# PACE Documentation Contribution Template

## Document Metadata

* **Author**: Patricia (Suzanne) Eastwood
* **Date Created**: 2025-04-29
* **Last Updated**: 2025-04-29
* **Applicable Clusters**: ICE

## D&O Litigation NLP: Multi-Model Legal Text Annotation Pipeline

### Overview

This NLP pipeline automates annotation and similarity scoring for legal documents related to D&O insurance litigation. It supports research on identifying coverage issues such as “late notice” and “what is a claim” by comparing document text chunks against gold standard annotations. The pipeline uses multiple model strategies including SBERT, Fuzzy Matching, LegalBERT, ModernBERT, CUADBERT, Saliency Maps, and Pairwise Sentence Comparison. It assists in automating tedious legal annotation tasks and is relevant to researchers working on legal AI, document classification, or insurance law analytics.

### Prerequisites

- Required access/permissions:
  - Active PACE account
  - ICE OnDemand access
- Software dependencies:
  - Python ≥ 3.10
  - sentence-transformers ≥ 2.2
  - transformers ≥ 4.30
  - pandas
  - tqdm
  - scikit-learn
  - matplotlib
  - thefuzz

- Storage requirements:
  - ≥ 5 GB (intermediate and merged results files are large)
  - Filepaths should be placed in /home/<gtID>/pace_dno_nlp/
- Other prerequisites:
  - Annotated gold chunks in gold_chunks.txt
  - Input data in chunks.csv

### Initial PACE Instructions

1. Download GlobalProtect VPN Client from vpn.gatech.edu
   a. See this link for more detailed install instructions: https://gatech.service-now.com/home?id=kb_article_view&sysparm_article=KB0042139

2. Open GlobalConnect and enter vpn.gatech.edu, then click Connect.

3. Enter your school credentials and perform your preferred 2FA. It should be identical to logging into buzzport. Example Username: peastwood3

4. The end result will show it to be connected
5. You can then connect to ICE clusters via your preferred command line using ‘ssh <username>@login-ice.pace.gatech.edu. You will be prompted to enter your password.
   In this example, the following would be entered into the command prompt:
   >ssh peastwood3@login-ice.pace.gatech.edu
   Note that your password should be identical to your Buzzport password.

~~~~
###

### Step-by-Step Instructions

1. **Log in and Launch VS Code on ICE**
   # Go to https://ondemand-ice.pace.gatech.edu
   # Choose Interactive Apps → VS Code → Launch with 4 cores / 16 GB

2. **Clone your repo or upload local files**
   # Create a folder pace_dno_nlp/ and place all .py, .csv, and .txt
   files inside.
```bash
   git clone https://github.gatech.edu/calexander97/law-data-design.git
   cd law-data-design/dno
   pip install -e .


3. **Python environment**
   # Create a new Python environment
```bash
   conda create -n dno python=3.11
   conda activate dno
   # In the future, just run conda activate dno


3. **Download the Cached Data**
```bash
   dno download-cached-data

4. **Install necessary packages**
```bash
   pip install -r requirements.txt

5. **Merge Outputs Run and Visualize**
```bash
   python modelmerge.py
   python plot_model_scores.py


### Configuration Details
1. Configuration File Setup
   ```yaml
   # Models:
~~~~

  - modernbert: "all-MiniLM-L6-v2"
  - legalbert: "nlpaueb/legal-bert-base-uncased"
  - fuzzy: Levenshtein ratio (thresholds 62, 80)
  - pairwise: Cosine similarity vs gold annotations
  - saliency: Precomputed attention-weighted scores   ```

### Troubleshooting

#### Common Issue 1: Output file >2GB crashes Excel
Error Message:
```
Nothing happens – it just freezes
```

Resolution:
1. Use df.sample(n=1000) to debug on a subset
2. df.to_csv("final_combined_sample.csv", index=False, compression="gzip")

#### Common Issue 2: CUDA Memory Overflow
```
Nothing happens – it just freezes
```
Resolution:
1. Lower batch size or use CPU-only mode
2. Ensure no other jobs are running on the node

### Storage and Resource Considerations
- Disk Space Requirements:
  - Temporary storage: 50 GB
  - Permanent storage: 20 GB
- Memory Usage:
  - Minimum: 8 GB
  - Recommended: 16-32 GB
- CPU Requirements:
  - Minimum cores: 4
  - Optimal performance: 4-8 cores recommended
- Quota Impact:
  - Can potentially time out

### Directory Structure
```
project/
pace_dno_nlp/
├── gold_chunks.txt
├── chunks.csv
├── generate_modernbert_late_notice_and_claim.py
├── modelmerge.py
├── outputs/
│   ├── final_combined_all_models.csv
│   ├── saliency_scores.csv
│   └```
│

### Additional Resources
```

- Internal PACE Documentation:
  - https://docs.pace.gatech.edu/ice/
- External Resources:
- SentenceTransformers Documentation — Sentence Transformers documentation
- GitHub - TheAtticusProject/cuad: CUAD (NeurIPS 2021)
- GitHub - seatgeek/thefuzz: Fuzzy String Matching in Python

### Complete Working Example
```bash
# Generate ModernBERT results
python generate_modernbert_late_notice_and_claim.py

# Merge all models
python modelmerge.py

# Visualize
python plot_model_scores.py```

Expected workflow and output:
```
ModernBERT F1: 0.69
Pairwise F1: 0.66
CUADBERT F1: 0.64```

NOTE: Important information that users should know goes here.
## Clean node_id duplication before merging and use drop_duplicates

WARNING: Critical warnings about potential issues go here.
## Large merge files (e.g. 2M+ rows) can causes crashes; always test on a smaller sample before exporting full results.

### Version Information
- sentence-transformers: 2.2.2
- transformers: 4.41.0
- pandas: 2.2.3
- Python: 3.11.6
- Last tested on PACE: Spring 2025
- Compatible PACE environments: ICE (VS Code session)