# The Basic Building Blocks of Analysis



Mark Nunnikhoven AWS COMMUNITY HERO @marknca markn.ca

#### Overview

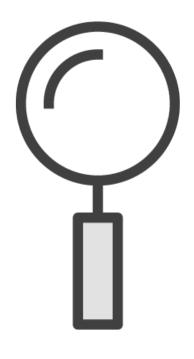
Language detection

Determining the tone of text using sentiment analysis

Locating part-of-speech words within a text

## What Language Is This?

## Language Support



Can identify 100 languages



Can understand 6 languages

### Language Detection Examples

Hello my friend, this is in English. English is one of the 6 supported languages

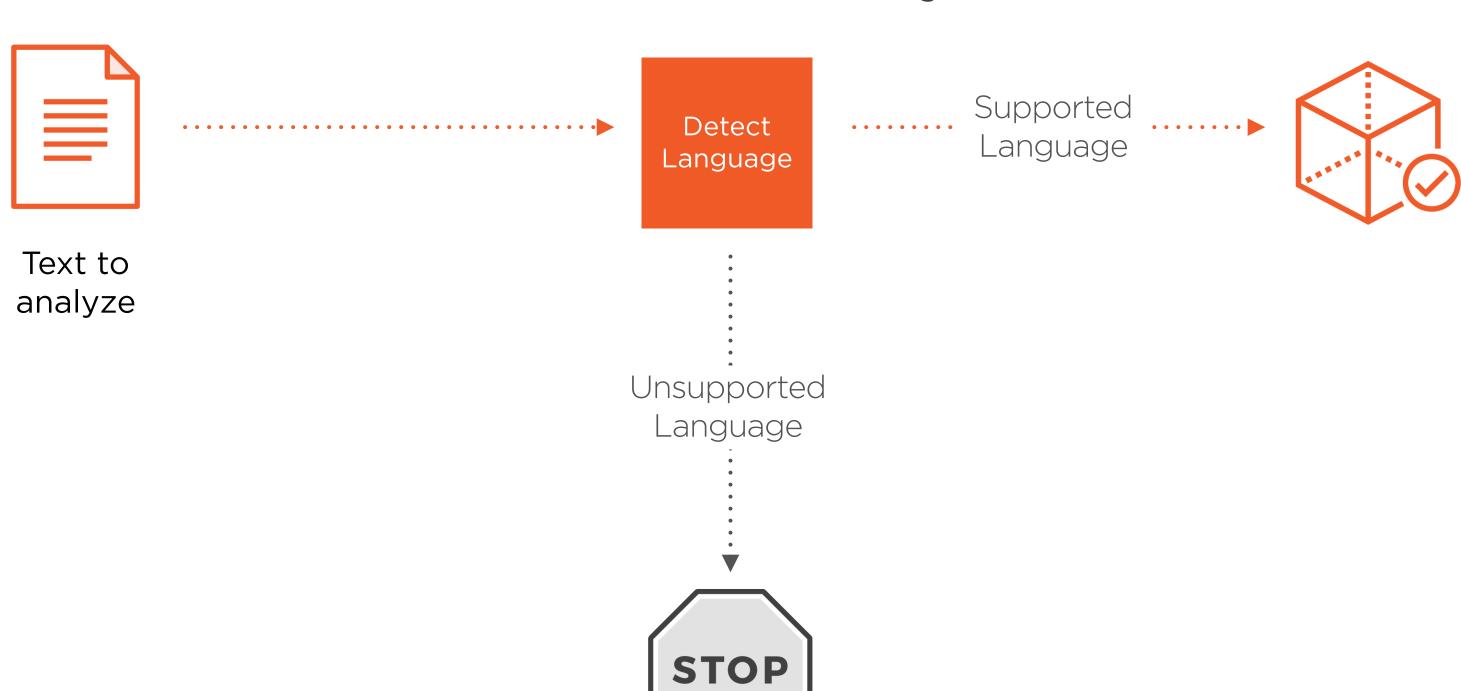
Bonjour mon ami, c'est en français. Le français est l'une des 6 langues supportées Hola mi amigo esto esta en español. Español es uno de los 6 idiomas soportados

*en* 0.997899

*fr* 0.999484

*es* 0.990634

## Document Analysis



# DetectDominantLanguage

A synchronous API call that provides the two digit language code(s) for the specified text

```
{

"Text": "string"

}
```

#### Request Format

The "string" must be at least 1 character but should be 20 or more for an accurate determination.

