Title: N-gram

URL: https://en.wikipedia.org/wiki/N-gram

PageID: 986182

Categories: Category:Computational linguistics, Category:Corpus linguistics, Category:Language modeling, Category:Natural language processing, Category:Probabilistic models, Category:Speech recognition

-----

An n -gram is a sequence of n adjacent symbols in a particular order. [1] The symbols may be n adjacent letters (including punctuation marks and blanks), syllables , or rarely whole words found in a language dataset; or adjacent phonemes extracted from a speech-recording dataset, or adjacent base pairs extracted from a genome. They are collected from a text corpus or speech corpus .

If Latin numerical prefixes are used, then n -gram of size 1 is called a "unigram", size 2 a " bigram " (or, less commonly, a "digram") etc. If, instead of the Latin ones, the English cardinal numbers are furtherly used, then they are called "four-gram", "five-gram", etc. Similarly, using Greek numerical prefixes such as "monomer", "dimer", "trimer", "tetramer", "pentamer", etc., or English cardinal numbers, "one-mer", "two-mer", "three-mer", etc. are used in computational biology, for polymers or oligomers of a known size, called k -mers . When the items are words, n -grams may also be called shingles . [2]

In the context of natural language processing (NLP), the use of n -grams allows bag-of-words models to capture information such as word order, which would not be possible in the traditional bag of words setting.

Examples

In 1951, Shannon [3] discussed n -gram models of English. For example:

3-gram character model (random draw based on the probabilities of each trigram): in no ist lat whey cratict froure birs grocid pondenome of demonstures of the retagin is regiactiona of cre

2-gram word model (random draw of words taking into account their transition probabilities): the head and in frontal attack on an english writer that the character of this point is therefore another method for the letters that the time of who ever told the problem for an unexpected

Figure 1 shows several example sequences and the corresponding 1-gram, 2-gram and 3-gram sequences.

Here are further examples; these are word-level 3-grams and 4-grams (and counts of the number of times they appeared) from the Google n -gram corpus. [4]

3-grams

ceramics collectables collectibles (55)

ceramics collectables fine (130)

ceramics collected by (52)

ceramics collectible pottery (50)

ceramics collectibles cooking (45)

4-grams

serve as the incoming (92)

serve as the incubator (99)

serve as the independent (794)

serve as the index (223)

serve as the indication (72)

serve as the indicator (120)

References

Further reading

Manning, Christopher D.; Schütze, Hinrich; Foundations of Statistical Natural Language Processing , MIT Press: 1999, ISBN 0-262-13360-1

White, Owen; Dunning, Ted; Sutton, Granger; Adams, Mark; Venter, J. Craig; Fields, Chris (1993). "A quality control algorithm for dna sequencing projects" . Nucleic Acids Research . 21 (16): 3829–3838. doi : 10.1093/nar/21.16.3829 . PMC 309901 . PMID 8367301 .

Damerau, Frederick J.; Markov Models and Linguistic Theory , Mouton, The Hague, 1971

Figueroa, Alejandro; Atkinson, John (2012). "Contextual Language Models For Ranking Answers To Natural Language Definition Questions". Computational Intelligence . 28 (4): 528– 548. doi : 10.1111/j.1467-8640.2012.00426.x .

Brocardo, Marcelo Luiz; Traore, Issa; Saad, Sherif; Woungang, Isaac (2013). "Authorship verification for short messages using stylometry". 2013 International Conference on Computer, Information and Telecommunication Systems (CITS) . pp. 1– 6. doi : 10.1109/CITS.2013.6705711 . ISBN 978-1-4799-0168-5 .

See also

Google Books Ngram Viewer

External links

Ngram Extractor: Gives weight of n -gram based on their frequency.

Google's Google Books n -gram viewer and Web n -grams database (September 2006)

STATOPERATOR N-grams Project Weighted n -gram viewer for every domain in Alexa Top 1M

1,000,000 most frequent 2,3,4,5-grams from the 425 million word Corpus of Contemporary American English

Peachnote's music ngram viewer

Stochastic Language Models (n -Gram) Specification (W3C)

Michael Collins's notes on n -Gram Language Models

OpenRefine: Clustering In Depth

v

t

e

AI-complete

Bag-of-words

n -gram Bigram Trigram

Bigram

Trigram

Computational linguistics

Natural language understanding

Stop words

Text processing

Argument mining

Collocation extraction

Concept mining

Coreference resolution

Deep linguistic processing

Distant reading

Information extraction

Named-entity recognition

Ontology learning

Parsing Semantic parsing Syntactic parsing

Semantic parsing

Syntactic parsing

Part-of-speech tagging

Semantic analysis

Semantic role labeling

Semantic decomposition

Semantic similarity

Sentiment analysis

Terminology extraction

Text mining

Textual entailment

Truecasing

Word-sense disambiguation

Word-sense induction

Compound-term processing

Lemmatisation

Lexical analysis

Text chunking

Stemming

Sentence segmentation

Word segmentation

Multi-document summarization

Sentence extraction

Text simplification

Computer-assisted

Example-based

Rule-based

Statistical

Transfer-based

Neural

BERT

Document-term matrix

Explicit semantic analysis

fastText

GloVe

Language model (large)

Latent semantic analysis

Seq2seq

Word embedding

Word2vec

Corpus linguistics

Lexical resource

Linguistic Linked Open Data

Machine-readable dictionary

Parallel text

PropBank

Semantic network

Simple Knowledge Organization System

Speech corpus

Text corpus

Thesaurus (information retrieval)

Treebank

Universal Dependencies

BabelNet

Bank of English

DBpedia

FrameNet

Google Ngram Viewer

UBY

WordNet

Wikidata

Speech recognition

Speech segmentation

Speech synthesis

Natural language generation

Optical character recognition

Document classification

Latent Dirichlet allocation

Pachinko allocation

Automated essay scoring

Concordancer

Grammar checker

Predictive text

Pronunciation assessment

Spell checker

Chatbot

Interactive fiction

Question answering

Virtual assistant

Voice user interface

Formal semantics

Hallucination

Natural Language Toolkit

spaCy