

0.1 Rappresentazione Modelli Stringhe

La prima cosa da fare quando lavoriamo con lambeq è processare la stringa di input e convertirla in uno string diagram.

I modelli che possono essere utilizzati per la modellazione delle stringhe sono:

- **Syntax-based model: DisCoCat**

Per ottenere un risultato simile alla notazione DisCoCat, possiamo usare il parser di default BobcatParser. Fornisce un modello syntax-based ovvero completamente basato sulla sintassi della frase e dei collegamenti tra le componenti della stringa.

- **Bag-of-Words model: spiders reader**

Modello di tipo bag-of-words, vuol dire che non tiene conto delle connessioni e le relazioni tra i termini della stringa ma tratta la stringa stessa come un unico insieme di parole da analizzare separatamente.

- **Word-Sequence model: cups and stairs readers**

Questi modelli tengono conto dell'ordine in cui le parole vengono presentate ma ignorano totalmente la sintassi e le relazioni tra esse.

- **Tree-reader**

Le stringhe vengono divise in monoidi e collegate in una struttura ad albero in base alle relazioni e alla sintassi che mano a mano viene identificata.

Nel contesto da realizzare, ovvero il riconoscimento e la classificazione dei requisiti di security, credo sia necessario utilizzare un approccio che utilizza i modelli che tengono conto della sintassi (anche in parte).

Questo perché se usassimo un modello di tipo bag of words non possiamo identificare realmente frasi sconnesse con parole messe a caso da periodi finiti e ben strutturati.

Inoltre bisogna sfruttare la sintassi per capire meglio il corpo di ciascuna frase, anche perché si potrebbe pensare di estrarre dei concetti e non basarci solamente sulle parole in input per classificare.

(esempio: una frase che parla del napoli potrebbe parlare di calcio o di sport)
(esempio: una frase potrebbe implicitamente dire il suo contesto di classificazione).

Si potrebbe pensare quindi di approcciare per una modellazione basata su DisCoCat oppure, se risulta essere più lightweight dal punto di vista computazionale e se ottiene a grandi linee gli stessi risultati, il modello basato su word-sequence, in modo da non per forza analizzare tutta la frase per estrarne la categoria di appartenenza.