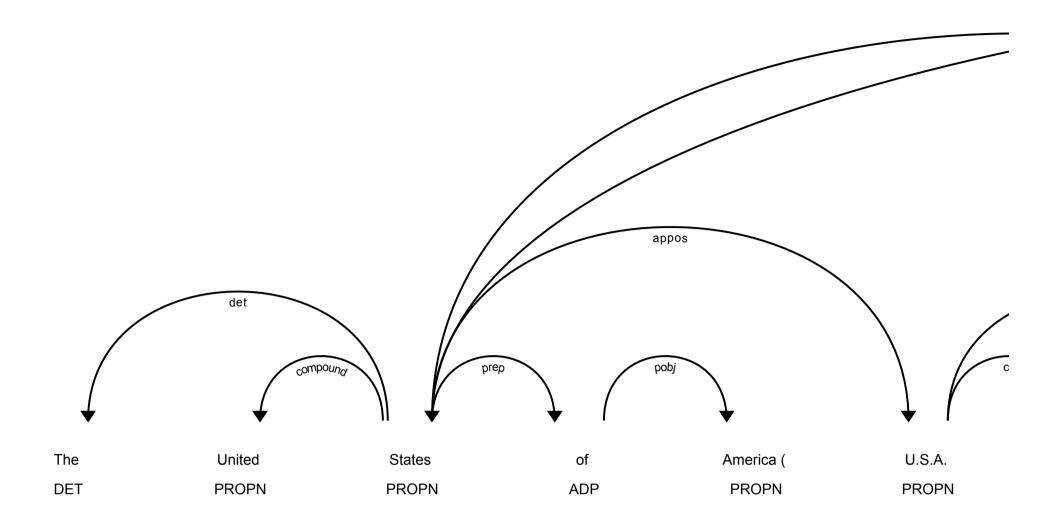
In the field of computational linguistics, understanding parts-of-speech is essential. SpaCy offers an easy way to parse a text and identify its parts of speech. Below, we will iterate across each token (word or punctuation) in the text and identify its part of speech.

```
for token in sentence1:
    print (token.text, token.pos_, token.dep_)
```

```
The DET det
United PROPN compound
States PROPN nsubj
of ADP prep
America PROPN pobj
( PUNCT punct
U.S.A. PROPN appos
or CCONJ cc
USA PROPN conj
) PUNCT punct
, PUNCT punct
commonly ADV advmod
known VERB acl
as ADP prep
the DET det
United PROPN compound
States PROPN pobj
( PUNCT punct
U.S. PROPN appos
or CCONJ cc
US PROPN conj
) PUNCT punct
or CCONJ cc
America PROPN conj
, PUNCT punct
is AUX ROOT
a DET det
country NOUN attr
primarily ADV advmod
located VERB acl
in ADP prep
North PROPN compound
America PROPN pobj
. PUNCT punct
```

Here, we can see two vital pieces of information: the string and the corresponding part-of-speech (pos). For a complete list of the pos labels, see the spaCy documentation (https://spacy.io/api/annotation#pos-tagging). Most of these, however, should be apparent, i.e. PROPN is proper noun, AUX is an auxiliary verb, ADJ, is adjective, etc. We can visualize this sentence with a diagram through spaCy's displaCy Notebook feature.

```
from spacy import displacy
displacy.render(sentence1, style="dep")
```



2.6. Named Entity Recognition

Another essential task of NLP, is named entity recognition, or NER. I spoke about NER in the last notebook. Here, I'd like to demonstrate how to perform basic NER via spaCy. Again, we will iterate over the doc object as we did above, but instead of iterating over doc.sents, we will iterate over doc.ents. For our purposes right now, I simply want to print off each entity's text (the string itself) and its corresponding label (note the _ after label). I will be explaining this process in much greater detail in the next two notebooks.

```
for ent in doc.ents:
    print (ent.text, ent.label_)
```

```
The United States of America GPE
U.S.A. GPE
USA GPE
the United States GPE
U.S. GPE
US GPE
America GPE
North America LOC
50 CARDINAL
five CARDINAL
326 CARDINAL
Indian NORP
3.8 million square miles QUANTITY
9.8 million square kilometers QUANTITY
fourth ORDINAL
The United States GPE
Canada GPE
Mexico GPE
Bahamas GPE
Cuba GPE
more than 331 million CARDINAL
third ORDINAL
Washington GPE
D.C. GPE
New York GPE
Paleo-Indians NORP
Siberia LOC
North American NORP
at least 12,000 years ago DATE
European NORP
the 16th century DATE
The United States GPE
thirteen CARDINAL
British NORP
the East Coast LOC
Great Britain GPE
the American Revolutionary War ORG
the late 18th century DATE
U.S. GPE
North America LOC
Native Americans NORP
1848 DATE
the United States GPE
United States GPE
the second half of the 19th century DATE
the American Civil War EVENT
The Spanish–American War and World War EVENT
U.S. GPE
World War II EVENT
the Cold War EVENT
the United States GPE
the Korean War EVENT
the Vietnam War EVENT
the Soviet Union GPE
two CARDINAL
the Space Race FAC
1969 DATE
first ORDINAL
The Soviet Union's GPE
1991 DATE
the Cold War EVENT
the United States GPE
The United States GPE
three CARDINAL
the United Nations ORG
World Bank ORG
International Monetary Fund ORG
Organization of American States ORG
NATO ORG
the United Nations Security Council ORG
centuries DATE
The United States GPE
approximately a quarter DATE
the United States GPE
second ORDINAL
only 4.2% PERCENT
29.4% PERCENT
more than a third CARDINAL
```

