Homework 5: Named Entity Recognition (NER)

Points: 20 | **Due:** Sunday, March 1, 2026 @ 11pm Pacific

Author: Richard Young, Ph.D. | UNLV Lee Business School

Compute: CPU (free tier) — GPU recommended for bonus

Learning Objectives

1. **Understand** Named Entity Recognition concepts
2. **Use** spaCy for entity extraction
3. **Extract** business-relevant entities from text data
4. **Analyze** entity patterns in reviews
5. **Visualize** entity distributions

Why This Matters for Business

Brand Monitoring: Companies like Coca-Cola use NER to automatically track mentions of their brands, competitors, and products across millions of social media posts daily. Manual tracking at this scale is impossible.

Financial Analysis: Bloomberg's terminal uses NER to extract company names, executives, and monetary values from news articles—feeding trading algorithms and analyst dashboards in real-time.

Legal Discovery: Law firms use NER to scan thousands of documents for people, organizations, and dates relevant to a case. What took paralegals weeks now takes hours.

Supply Chain Intelligence: Walmart monitors news for location entities (GPE) to predict disruptions—a hurricane heading toward a supplier's city triggers automatic inventory adjustments.

Grading

Component	Points	Effort	What We're Looking For
Setup & Data	3	*	spaCy loaded, data ready
NER Demo	4	*	Understand entity types
Entity Extraction	6	**	Extract ORG, PRODUCT, GPE from reviews
Analysis	4	**	Identify patterns in entity mentions
Visualization	3	*	Clear bar charts of top entities
Total	**20**		

Effort Key: * Straightforward | ** Requires thinking | *** Challenge

---

Entity Types

spaCy recognizes these business-relevant entities:

Entity	Description	Example
ORG	Organizations	“Apple”, “Amazon”
PRODUCT	Products	“iPhone”, “Kindle”
GPE	Countries, cities	“New York”, “USA”
PERSON	People names	“Tim Cook”
MONEY	Monetary values	“$50”
DATE	Dates and periods	“September 2024”
EVENT	Named events	“Olympics”, “Black Friday”

