{

"nbformat": 4,

"nbformat\_minor": 0,

"metadata": {

"colab": {

"name": "Aula 4 - NLTK -Python para PLN.ipynb",

"provenance": [],

"collapsed\_sections": []

},

"kernelspec": {

"name": "python3",

"display\_name": "Python 3"

},

"language\_info": {

"name": "python"

}

},

"cells": [

{

"cell\_type": "code",

"metadata": {

"id": "PAhFdxNexiye"

},

"source": [

"import nltk"

],

"execution\_count": null,

"outputs": []

},

{

"cell\_type": "code",

"metadata": {

"colab": {

"base\_uri": "https://localhost:8080/"

},

"id": "rjYDsdnhyIK7",

"outputId": "0100b7c8-14dc-4402-c51c-fac762677f5c"

},

"source": [

"nltk.download()"

],

"execution\_count": null,

"outputs": [

{

"name": "stdout",

"output\_type": "stream",

"text": [

"NLTK Downloader\n",

"---------------------------------------------------------------------------\n",

" d) Download l) List u) Update c) Config h) Help q) Quit\n",

"---------------------------------------------------------------------------\n",

"Downloader> l\n",

"\n",

"Packages:\n",

" [ ] abc................. Australian Broadcasting Commission 2006\n",

" [ ] alpino.............. Alpino Dutch Treebank\n",

" [ ] averaged\_perceptron\_tagger Averaged Perceptron Tagger\n",

" [ ] averaged\_perceptron\_tagger\_ru Averaged Perceptron Tagger (Russian)\n",

" [ ] basque\_grammars..... Grammars for Basque\n",

" [ ] biocreative\_ppi..... BioCreAtIvE (Critical Assessment of Information\n",

" Extraction Systems in Biology)\n",

" [ ] bllip\_wsj\_no\_aux.... BLLIP Parser: WSJ Model\n",

" [ ] book\_grammars....... Grammars from NLTK Book\n",

" [ ] brown............... Brown Corpus\n",

" [ ] brown\_tei........... Brown Corpus (TEI XML Version)\n",

" [ ] cess\_cat............ CESS-CAT Treebank\n",

" [ ] cess\_esp............ CESS-ESP Treebank\n",

" [ ] chat80.............. Chat-80 Data Files\n",

" [ ] city\_database....... City Database\n",

" [ ] cmudict............. The Carnegie Mellon Pronouncing Dictionary (0.6)\n",

" [ ] comparative\_sentences Comparative Sentence Dataset\n",

" [ ] comtrans............ ComTrans Corpus Sample\n",

" [ ] conll2000........... CONLL 2000 Chunking Corpus\n",

" [ ] conll2002........... CONLL 2002 Named Entity Recognition Corpus\n",

"Hit Enter to continue: \n",

" [ ] conll2007........... Dependency Treebanks from CoNLL 2007 (Catalan\n",

" and Basque Subset)\n",

" [ ] crubadan............ Crubadan Corpus\n",

" [ ] dependency\_treebank. Dependency Parsed Treebank\n",

" [ ] dolch............... Dolch Word List\n",

" [ ] europarl\_raw........ Sample European Parliament Proceedings Parallel\n",

" Corpus\n",

" [ ] floresta............ Portuguese Treebank\n",

" [ ] framenet\_v15........ FrameNet 1.5\n",

" [ ] framenet\_v17........ FrameNet 1.7\n",

" [ ] gazetteers.......... Gazeteer Lists\n",

" [ ] genesis............. Genesis Corpus\n",

" [ ] gutenberg........... Project Gutenberg Selections\n",

" [ ] ieer................ NIST IE-ER DATA SAMPLE\n",

" [ ] inaugural........... C-Span Inaugural Address Corpus\n",

" [ ] indian.............. Indian Language POS-Tagged Corpus\n",

" [ ] jeita............... JEITA Public Morphologically Tagged Corpus (in\n",

" ChaSen format)\n",

" [ ] kimmo............... PC-KIMMO Data Files\n",

" [ ] knbc................ KNB Corpus (Annotated blog corpus)\n",

" [ ] large\_grammars...... Large context-free and feature-based grammars\n",

" for parser comparison\n",

"Hit Enter to continue: \n",

" [ ] lin\_thesaurus....... Lin's Dependency Thesaurus\n",

" [ ] mac\_morpho.......... MAC-MORPHO: Brazilian Portuguese news text with\n",

" part-of-speech tags\n",

" [ ] machado............. Machado de Assis -- Obra Completa\n",

" [ ] masc\_tagged......... MASC Tagged Corpus\n",

" [ ] maxent\_ne\_chunker... ACE Named Entity Chunker (Maximum entropy)\n",

" [ ] maxent\_treebank\_pos\_tagger Treebank Part of Speech Tagger (Maximum entropy)\n",

