LABS: Treetagger and POS-Tagging for *that*

Objective

This project involves:

1. Using Treetagger to analyze all categories of *that* in test files.
2. Evaluating and reporting precision for each realization.
3. Retraining Treetagger with a customized tagset to distinguish different uses of *that*.

Part 1: Using Treetagger for POS-tagging

1. **Task**:
    - Run Treetagger with test files for all categories of *that*.
    - Example tags:
        - Adverbs: `AVO` (BNC.par, CLAWS5 tagset) or `RB` (Penn Treebank tagset).
    - Test files include various datasets for evaluation.
2. **Evaluation**:
    - Report precision for each test set.
    - **Bonus**: Compare the performance of the Penn and BNC tagsets.

Part 2: Retraining Treetagger

Guidelines

- Use a specific tagset to distinguish different uses of *that*.
- Report precision and recall with the specific datasets.

Example

Refer to this French retraining example.

Custom Tagset: C8 adapted CLAWS8

Tag	Description
WPR	Relative pronoun (*that* in *The man that I saw*).
CST	Subordinating conjunction (*that* in *the fact that I saw*).
CJT	Conjunction for verbs (*I think that you are right*).
DT	Singular determiner (*this, that, another*).
RB	Adverb (*It's not that difficult*).

Deliverables

1. **System Description**:

    - PDF documenting methods, results, and analysis.
2. **Code and Examples**:
    - Include examples of annotated and re-annotated *that* instances.
3. **Team Contributions**:
    - Explain contributions using the CRediT system.

Submission

- Upload a `.zip` file containing:
    - PDF report.
    - Annotated examples as a `.txt` file.
    - Codebase for retraining and evaluation.
- Submission portal:
    - IFHFBU41 Data Science avancée (M1 en alternance).
    - IFABY030 Data Science avancée (M1 en formation initiale).

Timeline

- **8 January 2025**: Lab sessions (9h-12h30, 14h-17h30).
- **15 January 2025**: Additional lab sessions (9h-12h30, 14h-17h30).
- **Deadline**: Submit all deliverables by **9 February 2025, 23:59**.

Project Plan and Structure

Introduction

1. Examples and properties of *that* in English.
2. Problem statement and research objectives.

---

1. Literature Review

- Overview of tagsets (e.g., CLAWS8, Penn Treebank, Universal Dependencies).
- Key distinctions in *that* tagging:
    - Relative pronoun vs. noun complement conjunction.

---

2. Methods and Tools

- **Corpus**: Brown corpus (filtered for *that*).
- **Annotation**:
    - Examples of re-annotated *that*.
    - Use confusion matrices to compare real and predicted tags.
- **Tools**:
    - Treetagger: Custom retraining using the adapted C8 tagset.
    - UDpipe: Dependency parsing for additional analyses.

---

3. Results

1. **Baseline**:
    - Precision with default Treetagger parameters.
2. **Re-annotation**:

- Partial re-annotation: Precision results.
- Full re-annotation (~60% of Brown corpus): Accuracy comparison.

---

4. Discussion

1. **Precision Gains**:
   - Graph: Corpus size (x-axis) vs. Precision (y-axis).
2. **Overfitting**:
   - Compare performance across different corpus categories (e.g., press vs. technical English).
3. **Undervalued Features**:
   - Singular/plural distinction.
   - *That* with/without adjacent nouns.

Conclusion

- Key findings and limitations.
- Suggestions for future work.

Appendices

1. Examples of re-annotated *that* phrases.
2. Model parameters and training configurations.

References

1. CLAWS5 Tagset
2. CLAWS8 Tagset
3. Penn Treebank
4. Manning et al. (2014). CoreNLP: Natural Language Processing Toolkit.
5. Wisniewski et al. (2019). Evaluating Annotation Divergence in Universal Dependencies.

Bonus Tasks

- Investigate *zero complement clauses* (e.g., *the fact I'm a coward*).
- Semantic feature engineering using Skweak.

Good Luck!