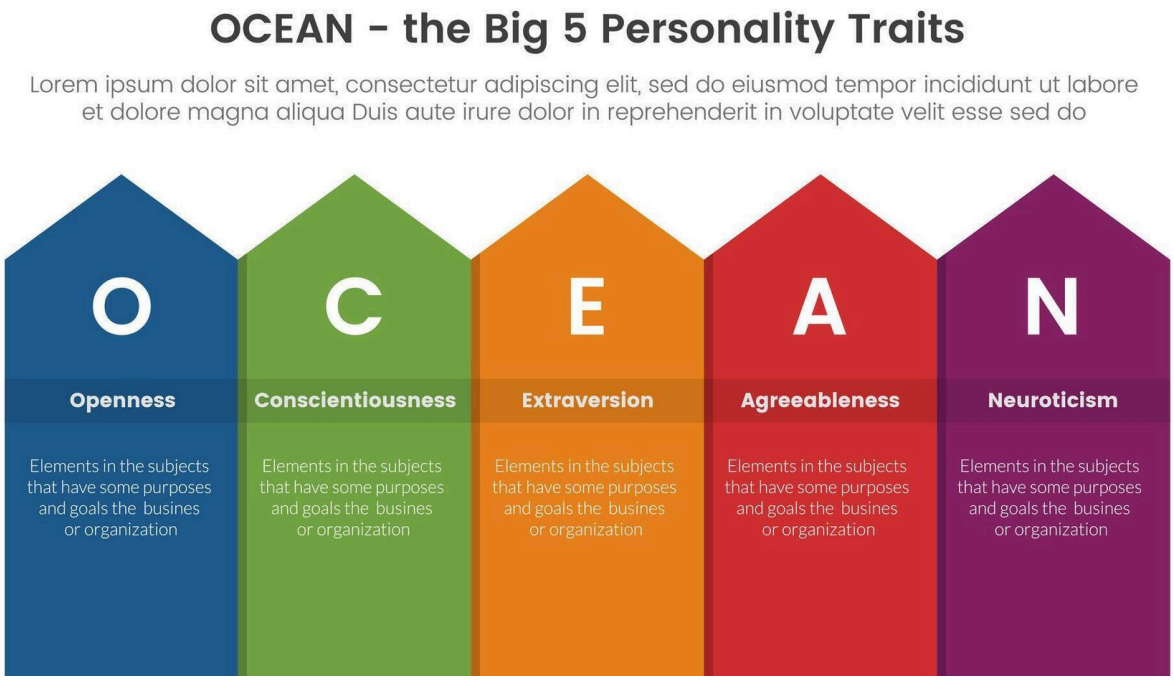
Derick Amadeus Budiono

Project Report: Knowledge Graph Construction from Text with Personality Modeling

Date: October 24, 2025

1. Introduction & Objective

This report details a project focused on the automated construction of Knowledge Graphs (KGs) from unstructured text documents. The primary objective was to develop a Python-based pipeline capable of extracting not only factual entities and relationships but also inferring and representing personality traits of individuals mentioned in the text. The resulting KG aims to provide a structured, queryable, and visual representation of the knowledge contained within the source documents, with a specific emphasis on modeling human personality characteristics according to the Big Five model (OCEAN).



2. Design & Implementation

The system is implemented as a modular Python script utilizing the spaCy library for core Natural Language Processing (NLP) tasks and NetworkX for graph representation, complemented by Pyvis for interactive visualization. The pipeline processes text files (.txt) and executes the following stages:

1. **Text Preprocessing & Entity Recognition:** Input text is processed using spaCy's en_core_web_sm model to perform sentence segmentation, tokenization, part-of-speech tagging, dependency parsing, and Named Entity Recognition (NER).
2. **Coreference Resolution & Entity Linking:** A custom function (create_canonical_name_map) identifies variations of PERSON entities (e.g., "Dr. Vance", "Elias Vance", "Vance") by stripping titles (using clean_name) and comparing name components. It establishes a canonical name (typically the longest variation) for each individual and creates a mapping from all detected variations and basic pronouns ("He", "She") to this canonical name. This map is crucial for unifying nodes in the graph.
3. **Factual Relation Extraction:** A rule-based approach (extract_triples_spacy_enhanced) leverages spaCy's dependency parse tree. It primarily identifies Subject-Predicate-Object (SPO) triples where the predicate is a verb. Subjects (nsubj, nsubjpass) and direct objects/attributes (dobj, attr, acomp) are extracted. Prepositional phrases attached to verbs are also processed to extract contextual relationships (e.g., location using 'IN'). "IS_A" relationships (copula verbs) and possessive relationships ("OWNS") are handled explicitly. All extracted entity strings are mapped to their canonical names using the pre-computed map before forming the final triples.
4. **Personality Trait Modeling:** A rule-based function (assess_personality_from_text) assigns scores (1-5 scale) for the Big Five personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) to the identified canonical person entities. It scans sentences containing person mentions for predefined keywords associated with each trait (defined in DEFAULT_TRAIT_CLUES). Scores are adjusted based on keyword presence, and the keywords serve as textual evidence ('basis'). Inverse traits (like 'Introversion' keywords affecting 'Extraversion') are handled.
5. **Graph Construction:** A NetworkX DiGraph object is constructed using the Labeled Property Graph (LPG) model. Canonical entity names form the nodes. Factual relationships become edges labeled with the predicate type. Personality traits are added as distinct 'Trait' nodes, connected to 'Person' nodes via "HAS_TRAIT" edges. Crucially, the personality score and textual basis are stored as *properties* on these "HAS_TRAIT" edges. Isolated nodes are removed post-construction.
6. **Evaluation:** A dedicated function (evaluate_knowledge_graph) calculates metrics on the generated graph, including structural statistics (node/edge counts, density, degree), counts of different entity types, trait coverage (percentage of persons with traits), and intrinsic quality scores assessing coherence, richness, completeness, entity quality, and relationship diversity. Optional comparison against manually defined ground truth (precision, recall, F1 for entities and relations) is also supported.
7. **Visualization:** The final graph is rendered as an interactive HTML file using the pyvis library (visualize_knowledge_graph_interactive), allowing node dragging, zooming, and hover-over inspection of node/edge properties.