" [ ] moses\_sample........ Moses Sample Models\n",

" [ ] movie\_reviews....... Sentiment Polarity Dataset Version 2.0\n",

" [ ] mte\_teip5........... MULTEXT-East 1984 annotated corpus 4.0\n",

" [ ] mwa\_ppdb............ The monolingual word aligner (Sultan et al.\n",

" 2015) subset of the Paraphrase Database.\n",

" [ ] names............... Names Corpus, Version 1.3 (1994-03-29)\n",

" [ ] nombank.1.0......... NomBank Corpus 1.0\n",

" [ ] nonbreaking\_prefixes Non-Breaking Prefixes (Moses Decoder)\n",

" [ ] nps\_chat............ NPS Chat\n",

" [ ] omw................. Open Multilingual Wordnet\n",

" [ ] opinion\_lexicon..... Opinion Lexicon\n",

" [ ] panlex\_swadesh...... PanLex Swadesh Corpora\n",

" [ ] paradigms........... Paradigm Corpus\n",

" [ ] pe08................ Cross-Framework and Cross-Domain Parser\n",

" Evaluation Shared Task\n",

"Hit Enter to continue: \n",

" [ ] perluniprops........ perluniprops: Index of Unicode Version 7.0.0\n",

" character properties in Perl\n",

" [ ] pil................. The Patient Information Leaflet (PIL) Corpus\n",

" [ ] pl196x.............. Polish language of the XX century sixties\n",

" [ ] porter\_test......... Porter Stemmer Test Files\n",

" [ ] ppattach............ Prepositional Phrase Attachment Corpus\n",

" [ ] problem\_reports..... Problem Report Corpus\n",

" [ ] product\_reviews\_1... Product Reviews (5 Products)\n",

" [ ] product\_reviews\_2... Product Reviews (9 Products)\n",

" [ ] propbank............ Proposition Bank Corpus 1.0\n",

" [ ] pros\_cons........... Pros and Cons\n",

" [ ] ptb................. Penn Treebank\n",

" [ ] punkt............... Punkt Tokenizer Models\n",

" [ ] qc.................. Experimental Data for Question Classification\n",

" [ ] reuters............. The Reuters-21578 benchmark corpus, ApteMod\n",

" version\n",

" [ ] rslp................ RSLP Stemmer (Removedor de Sufixos da Lingua\n",

" Portuguesa)\n",

" [ ] rte................. PASCAL RTE Challenges 1, 2, and 3\n",

" [ ] sample\_grammars..... Sample Grammars\n",

" [ ] semcor.............. SemCor 3.0\n",

"Hit Enter to continue: \n",

" [ ] senseval............ SENSEVAL 2 Corpus: Sense Tagged Text\n",

" [ ] sentence\_polarity... Sentence Polarity Dataset v1.0\n",

" [ ] sentiwordnet........ SentiWordNet\n",

" [ ] shakespeare......... Shakespeare XML Corpus Sample\n",

" [ ] sinica\_treebank..... Sinica Treebank Corpus Sample\n",

" [ ] smultron............ SMULTRON Corpus Sample\n",

" [ ] snowball\_data....... Snowball Data\n",

" [ ] spanish\_grammars.... Grammars for Spanish\n",

" [ ] state\_union......... C-Span State of the Union Address Corpus\n",

" [ ] stopwords........... Stopwords Corpus\n",

" [ ] subjectivity........ Subjectivity Dataset v1.0\n",

" [ ] swadesh............. Swadesh Wordlists\n",

" [ ] switchboard......... Switchboard Corpus Sample\n",

" [ ] tagsets............. Help on Tagsets\n",

" [ ] timit............... TIMIT Corpus Sample\n",

" [ ] toolbox............. Toolbox Sample Files\n",

" [ ] treebank............ Penn Treebank Sample\n",

" [ ] twitter\_samples..... Twitter Samples\n",

" [ ] udhr2............... Universal Declaration of Human Rights Corpus\n",

" (Unicode Version)\n",

" [ ] udhr................ Universal Declaration of Human Rights Corpus\n",

"Hit Enter to continue: \n",

" [ ] unicode\_samples..... Unicode Samples\n",

" [ ] universal\_tagset.... Mappings to the Universal Part-of-Speech Tagset\n",

