Capstone Project: Predicting Career Domain and Seniority from LinkedIn Profiles

Predicting Career Domain and Seniority from LinkedIn Profiles

Project Overview:

In this semester's capstone project, your task is to develop an end-to-end machine-learning pipeline that predicts (1) the current professional domain and (2) the current seniority level of an individual based solely on the information contained in their LinkedIn CV. Your models will be evaluated using a hand-labeled dataset provided by SnapAddy. The project encourages you to creatively combine modern NLP techniques, programmatic labeling strategies, and supervised or zero-shot approaches to extract meaningful signals from semi-structured career data.Details:

- The target is to predict the characteristics (domain, seniority) of the current job. The current job is labeled as "ACTIVE" in the status in the CVs.

Possible Approaches (non-exhaustive)

1. **Rule-based matching (baseline):** Identify relevant job titles and text passages using predefined label lists and assign domain and seniority accordingly.
2. **Embedding-based labeling:** Use the provided label lists to generate embeddings (e.g., via LLMs or sentence transformers). Compute similarity between profile text and label embeddings and perform zero-shot classification.
3. **Fine-tuned classification model.** Use the csv files to fine-tune a pre-trained classification model. Apply the model to the linked-in data
4. **Programmatic labeling + supervised learning: Use rule-based or embedding-based predictions to create pseudo-labels for a large set of LinkedIn profiles, then fine-tune a classifier on this expanded dataset.**
5. **Feature engineering and conventional machine learning. Look at the linked-In data and generate meaningful features (e.g. number of previous jobs as an indicator for seniority, etc.) . Then train conventional algorithms (e.g. random forests) to predict the labels.**
6. **Simple interpretable baseline: E.g. a bag-of-words and TF–IDF + logistic regression classifier for domain or seniority.**
7. **Your own approach: Be creative and find your own solution.**

Note that for each of these approaches, two models are required: one for predicting the department and one for predicting the seniority.

Optional Extensions:

- All of the previous approaches (1-7) can be conducted using only the current position as input. One promising extension is to include the information from previous positions as well. Thereby, you can incorporate assumptions such as:
  - The seniority level typically increases over time
  - It is rather unlikely that persons change the department

    - The name of the organization may provide information about the seniority level and/or departments (for example, if persons work at a NGO the likelihood of having an unpaid job increases)
    - Etc. (identify useful assumptions on your own)
- Deploy the model as a simple prototype application or dashboard for SnapAddy.
- Explainability: Implement some methods to explain the model predictions to the user. E.g. by providing a summary of top cosine similarities, by using LIME / SHAP or other explainability techniques.

Subgoals

- Exploratory Data Analysis of the LinkedIn text data
- Construction of a clean data pipeline
- Implementation and comparison of different learning approaches (see above).
- Evaluation of model performance on the SnapAddy labeled dataset

Evaluation Criteria

- Quality of the presentation
- Documentation quality for both code and final document. There must be one final PDF document. This document may refer to additional Jupyter Notebooks and/or another documentation, such as https://about.readthedocs.com/ or a documented github repository.
- Variety of models or modeling approaches applied. For passing, at least the baseline (approach 1) and one additional approach have to be implemented with sufficient care.
- Originality of the approaches (i.e. how many own thoughts did you integrate)
- Predictive performance (i.e. accuracy on the evaluation set)
- Documentation of potential model failures, potential improvements and next steps
- Functionality of the frontend/dashboard, if implemented
- The use of GenAI must be documented in a dedicated section (corresponding sections in the document and prompts used.)

Deliverables & Deadlines

– Submission of all documentation: by Jan 31, 2026 23:59pm
– Presentations 10 mins per group February 2 (during the lecture time)
– Group size: 3 students

All final reports must include a section explaining the role of each group member.

Notes -> from lecture on 08.12.25

- Endgoal: pipeline die CV als input nimmt & als output predicted automatisiert
-> predicted seniority & which department it is in

- Possible approaches: -> do not have to do all of them (2 required for passing, more required for good grade) (approach 4-> most advanced)
- Requried: Rule-based: > we have to do as simple baseline (no ml involved) -> all other ones have to be better