"String" cannot be more than 5,000 bytes and should be in UTF-8 format.

### Response Format

Under the "Languages" key, the response includes an array of simple a structure.

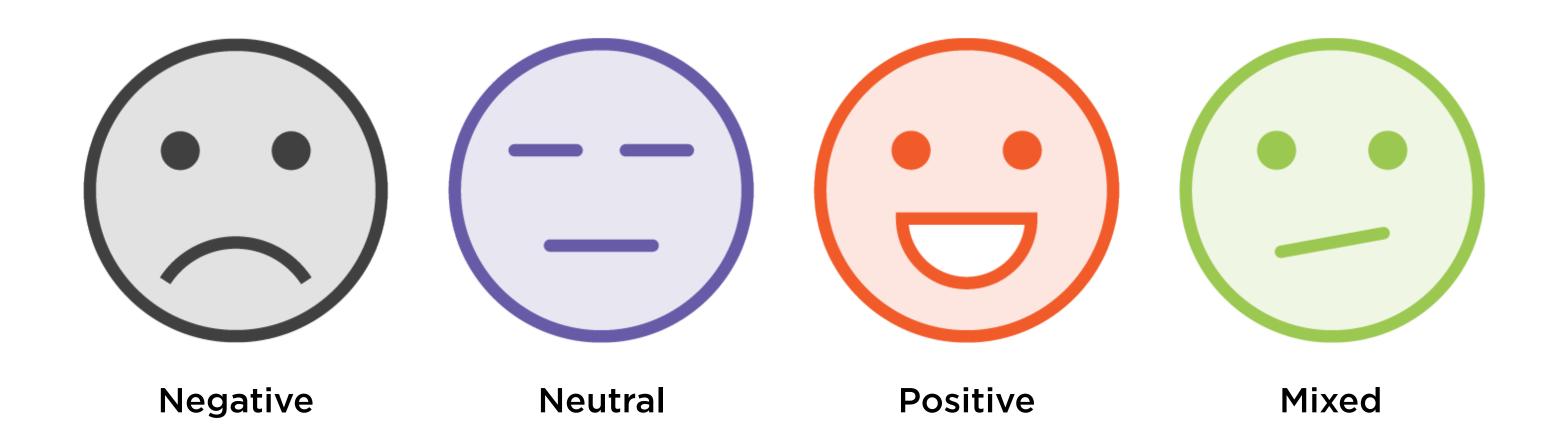
The "LanguageCode" follows RFC 5646 (e.g., "en" or "fr") and the score represents the confidence which with that language has been identified.

