indicngram Documentation

Release 0.3

Santhosh Thottingal

September 16, 2013

CONTENTS

An n-gram generator for indic languages

CHAPTER

ONE

WHAT IS NGRAM?

An n-gram model is a type of probabilistic model for predicting the next item in a sequence. n-grams are used in various areas of statistical natural language processing and genetic sequence analysis.

An n-gram is a subsequence of n items from a given sequence. The items in question can be phonemes, syllables, letters, words or base pairs according to the application.

An n-gram of size 1 is referred to as a “unigram”; size 2 is a “bigram” (or, less commonly, a “digram”); size 3 is a “trigram”; and size 4 or more is simply called an “n-gram”.

CHAPTER

TWO

API REFERENCE

class indicngram.core.**Ngram**

Ngram class.You need to create an object to use the function

get_info()

returns info on the module

get_module_name()

returns the module's name

letterNgram(*word*, *window_size=2*)

Parameters

- **word** (*str.*) – The word to be split into ngrams.
- **window_size** (*int.*) – window size to be used while making the ngrams.

Returns list of ngrams.

syllableNgram(*text*, *window_size=2*)

Parameters

- **text** – The text to be split into ngrams.
- **window_size** (*int.*) – window size to be used while making the ngrams.

Returns list of syllable ngrams.

wordNgram(*text*, *window_size=2*)

Parameters

- **text** – The text to be split into ngrams.
- **window_size** (*int.*) – window size to be used while making the ngrams.

Returns list of word ngrams.

CHAPTER

THREE

INDICES AND TABLES

- *genindex*
- *modindex*
- *search*

PYTHON MODULE INDEX