# Homework 3: Text Processing Fundamentals

**Points:** 20 | **Due:** Sunday, February 15, 2026 @ 11pm Pacific

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

**Compute:** CPU (free tier)

---

## Learning Objectives

1. **Install and use** NLP libraries (spaCy, NLTK) in Google Colab
2. **Understand** WHY we preprocess text (not just how)
3. **Create** domain-specific stopwords for your data
4. **Measure** the impact of text cleaning on analysis
5. **Identify** cases where cleaning hurts your analysis

---

## Grading

| Component | Points | What We're Looking For |
|---|---|---|
| Environment Setup | 3 | NLP libraries installed and working |
| Standard Stopwords | 4 | Applied and analyzed impact |
| Domain Stopwords | 5 | Created 10+ with justification |
| Negation Analysis | 5 | Smart stopwords preserving meaning |
| Visualization | 3 | Word cloud comparison |
| **Total** | **20** | |

---

## The Big Picture

Text preprocessing is like preparing ingredients before cooking. But just like cooking, **the same preparation isn't right for every dish.**

Standard stopword removal can destroy important meaning (like negations in sentiment analysis).

---

## Instructions

1. Open `MIS769_HW3_Text_Processing.ipynb` in Google Colab
2. Load your dataset (HuggingFace, Kaggle, or your own)
3. Apply standard stopword removal and analyze what was removed
4. Create your own domain-specific stopwords with justification
5. Analyze the negation problem and create a "smart" stopword list
6. Generate word cloud visualizations comparing approaches

---

## Questions to Answer

- **Q1:** Which removed stopwords might carry meaning in your domain?
- **Q2:** Why did you choose each domain-specific stopword?
- **Q3:** When should you preserve vs. remove negations?

---

## Submission

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

---

— *Richard Young, Ph.D.*