# Natural Language Query Detection: Methodology & Metrics

## 1. Introduction

This document outlines the approach used to determine whether an input query is a Natural Language Query (NLQ) or not. The system utilizes spaCy, a popular NLP library, to extract linguistic features and compute relevant scores.

## 2. Steps for Classification

1. Preprocessing: The input text is tokenized and processed using the spaCy NLP pipeline.
2. Feature Extraction: Key linguistic features are identified:
3. - POS Tagging (Part-of-Speech Analysis)
4. - Named Entity Recognition (NER)
5. - Stopword Analysis
6. Score Computation: Metrics are calculated based on extracted features.
7. Final Decision: A weighted formula is applied to classify the query as NLQ or Non-NLQ.

## 3. Metrics & Formulae

### 3.1 POS Score (Part-of-Speech Score)

Definition: Measures the proportion of meaningful words (NOUNs, VERBs, ADJs, ADV) in the text.

Formula:  
POS\_Score = (Count of (NOUN + VERB + ADJ + ADV) / Total Token Count) \* 100

### 3.2 NER Score (Named Entity Recognition Score)

Definition: Measures the percentage of words identified as named entities (e.g., products, locations, brands).

Formula:  
NER\_Score = (Number of Named Entities / Total Token Count) \* 100

### 3.3 Stopword Score

Definition: Measures the percentage of stopwords (e.g., 'the', 'is', 'to', 'me').

Formula:  
Stopword\_Score = (Number of Stopwords / Total Token Count) \* 100

### 3.4 Final Classification Score

A weighted sum of the individual metrics is used to determine if the query is a natural language query.

Formula:  
Final\_Score = (0.5 \* POS\_Score) + (0.3 \* NER\_Score) + (0.2 \* Stopword\_Score)

## 4. Decision Rule

If Final Score > 50% → Classified as a Natural Language Query ✅

Otherwise → Classified as Non-NLQ ❌

## 5. Example Calculations

### Example 1: Natural Language Query

Input: "Show me the best smartphone under 10k with 8GB RAM"

POS Score = 62.5%  
NER Score = 40%  
Stopword Score = 35%

Final Score = (0.5 \* 62.5) + (0.3 \* 40) + (0.2 \* 35) = 58.5% ✅ (NLQ)

### Example 2: Non-NLQ (SQL Query)

Input: "SELECT \* FROM users WHERE age > 25"

POS Score = 18.5%  
NER Score = 5%  
Stopword Score = 10%

Final Score = (0.5 \* 18.5) + (0.3 \* 5) + (0.2 \* 10) = 18.5% ❌ (Not NLQ)

## 6. Conclusion

This approach efficiently classifies queries as Natural Language Queries or Non-NLQs based on linguistic features. It is suitable for applications like chatbots, voice assistants, and search engines.

For improvements, Transformer-based models (like BERT) can be explored for enhanced accuracy.