" [ ] universal\_treebanks\_v20 Universal Treebanks Version 2.0\n",

" [ ] vader\_lexicon....... VADER Sentiment Lexicon\n",

" [ ] verbnet3............ VerbNet Lexicon, Version 3.3\n",

" [ ] verbnet............. VerbNet Lexicon, Version 2.1\n",

" [ ] webtext............. Web Text Corpus\n",

" [ ] wmt15\_eval.......... Evaluation data from WMT15\n",

" [ ] word2vec\_sample..... Word2Vec Sample\n",

" [ ] wordnet............. WordNet\n",

" [ ] wordnet\_ic.......... WordNet-InfoContent\n",

" [ ] words............... Word Lists\n",

" [ ] ycoe................ York-Toronto-Helsinki Parsed Corpus of Old\n",

" English Prose\n",

"\n",

"Collections:\n",

" [ ] all-corpora......... All the corpora\n",

" [ ] all-nltk............ All packages available on nltk\_data gh-pages\n",

" branch\n",

" [ ] all................. All packages\n",

" [ ] book................ Everything used in the NLTK Book\n",

" [ ] popular............. Popular packages\n",

"Hit Enter to continue: \n",

" [ ] tests............... Packages for running tests\n",

" [ ] third-party......... Third-party data packages\n",

"\n",

"([\*] marks installed packages)\n",

"\n",

"---------------------------------------------------------------------------\n",

" d) Download l) List u) Update c) Config h) Help q) Quit\n",

"---------------------------------------------------------------------------\n",

"Downloader> d\n",

"\n",

"Download which package (l=list; x=cancel)?\n",

" Identifier> all\n",

" Downloading collection 'all'\n",

" | \n",

" | Downloading package abc to /root/nltk\_data...\n",

" | Unzipping corpora/abc.zip.\n",

" | Downloading package alpino to /root/nltk\_data...\n",

" | Unzipping corpora/alpino.zip.\n",

" | Downloading package biocreative\_ppi to /root/nltk\_data...\n",

" | Unzipping corpora/biocreative\_ppi.zip.\n",

" | Downloading package brown to /root/nltk\_data...\n",

" | Unzipping corpora/brown.zip.\n",

" | Downloading package brown\_tei to /root/nltk\_data...\n",

" | Unzipping corpora/brown\_tei.zip.\n",

" | Downloading package cess\_cat to /root/nltk\_data...\n",

" | Unzipping corpora/cess\_cat.zip.\n",

" | Downloading package cess\_esp to /root/nltk\_data...\n",

" | Unzipping corpora/cess\_esp.zip.\n",

" | Downloading package chat80 to /root/nltk\_data...\n",

" | Unzipping corpora/chat80.zip.\n",

" | Downloading package city\_database to /root/nltk\_data...\n",

" | Unzipping corpora/city\_database.zip.\n",

" | Downloading package cmudict to /root/nltk\_data...\n",

" | Unzipping corpora/cmudict.zip.\n",

" | Downloading package comparative\_sentences to\n",

" | /root/nltk\_data...\n",

" | Unzipping corpora/comparative\_sentences.zip.\n",

" | Downloading package comtrans to /root/nltk\_data...\n",

" | Downloading package conll2000 to /root/nltk\_data...\n",

" | Unzipping corpora/conll2000.zip.\n",

" | Downloading package conll2002 to /root/nltk\_data...\n",

" | Unzipping corpora/conll2002.zip.\n",

" | Downloading package conll2007 to /root/nltk\_data...\n",

" | Downloading package crubadan to /root/nltk\_data...\n",

" | Unzipping corpora/crubadan.zip.\n",

" | Downloading package dependency\_treebank to /root/nltk\_data...\n",

" | Unzipping corpora/dependency\_treebank.zip.\n",

" | Downloading package dolch to /root/nltk\_data...\n",

" | Unzipping corpora/dolch.zip.\n",

" | Downloading package europarl\_raw to /root/nltk\_data...\n",