#### Demo

Using the CLI to send text to the API Evaluate the results

## Sentiment Analysis

## Document Tone



## Sentiment Examples

This is great

*Positive* 0.977699

This is horrible

*Negative* 0.951358

This is ok

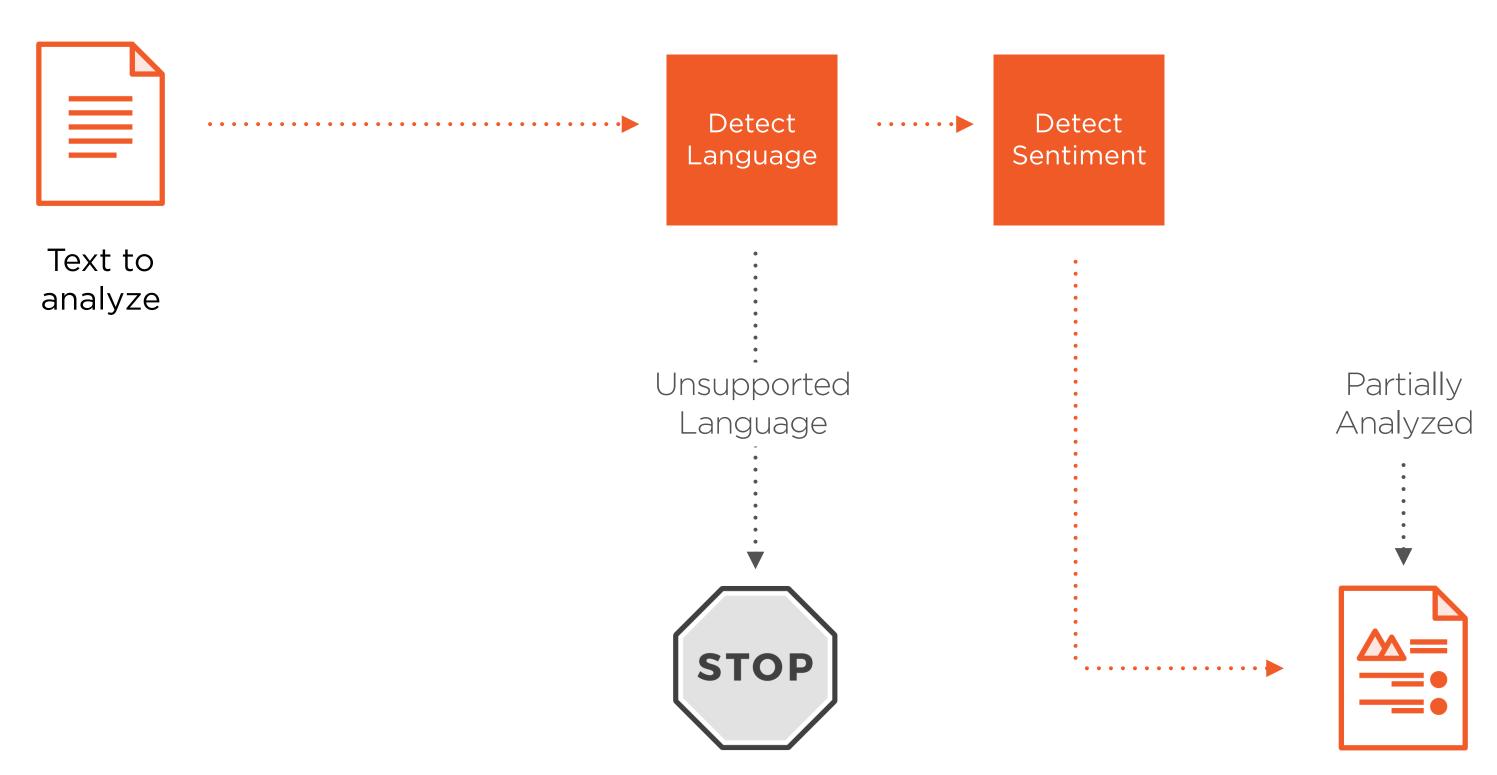
*Neutral* 0.623352







## Document Analysis



## DetectSentiment

A synchronous API call that provides the sentiment for the specified text

## BatchDetectSentiment

A synchronous API call that provides the sentiment for up to 25 samples of text in the same dominant language

#### Request Format

Each "string" must be at least 1 character but should be 20 or more for an accurate determination.

Each "String" cannot be more than 5,000 bytes, should be in UTF-8 format, and there must be 1—25 strings in "TextList"

#### Response Format—Errors

The "ErrorList" contains any errors that occur during the batch job.

The "ErrorCode" is a consistent numerical designation of the unique error, the "ErrorMessage" actually explains the error.

