N-gram analysis:

We ignore **Stopwords** (e.g., “এই”, “সে”, “আমি”, “এবং”) etc.

We also need to ignore **Named entities** (e.g., "রাহুল", "ঢাকা"). But how to do so?

**✅ Step 1: Frequency-based Target Word Selection**

* For each time period (e.g. 1950–1970, 1970–1990, etc.):
  + Tokenize and count all word frequencies.
  + Pick the **top 100 frequent content words** per period (after removing stopwords/named entities).
* You now have a pool of **candidate target words** to track.

**✅ Step 2: Find Top Collocates for Each Target Word in Each Time Period**

Yes — for each target word, you now need to:

* Go through each time slice,
* Look at the **neighboring words** around it (usually within a small window),
* Count or **score** those neighbors,
* Pick the **top-k collocates** (e.g. top 10 or 20) based on the score.

We are using the windowed collocate