" | Unzipping corpora/europarl\_raw.zip.\n",

" | Downloading package floresta to /root/nltk\_data...\n",

" | Unzipping corpora/floresta.zip.\n",

" | Downloading package framenet\_v15 to /root/nltk\_data...\n",

" | Unzipping corpora/framenet\_v15.zip.\n",

" | Downloading package framenet\_v17 to /root/nltk\_data...\n",

" | Unzipping corpora/framenet\_v17.zip.\n",

" | Downloading package gazetteers to /root/nltk\_data...\n",

" | Unzipping corpora/gazetteers.zip.\n",

" | Downloading package genesis to /root/nltk\_data...\n",

" | Unzipping corpora/genesis.zip.\n",

" | Downloading package gutenberg to /root/nltk\_data...\n",

" | Unzipping corpora/gutenberg.zip.\n",

" | Downloading package ieer to /root/nltk\_data...\n",

" | Unzipping corpora/ieer.zip.\n",

" | Downloading package inaugural to /root/nltk\_data...\n",

" | Unzipping corpora/inaugural.zip.\n",

" | Downloading package indian to /root/nltk\_data...\n",

" | Unzipping corpora/indian.zip.\n",

" | Downloading package jeita to /root/nltk\_data...\n",

" | Downloading package kimmo to /root/nltk\_data...\n",

" | Unzipping corpora/kimmo.zip.\n",

" | Downloading package knbc to /root/nltk\_data...\n",

" | Downloading package lin\_thesaurus to /root/nltk\_data...\n",

" | Unzipping corpora/lin\_thesaurus.zip.\n",

" | Downloading package mac\_morpho to /root/nltk\_data...\n",

" | Unzipping corpora/mac\_morpho.zip.\n",

" | Downloading package machado to /root/nltk\_data...\n",

" | Downloading package masc\_tagged to /root/nltk\_data...\n",

" | Downloading package moses\_sample to /root/nltk\_data...\n",

" | Unzipping models/moses\_sample.zip.\n",

" | Downloading package movie\_reviews to /root/nltk\_data...\n",

" | Unzipping corpora/movie\_reviews.zip.\n",

" | Downloading package names to /root/nltk\_data...\n",

" | Unzipping corpora/names.zip.\n",

" | Downloading package nombank.1.0 to /root/nltk\_data...\n",

" | Downloading package nps\_chat to /root/nltk\_data...\n",

" | Unzipping corpora/nps\_chat.zip.\n",

" | Downloading package omw to /root/nltk\_data...\n",

" | Unzipping corpora/omw.zip.\n",

" | Downloading package opinion\_lexicon to /root/nltk\_data...\n",

" | Unzipping corpora/opinion\_lexicon.zip.\n",

" | Downloading package paradigms to /root/nltk\_data...\n",

" | Unzipping corpora/paradigms.zip.\n",

" | Downloading package pil to /root/nltk\_data...\n",

" | Unzipping corpora/pil.zip.\n",

" | Downloading package pl196x to /root/nltk\_data...\n",

" | Unzipping corpora/pl196x.zip.\n",

" | Downloading package ppattach to /root/nltk\_data...\n",

" | Unzipping corpora/ppattach.zip.\n",

" | Downloading package problem\_reports to /root/nltk\_data...\n",

" | Unzipping corpora/problem\_reports.zip.\n",

" | Downloading package propbank to /root/nltk\_data...\n",

" | Downloading package ptb to /root/nltk\_data...\n",

" | Unzipping corpora/ptb.zip.\n",

" | Downloading package product\_reviews\_1 to /root/nltk\_data...\n",

" | Unzipping corpora/product\_reviews\_1.zip.\n",

" | Downloading package product\_reviews\_2 to /root/nltk\_data...\n",

" | Unzipping corpora/product\_reviews\_2.zip.\n",

" | Downloading package pros\_cons to /root/nltk\_data...\n",

" | Unzipping corpora/pros\_cons.zip.\n",

" | Downloading package qc to /root/nltk\_data...\n",

" | Unzipping corpora/qc.zip.\n",

" | Downloading package reuters to /root/nltk\_data...\n",

" | Downloading package rte to /root/nltk\_data...\n",

" | Unzipping corpora/rte.zip.\n",