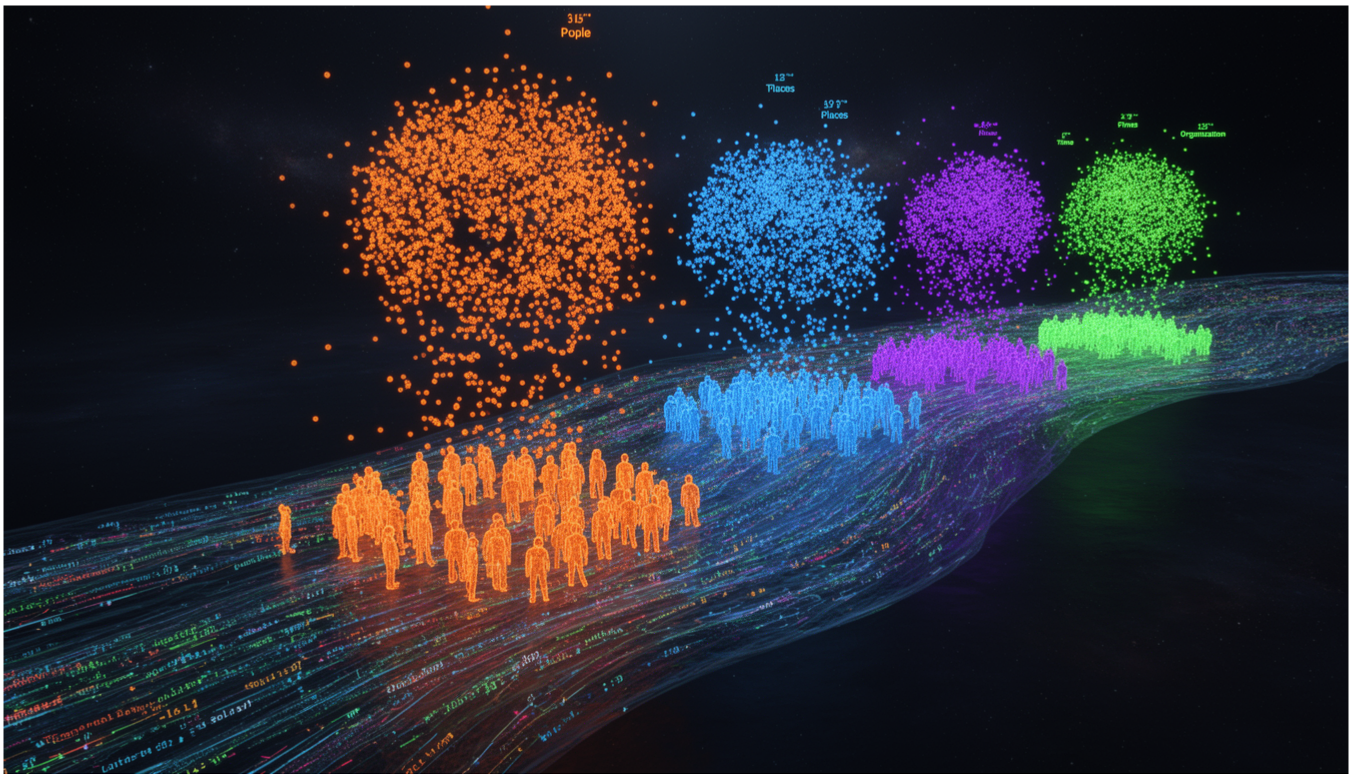

Figure 1: Named Entity Recognition

Instructions

1. Open `MIS769_HW5_NER.ipynb` in Google Colab
2. Load spaCy and understand entity types
3. Extract entities from your review dataset
4. Analyze: Which organizations/products are mentioned most?
5. Visualize the top entities by category

What Your Output Should Look Like

NER Demo Output:

```
  NAMED ENTITIES FOUND
------------------------------------------------------------
Apple                 | ORG        | Companies, agencies, institutions
Tim Cook              | PERSON     | People, including fictional
iPhone 15             | PRODUCT    | Objects, vehicles, foods, etc.
San Francisco         | GPE        | Countries, cities, states
$999                  | MONEY      | Monetary values
September 22, 2024    | DATE       | Absolute or relative dates
```

Entity Counts:

```
  Extraction complete!
   ORG: 8,234
   PRODUCT: 1,892
   GPE: 4,567
   PERSON: 12,453
```

Top Organizations:

```
  TOP ORGANIZATIONS
----------------------------------------
Hollywood                      | 1,247
BBC                            | 523
HBO                            | 412
Disney                         | 389
Netflix                        | 298
```

Common Mistakes (and How to Avoid Them)

Mistake	Symptom	Fix
Not downloading spaCy model	`OSError: Can't find model`	Run `!python -m spacy download en_core_web_sm`
Processing very long texts	Slow/crashes	Truncate: `text[:5000]`
Expecting perfect accuracy	Frustration	NER is ~85% accurate; document errors
Forgetting to cast to string	`TypeError` on null values	Use `str(text)`
Case sensitivity confusion	"apple" (fruit) vs "Apple" (company)	NER uses context, not just capitalization

If you see this error:

```
OSError: [E050] Can't find model 'en_core_web_sm'
```

Run:

```
!python -m spacy download en_core_web_sm
```

Then restart runtime (Runtime Restart runtime)

If extraction is too slow: - Process a sample: `df.sample(1000)` - Truncate long texts: `text[:5000]` - Use `nlp.pipe()` for batch processing (faster)

Questions to Answer

- **Q1:** Were the extracted entities accurate? What errors did you observe?
- **Q2:** What business insights can you derive from entity mentions?
- **Q3:** How could you improve NER for your specific domain?

Going Deeper (Optional Challenges)

Challenge A: Entity Co-occurrence Network

Build a network graph showing which entities appear together. Do certain companies get mentioned alongside certain people? Use `networkx` to visualize.

Challenge B: Sentiment by Entity

For each organization mentioned, calculate the average sentiment of reviews that mention it. Which companies have positive vs. negative associations?

Challenge C: Custom Entity Training

Train a custom NER model to recognize domain-specific entities not in spaCy's default model (e.g., FEATURE, COMPETITOR, PRODUCT_LINE). Requires the `en_core_web_trf` model and GPU.

Quick Reference

```python
# Load spaCy
import spacy
nlp = spacy.load("en_core_web_sm")

# Process text
doc = nlp("Apple CEO Tim Cook announced the new iPhone in San Francisco.")

# Extract entities
for ent in doc.ents:
    print(f"{ent.text:20} | {ent.label_:10} | {spacy.explain(ent.label_)}")

# Get specific entity types
orgs = [ent.text for ent in doc.ents if ent.label_ == "ORG"]
people = [ent.text for ent in doc.ents if ent.label_ == "PERSON"]
```

```
# Batch processing (faster for large datasets)
texts = ["Text 1", "Text 2", "Text 3"]
for doc in nlp.pipe(texts, batch_size=50):
    entities = [(ent.text, ent.label_) for ent in doc.ents]

# Count entities
from collections import Counter
all_orgs = []
for doc in nlp.pipe(texts):
    all_orgs.extend([ent.text for ent in doc.ents if ent.label_ == "ORG"])
org_counts = Counter(all_orgs).most_common(10)

# Visualize entities in notebook
from spacy import displacy
displacy.render(doc, style="ent", jupyter=True)
```

Entity Labels Cheat Sheet: | Label | Meaning | Business Use | |——-|———|————–| | ORG | Organization | Competitor tracking | | PERSON | Person name | Executive mentions | | GPE | Geo-political entity | Market analysis | | PRODUCT | Product name | Product mentions | | MONEY | Currency amounts | Pricing intelligence | | DATE | Date/time | Timeline analysis | | EVENT | Named event | Event impact | | NORP | Nationality/group | Demographic insights |

Submission

Upload to Canvas: - Your completed `.ipynb` notebook with all cells executed

---

— Richard Young, Ph.D.