5W1H Extraction: Cause & Method Extractor Documentation

1. Problem Statement

The task involves extracting the cause ("Why?") and method ("How?") information from Romanian news articles. This falls under the umbrella of Natural Language Processing (NLP) and involves identifying the reasons and methods described in the text. The solution will leverage dependency parsing to accurately determine the relationships between words and extract relevant information.

2. Proposed Solution

2.1 Theoretical Aspects

Dependency Parsing:
Dependency parsing involves analysing the grammatical structure of a sentence by establishing relationships between "head" words and words that modify those heads. It provides a clear representation of the syntactic structure of a sentence, making it suitable for extracting "Why?" and "How?" information, which often depends on understanding the relationships between verbs and their arguments or adjuncts.

In this project, we use dependency relations to identify the clauses and phrases that provide explanations (answers to "Why?") and methods (answers to "How?"). For example, in the sentence "Proiectul a fost anulat din cauza tăierilor de buget", the clause "din cauza tăierilor de buget" explains the reason for the delay.

Dependency Relations:

- **Why:** Typically linked with causal conjunctions like "pentru că" (because), "deoarece" (since), etc.
- **How:** Typically linked with phrases indicating manner or method, such as "în felul acesta" (in this way), "prin intermediul" (through), etc.

Algorithm:

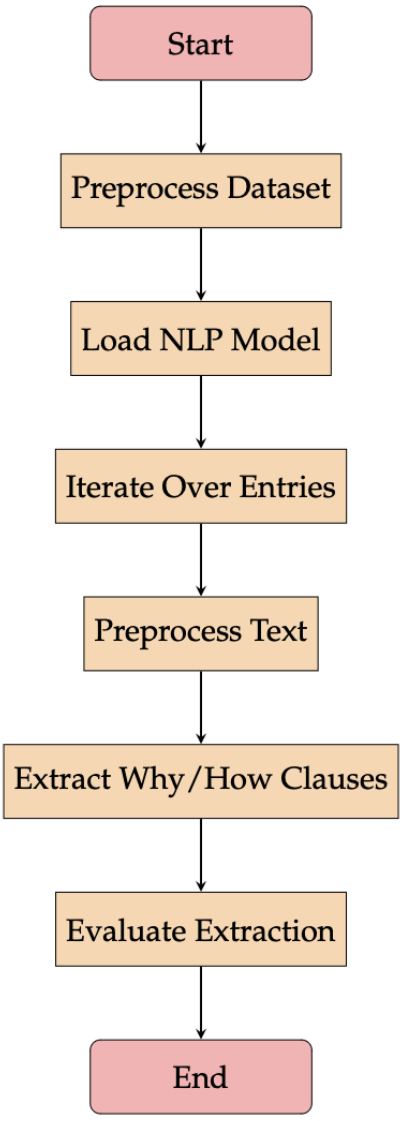
1. **Dependency Parsing**: Use a pre-trained Romanian NLP model (ro_core_news_lg) from SpaCy to perform dependency parsing.
2. **Identification of Clauses**:

- **Why Clauses**: Look for dependency labels such as "mark", "advcl", and "obl". Check for the presence of specific keywords indicative of causality (e.g., "din cauza", "datorită", "pentru că", "deoarece").
- **How Clauses**: Similar dependency labels are analyzed, but the keywords differ (e.g., "prin", "în felul acesta", "astfel", "de această manieră").

2.2 Data Set Used in Application

The dataset used is the Romanian News Articles Dataset, sourced from: https://github.com/mhakan20/RomanianNewsArticlesDataset. The dataset comprises Romanian news articles with annotated "Why" and "How" clauses, suitable for both training and evaluation of the model.

2.3 Application



3. Implementation Details

Libraries/Functions Used:

- **SpaCy**: For NLP tasks including tokenization, part-of-speech tagging, dependency parsing.
- **Scikit-learn**: For evaluation metrics like precision, recall, and F1-score.
- **Logging**: Custom logging setup for tracking progress and errors.

Original Contribution:

- Custom extraction functions tailored for large Romanian news language constructs.
- Utilising dependency parsing to accurately identify and extract relevant clauses.

4. Experiments and Results

Evaluation Metrics

Precision, Recall, and F1-Score are calculated for both "Why" and "How" clauses. These metrics provide insights into the performance of the extraction process, specifically focusing on:

- **Precision**: The proportion of correctly identified clauses among all clauses identified.
- **Recall**: The proportion of correctly identified clauses among all actual clauses.
- **F1-Score**: The harmonic mean of precision and recall, providing a single metric to evaluate the performance.

Extracted Clauses and Performance Metrics:

Average Results for "Why" Clauses:

- **Precision**: (0.33 + 0.00 + 0.50 + 0.50 + 0.00) / 5 = 0.27
- **Recall**: (1.00 + 0.00 + 1.00 + 1.00 + 0.00) / 5 = 0.60
- **F1-Score**: (0.50 + 0.00 + 0.67 + 0.67 + 0.00) / 5 = 0.37

Average Results for "How" Clauses:

- **Precision**: (0.50 + 0.00 + 1.00 + 0.33 + 0.00) / 5 = 0.37
- **Recall**: (1.00 + 0.00 + 1.00 + 1.00 + 0.00) / 5 = 0.60
- **F1-Score**: (0.67 + 0.00 + 1.00 + 0.50 + 0.00) / 5 = 0.43

Summary of Average Results

The average performance metrics for the model across the five entries are as follows:

- **Why Clauses**:
  - **Precision**: 0.27
  - **Recall**: 0.60
  - **F1-Score**: 0.37
- **How Clauses**:
  - **Precision**: 0.37
  - **Recall**: 0.60
  - **F1-Score**: 0.43

These average results indicate that the model performs better in terms of recall than precision for both "Why" and "How" clauses, suggesting that while it identifies a significant portion of relevant clauses, it also includes a number of incorrect ones. Further refinement and tuning of the extraction methods may improve these metrics.