" | Downloading package semcor to /root/nltk\_data...\n",

" | Downloading package senseval to /root/nltk\_data...\n",

" | Unzipping corpora/senseval.zip.\n",

" | Downloading package sentiwordnet to /root/nltk\_data...\n",

" | Unzipping corpora/sentiwordnet.zip.\n",

" | Downloading package sentence\_polarity to /root/nltk\_data...\n",

" | Unzipping corpora/sentence\_polarity.zip.\n",

" | Downloading package shakespeare to /root/nltk\_data...\n",

" | Unzipping corpora/shakespeare.zip.\n",

" | Downloading package sinica\_treebank to /root/nltk\_data...\n",

" | Unzipping corpora/sinica\_treebank.zip.\n",

" | Downloading package smultron to /root/nltk\_data...\n",

" | Unzipping corpora/smultron.zip.\n",

" | Downloading package state\_union to /root/nltk\_data...\n",

" | Unzipping corpora/state\_union.zip.\n",

" | Downloading package stopwords to /root/nltk\_data...\n",

" | Unzipping corpora/stopwords.zip.\n",

" | Downloading package subjectivity to /root/nltk\_data...\n",

" | Unzipping corpora/subjectivity.zip.\n",

" | Downloading package swadesh to /root/nltk\_data...\n",

" | Unzipping corpora/swadesh.zip.\n",

" | Downloading package switchboard to /root/nltk\_data...\n",

" | Unzipping corpora/switchboard.zip.\n",

" | Downloading package timit to /root/nltk\_data...\n",

" | Unzipping corpora/timit.zip.\n",

" | Downloading package toolbox to /root/nltk\_data...\n",

" | Unzipping corpora/toolbox.zip.\n",

" | Downloading package treebank to /root/nltk\_data...\n",

" | Unzipping corpora/treebank.zip.\n",

" | Downloading package twitter\_samples to /root/nltk\_data...\n",

" | Unzipping corpora/twitter\_samples.zip.\n",

" | Downloading package udhr to /root/nltk\_data...\n",

" | Unzipping corpora/udhr.zip.\n",

" | Downloading package udhr2 to /root/nltk\_data...\n",

" | Unzipping corpora/udhr2.zip.\n",

" | Downloading package unicode\_samples to /root/nltk\_data...\n",

" | Unzipping corpora/unicode\_samples.zip.\n",

" | Downloading package universal\_treebanks\_v20 to\n",

" | /root/nltk\_data...\n",

" | Downloading package verbnet to /root/nltk\_data...\n",

" | Unzipping corpora/verbnet.zip.\n",

" | Downloading package verbnet3 to /root/nltk\_data...\n",

" | Unzipping corpora/verbnet3.zip.\n",

" | Downloading package webtext to /root/nltk\_data...\n",

" | Unzipping corpora/webtext.zip.\n",

" | Downloading package wordnet to /root/nltk\_data...\n",

" | Unzipping corpora/wordnet.zip.\n",

" | Downloading package wordnet\_ic to /root/nltk\_data...\n",

" | Unzipping corpora/wordnet\_ic.zip.\n",

" | Downloading package words to /root/nltk\_data...\n",

" | Unzipping corpora/words.zip.\n",

" | Downloading package ycoe to /root/nltk\_data...\n",

" | Unzipping corpora/ycoe.zip.\n",

" | Downloading package rslp to /root/nltk\_data...\n",

" | Unzipping stemmers/rslp.zip.\n",

" | Downloading package maxent\_treebank\_pos\_tagger to\n",

" | /root/nltk\_data...\n",

" | Unzipping taggers/maxent\_treebank\_pos\_tagger.zip.\n",

" | Downloading package universal\_tagset to /root/nltk\_data...\n",

" | Unzipping taggers/universal\_tagset.zip.\n",

" | Downloading package maxent\_ne\_chunker to /root/nltk\_data...\n",

" | Unzipping chunkers/maxent\_ne\_chunker.zip.\n",

" | Downloading package punkt to /root/nltk\_data...\n",

" | Unzipping tokenizers/punkt.zip.\n",

" | Downloading package book\_grammars to /root/nltk\_data...\n",