The "Index" corresponds position of the "string" in the Request objects "TextList" array.

```
"ResultList": [
     "Index": number,
     "Sentiment": "string",
     "SentimentScore": {
       "Mixed": number,
       "Negative": number,
       "Neutral": number,
       "Positive": number,
```

- Corresponds to input "TextList"
- Overall sentiment
- ◆ Confidence score for each type of sentiment for the given text. This spread more accurately represents the expressed sentiment in the text.

#### Demo

Use code to prepare the text for analysis as a batch

Analyze the sentiment of each text block

Review the results

## Deconstructing Syntax

# Syntax classification identifies the **part-of-speech** for each word in a text.



Adjective Numeral

Adposition Particle

Adverb Pronoun

Auxiliary Proper noun

**Determiner** Punctuation

Interjection Symbol

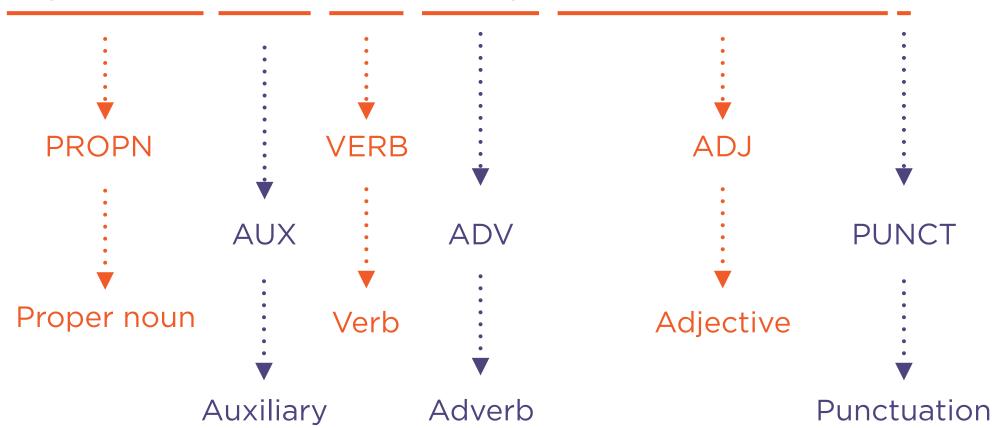
Noun Verb

Subordinating conjunction

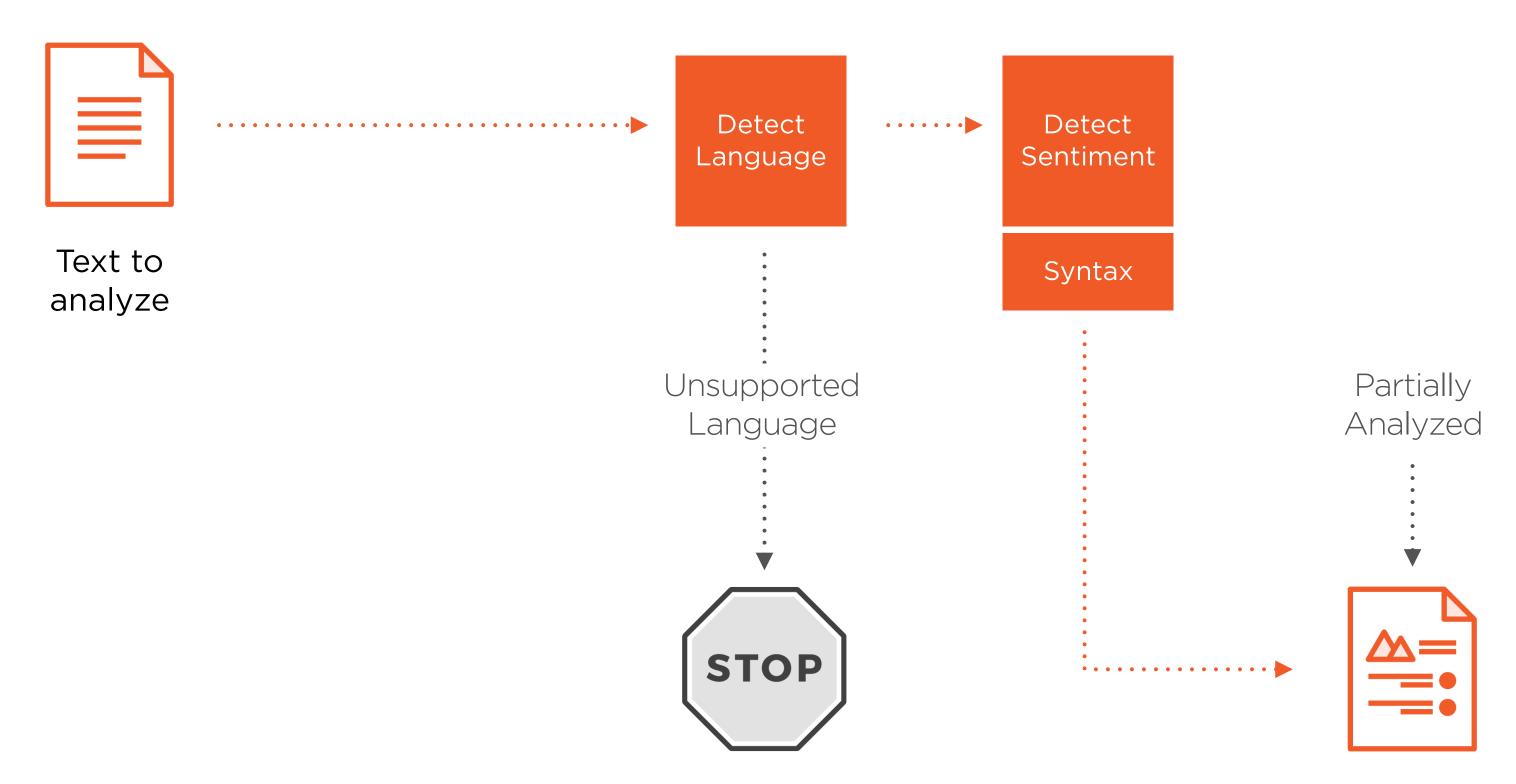
Coordinating conjunction

### Syntax Examples

#### Syntax can be very complicated.



## Document Analysis



# DetectSyntax

A synchronous API call that inspects text looking for various part-of-speech words

# BatchDetectSyntax

A synchronous API call inspects text looking for various part-of-speech words up to 25 samples of text in the same dominant language