3. Justifications

- **Rule-Based Extraction (spaCy Dependencies & Keywords):** This approach was chosen for its interpretability and control. Dependency parsing provides a structured way to extract core grammatical relationships, while keyword matching for personality allows explicit definition of the behavioral indicators being targeted. This avoids the "black box" nature and potential unpredictability of relying solely on large generative models for extraction in this iteration.
- **Big Five Personality Model:** The OCEAN model is a widely accepted and validated framework in psychology, providing a standardized structure for representing personality dimensions.
- **Labeled Property Graph (LPG) & Edge Properties:** The LPG model was selected because it naturally allows properties (key-value pairs) to be stored directly on edges. This is ideal for representing personality scores, where the score is a characteristic *of the relationship* between a person and a trait, rather than an attribute of the person or trait node itself.
- **Interactive Visualization (pyvis):** While static plots are useful, an interactive visualization significantly aids exploration and debugging of the potentially complex graphs generated, especially for identifying connection patterns or extraction errors.

4. Evaluation

The system's performance was evaluated using a combination of methods:

- **Structural Metrics:** Quantitative measures calculated directly from the graph structure (node/edge count, density, average degree, component count) provide an objective assessment of the graph's size and connectivity.
- **Intrinsic Quality Scores:** Custom composite scores (Graph Coherence, Information Richness, Semantic Completeness, Entity Recognition Quality, Relationship Diversity, Overall Quality Score) were developed to provide a heuristic measure of the graph's quality without requiring manual annotation, assessing aspects like connectivity, extraction rate per sentence, trait assignment completeness, and relationship variety.
- **Manual Inspection:** Visual inspection of the interactive graphs generated from synthetic test documents (1.txt, 2.txt, 3.txt) was crucial for identifying qualitative issues, such as incorrect relationships (e.g., "Bob MEET London") or failed entity unification (e.g., duplicate "Vance" nodes), which guided iterative refinement.
- **Ground Truth Comparison (Optional):** The evaluation framework includes functionality to compare extracted entities and relations against a predefined ground truth dictionary, calculating Precision, Recall, and F1-score for a more formal accuracy assessment when annotations are available.

5. Insights & Limitations

- **Insights:**
    - spaCy's dependency parser provides a solid foundation for extracting simple factual triples.

    - The canonical name mapping significantly improved entity consistency compared to initial versions.
    - Rule-based personality assessment, while basic, successfully links textual evidence to trait scores.
    - The LPG model effectively captures nuanced information like scores on relationships.
    - Interactive visualization proved invaluable for understanding and debugging the extraction process.
- **Limitations:**
    - **Extraction Brittleness:** The rule-based dependency parsing struggles with complex sentence structures, passive voice, and implicit relationships, leading to missed facts or incorrect extractions (like the "Bob MEET London" error).
    - **Coreference/Entity Linking:** The current heuristic approach to name mapping and pronoun resolution is imperfect and can fail with more complex references or less common name variations. A dedicated coreference model would be more robust.
    - **Personality Modeling Simplicity:** Keyword matching is highly dependent on the DEFAULT_TRAIT_CLUES configuration and lacks deeper semantic understanding or context sensitivity. It cannot capture sarcasm, negation subtleties, or infer traits from actions not explicitly listed.
    - **Evaluation:** Intrinsic metrics provide useful heuristics but are not a substitute for thorough evaluation against human-annotated ground truth, which was only partially implemented via the optional dictionary.

6. Conclusion

This project successfully demonstrated the feasibility of constructing a Knowledge Graph from text, incorporating both factual information and rule-based personality assessments, using spaCy and NetworkX. The implementation highlighted the strengths of the LPG model for representing complex attributes on relationships and the utility of interactive visualizations. Key challenges encountered involved the inherent difficulties of robust relation extraction and entity linking using purely rule-based methods.

Future work should focus on integrating more advanced NLP techniques. Incorporating a dedicated coreference resolution model and exploring machine learning-based Relation Extraction models could significantly improve the accuracy and coverage of the factual graph. For personality modeling, leveraging transformer-based models or fine-tuning language models specifically for trait inference from text could yield more nuanced and context