" | Unzipping grammars/book\_grammars.zip.\n",

" | Downloading package sample\_grammars to /root/nltk\_data...\n",

" | Unzipping grammars/sample\_grammars.zip.\n",

" | Downloading package spanish\_grammars to /root/nltk\_data...\n",

" | Unzipping grammars/spanish\_grammars.zip.\n",

" | Downloading package basque\_grammars to /root/nltk\_data...\n",

" | Unzipping grammars/basque\_grammars.zip.\n",

" | Downloading package large\_grammars to /root/nltk\_data...\n",

" | Unzipping grammars/large\_grammars.zip.\n",

" | Downloading package tagsets to /root/nltk\_data...\n",

" | Unzipping help/tagsets.zip.\n",

" | Downloading package snowball\_data to /root/nltk\_data...\n",

" | Downloading package bllip\_wsj\_no\_aux to /root/nltk\_data...\n",

" | Unzipping models/bllip\_wsj\_no\_aux.zip.\n",

" | Downloading package word2vec\_sample to /root/nltk\_data...\n",

" | Unzipping models/word2vec\_sample.zip.\n",

" | Downloading package panlex\_swadesh to /root/nltk\_data...\n",

" | Downloading package mte\_teip5 to /root/nltk\_data...\n",

" | Unzipping corpora/mte\_teip5.zip.\n",

" | Downloading package averaged\_perceptron\_tagger to\n",

" | /root/nltk\_data...\n",

" | Unzipping taggers/averaged\_perceptron\_tagger.zip.\n",

" | Downloading package averaged\_perceptron\_tagger\_ru to\n",

" | /root/nltk\_data...\n",

" | Unzipping taggers/averaged\_perceptron\_tagger\_ru.zip.\n",

" | Downloading package perluniprops to /root/nltk\_data...\n",

" | Unzipping misc/perluniprops.zip.\n",

" | Downloading package nonbreaking\_prefixes to\n",

" | /root/nltk\_data...\n",

" | Unzipping corpora/nonbreaking\_prefixes.zip.\n",

" | Downloading package vader\_lexicon to /root/nltk\_data...\n",

" | Downloading package porter\_test to /root/nltk\_data...\n",

" | Unzipping stemmers/porter\_test.zip.\n",

" | Downloading package wmt15\_eval to /root/nltk\_data...\n",

" | Unzipping models/wmt15\_eval.zip.\n",

" | Downloading package mwa\_ppdb to /root/nltk\_data...\n",

" | Unzipping misc/mwa\_ppdb.zip.\n",

" | \n",

" Done downloading collection all\n",

"\n",

"---------------------------------------------------------------------------\n",

" d) Download l) List u) Update c) Config h) Help q) Quit\n",

"---------------------------------------------------------------------------\n",

"Downloader> q\n"

]

},

{

"output\_type": "execute\_result",

"data": {

"text/plain": [

"True"

]

},

"metadata": {},

"execution\_count": 3

}

]

},

{

"cell\_type": "code",

"metadata": {

"colab": {

"base\_uri": "https://localhost:8080/"

},

"id": "HZEiXffDz3Do",

"outputId": "8ec849e7-d54b-4ec1-ef77-fcf6b4e0354a"

},

"source": [

"nltk.corpus.mac\_morpho.words()"

],

"execution\_count": null,

"outputs": [

{

"output\_type": "execute\_result",

"data": {

"text/plain": [

"['Jersei', 'atinge', 'média', 'de', 'Cr$', '1,4', ...]"

]

},

"metadata": {},

"execution\_count": 7

}

]

},

{

"cell\_type": "code",

"metadata": {

"colab": {

"base\_uri": "https://localhost:8080/"

},

"id": "IvF41ORQ0vKL",

"outputId": "f446e8dd-eab7-4a66-dc14-7cdf14db3818"

},

"source": [

"nltk.corpus.mac\_morpho.sents()[1]"

],

"execution\_count": null,

"outputs": [

{

"output\_type": "execute\_result",

"data": {

"text/plain": [

"['Programe',\n",

" 'sua',\n",

" 'viagem',\n",

" 'a',\n",

" 'a',\n",

" 'Exposição',\n",

" 'Nacional',\n",

" 'do',\n",

" 'Zebu',\n",

" ',',\n",

" 'que',\n",

" 'começa',\n",

" 'dia',\n",

" '25']"