#### Request Format

Each "string" must be at least 1 character but should be 20 or more for an accurate determination.

Each "String" cannot be more than 5,000 bytes, should be in UTF-8 format, and there must be 1—25 strings in "TextList"

## Response Format

```
"ResultList": [
     "Index": number,
     "SyntaxTokens": [
         "BeginOffset": number,
         "EndOffset": number,
         "PartOfSpeech": {
            "Score": number,
           "Tag": "string",
         "Text": "string",
      "TokenId": number,
  },
```

Corresponds to the input document

- **■** Starting character in the text
- **■** End character in the text
- **◄** Confidence score for classification
- ◆ Part-of-speech abbreviation (e.g., ADV)

■ Unique ID for this token

#### Demo

Use code to break up a sample text into smaller segments under 5,000 byte limit

Analyze the syntax of each segment

Review the results

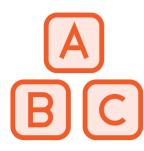
#### Review



Use language detection to discover material in 100 languages and determine which contain the 6 support analysis languages



Sentiment analysis provides insights about the tone of a text



Syntax discovery provides a part-of-speech breakdown of the text, identifying adverbs, adjectives, nouns, etc.



Most classifiers have a direct call and a batch method. The batch method takes up to 25 texts to analyze via one call