Sometimes it can be difficult to read this output as raw data. In this case, we can again leverage spaCy's displaCy feature. Notice that this time we are altering the keyword argument, style, with the string "ent". This tells displaCy to display the text as NER annotations

```
displacy.render(doc, style="ent")
```

The United States of America GPE (U.S.A. GPE or USA GPE), commonly known as the United States GPE (U.S. GPE or US GPE) or America GPE , is a country primarily located in North America LOC . It consists of 50 CARDINAL states, a federal district, five CARDINAL major unincorporated territories, 326 CARDINAL Indian NORP reservations, and some minor possessions.[j] At 3.8 million square miles QUANTITY (9.8 million square kilometers QUANTITY), it is the world's third- or fourth ORDINAL -largest country by total area. [d] The United States GPE shares significant land borders with Canada GPE to the north and Mexico GPE to the south, as well as limited maritime borders with the Bahamas GPE , Cuba GPE , and Russia.[22] With a population of more than 331 million CARDINAL people, it is the third ORDINAL most populous country in the world. The national capital is Washington GPE ,

Paleo-Indians NORP migrated from Siberia Loc to the North American NORP mainland at least 12,000 years ago DATE , and European NORP colonization began in the 16th century DATE . The United States GPE emerged from the thirteen CARDINAL British NORP colonies established along the East Coast Loc . Disputes over taxation and political representation with Great Britain GPE led to the American Revolutionary War ORG (1775–1783), which established independence. In the late 18th century DATE , the U.S. GPE began expanding across North America Loc , gradually obtaining new territories, sometimes through war, frequently displacing Native Americans NORP , and admitting new states; by 1848 DATE , the United States GPE spanned the continent. Slavery was legal in the southern United States GPE until the second half of the 19th century DATE when the American Civil War EVENT led to its abolition. The Spanish–American War and World War EVENT I established the U.S. GPE as a world power, a status confirmed by the outcome of World War II

During the Cold War EVENT , the United States GPE fought the Korean War EVENT and the Vietnam War EVENT but avoided direct military conflict with the Soviet Union GPE . The two CARDINAL superpowers competed in the Space Race FAC , culminating in the 1969

DATE spaceflight that first ORDINAL landed humans on the Moon. The Soviet Union's GPE dissolution in 1991 DATE ended the Cold War EVENT , leaving the United States GPE as the world's sole superpower.

The United States GPE is a federal republic and a representative democracy with three CARDINAL separate branches of government, including a bicameral legislature. It is a founding member of the United Nations ORG, World Bank ORG, International Monetary Fund ORG, Organization of American States ORG, NATO ORG, and other international organizations. It is a permanent member of the United Nations Security Council ORG. Considered a melting pot of cultures and ethnicities, its population has been profoundly shaped by centuries DATE of immigration. The country ranks high in international measures of economic freedom, quality of life, education, and human rights, and has low levels of perceived corruption. However, the country has

received criticism concerning inequality related to race, wealth and income, the use of capital punishment, high incarceration rates, and lack of universal health care.

The United States GPE is a highly developed country, accounts for approximately a quarter DATE of global GDP, and is the world's largest economy. By value, the United States GPE is the world's largest importer and the second ORDINAL -largest exporter of goods. Although its population is only 4.2% PERCENT of the world's total, it holds 29.4% PERCENT of the total wealth in the world, the largest share held by any country. Making up more than a third CARDINAL of global military spending, it is the foremost military power in the world; and it is a leading political, cultural, and scientific force internationally.[23]

2.7. Conclusion

I recommend spending a little bit of time going through this notebook a few times. The information covered throughout this notebook will be reinforced as we explore each of these areas in more depth with real-world examples of how to implement them to tackle different problems.

By William Mattingly
© Copyright 2021.