]

},

"metadata": {},

"execution\_count": 10

}

]

},

{

"cell\_type": "code",

"metadata": {

"colab": {

"base\_uri": "https://localhost:8080/"

},

"id": "jHMm8CPz1SIp",

"outputId": "72a8a6be-c746-4e8e-c667-7b5fa57686d6"

},

"source": [

"nltk.corpus.mac\_morpho.tagged\_words()"

],

"execution\_count": null,

"outputs": [

{

"output\_type": "execute\_result",

"data": {

"text/plain": [

"[('Jersei', 'N'), ('atinge', 'V'), ('média', 'N'), ...]"

]

},

"metadata": {},

"execution\_count": 11

}

]

},

{

"cell\_type": "code",

"metadata": {

"colab": {

"base\_uri": "https://localhost:8080/"

},

"id": "UtCC0wPq1p8Z",

"outputId": "b555d512-a4da-4b0b-b64c-98e92fbcc32f"

},

"source": [

"nltk.corpus.mac\_morpho.tagged\_sents()"

],

"execution\_count": null,

"outputs": [

{

"output\_type": "execute\_result",

"data": {

"text/plain": [

"[[('Jersei', 'N'), ('atinge', 'V'), ('média', 'N'), ('de', 'PREP'), ('Cr$', 'CUR'), ('1,4', 'NUM'), ('milhão', 'N'), ('em', 'PREP|+'), ('a', 'ART'), ('venda', 'N'), ('de', 'PREP|+'), ('a', 'ART'), ('Pinhal', 'NPROP'), ('em', 'PREP'), ('São', 'NPROP'), ('Paulo', 'NPROP')], [('Programe', 'V'), ('sua', 'PROADJ'), ('viagem', 'N'), ('a', 'PREP|+'), ('a', 'ART'), ('Exposição', 'NPROP'), ('Nacional', 'NPROP'), ('do', 'NPROP'), ('Zebu', 'NPROP'), (',', ','), ('que', 'PRO-KS-REL'), ('começa', 'V'), ('dia', 'N'), ('25', 'N|AP')], ...]"

]

},

"metadata": {},

"execution\_count": 12

}

]

},

{

"cell\_type": "code",

"metadata": {

"colab": {

"base\_uri": "https://localhost:8080/"

},

"id": "QotSBAFa-yo2",

"outputId": "f5247e67-ad04-4ec0-ab56-b845585d8366"

},

"source": [

"nltk.word\_tokenize(\"Com um passe de Eli Manning para Plaxico Burress a 39 segundos do fim, o New York Giants anotou o touchdown decisivo e derrubou o favorito New England Patriots por 17 a 14 neste domingo, em Glendale, no Super Bowl XLII.\")"

],

"execution\_count": null,

"outputs": [

{

"output\_type": "execute\_result",

"data": {

"text/plain": [

"['Com',\n",

" 'um',\n",

" 'passe',\n",

" 'de',\n",

" 'Eli',\n",

" 'Manning',\n",

" 'para',\n",

" 'Plaxico',\n",

" 'Burress',\n",

" 'a',\n",

" '39',\n",

" 'segundos',\n",

" 'do',\n",

" 'fim',\n",

" ',',\n",

" 'o',\n",

" 'New',\n",

" 'York',\n",

" 'Giants',\n",

" 'anotou',\n",

" 'o',\n",

" 'touchdown',\n",

" 'decisivo',\n",

" 'e',\n",

" 'derrubou',\n",

" 'o',\n",

" 'favorito',\n",

" 'New',\n",

" 'England',\n",

" 'Patriots',\n",

" 'por',\n",

" '17',\n",

" 'a',\n",

" '14',\n",

" 'neste',\n",

" 'domingo',\n",

" ',',\n",

" 'em',\n",

" 'Glendale',\n",

" ',',\n",

" 'no',\n",

" 'Super',\n",

" 'Bowl',\n",

" 'XLII',\n",

" '.']"

]

},

"metadata": {},

"execution\_count": 13

}

]

},

{

"cell\_type": "code",

"metadata": {

"colab": {

"base\_uri": "https://localhost:8080/"

},

"id": "09jsmJtjBFyM",

"outputId": "ba18a5d5-7f39-4929-8fed-d4905481271e"

},

"source": [

"from nltk.tokenize import RegexpTokenizer\n",

"texto = \"Com um passe de Eli Manning para Plaxico Burress a 39 segundos do fim, o New York Giants anotou o touchdown decisivo e derrubou o favorito New England Patriots por 17 a 14 neste domingo, em Glendale, no Super Bowl XLII.\"\n",

"tokenizer = RegexpTokenizer(r'\\w+')\n",

"tokens = tokenizer.tokenize(texto)\n",

"tokens"

],

"execution\_count": null,

"outputs": [

{

"output\_type": "execute\_result",

"data": {

"text/plain": [

"['Com',\n",

" 'um',\n",

" 'passe',\n",

" 'de',\n",

" 'Eli',\n",

" 'Manning',\n",

" 'para',\n",

" 'Plaxico',\n",

" 'Burress',\n",

" 'a',\n",

" '39',\n",

" 'segundos',\n",

" 'do',\n",

" 'fim',\n",

" 'o',\n",

" 'New',\n",

" 'York',\n",

" 'Giants',\n",

" 'anotou',\n",

" 'o',\n",

" 'touchdown',\n",

" 'decisivo',\n",

" 'e',\n",

" 'derrubou',\n",

" 'o',\n",

" 'favorito',\n",

" 'New',\n",

" 'England',\n",

" 'Patriots',\n",

" 'por',\n",

" '17',\n",

" 'a',\n",

" '14',\n",

" 'neste',\n",

" 'domingo',\n",

" 'em',\n",

" 'Glendale',\n",

" 'no',\n",

" 'Super',\n",

" 'Bowl',\n",

" 'XLII']"

]

},

"metadata": {},

"execution\_count": 15

}

]

},

{

"cell\_type": "code",

"metadata": {

"colab": {

"base\_uri": "https://localhost:8080/"

},

"id": "CNMY7w9jDUO7",

"outputId": "c02e4918-0c30-4766-f6dc-92842b40950a"

},

"source": [

"from nltk.tokenize import RegexpTokenizer\n",

"texto = \"Com um passe de Eli Manning para Plaxico Burress a 39 segundos do fim, o New York Giants anotou o touchdown decisivo e derrubou o favorito New England Patriots por 17 a 14 neste domingo, em Glendale, no Super Bowl XLII.\"\n",

"tokenizer = RegexpTokenizer(r'[a-zA-Z]\\w+')\n",

"tokens = tokenizer.tokenize(texto)\n",

"tokens"

],

"execution\_count": null,

"outputs": [

{

"output\_type": "execute\_result",

"data": {

"text/plain": [

"['Com',\n",

" 'um',\n",

" 'passe',\n",

" 'de',\n",

" 'Eli',\n",

" 'Manning',\n",

" 'para',\n",

" 'Plaxico',\n",

" 'Burress',\n",

" 'segundos',\n",

" 'do',\n",

" 'fim',\n",

" 'New',\n",

" 'York',\n",

" 'Giants',\n",

" 'anotou',\n",

" 'touchdown',\n",

" 'decisivo',\n",

" 'derrubou',\n",

" 'favorito',\n",

" 'New',\n",

" 'England',\n",

" 'Patriots',\n",

" 'por',\n",

" 'neste',\n",

" 'domingo',\n",

" 'em',\n",

" 'Glendale',\n",

" 'no',\n",

" 'Super',\n",

" 'Bowl',\n",

" 'XLII']"

]

},

"metadata": {},

"execution\_count": 17

}

]

},

{

"cell\_type": "code",

"metadata": {

"colab": {

"base\_uri": "https://localhost:8080/"

},

"id": "TvahZUEmFV1g",

"outputId": "1398f03d-afd2-49c1-bde8-b99d7c29636c"

},

"source": [

"texto = \"Com um passe de Eli Manning para Plaxico Burress a 39 segundos do fim, o New York Giants anotou o touchdown decisivo e derrubou o favorito New England Patriots por 17 a 14 neste domingo, em Glendale, no Super Bowl XLII.\"\n",

"\n",

"tokens = nltk.word\_tokenize(texto)\n",

"\n",

"frequencia = nltk.FreqDist(tokens)\n",

"frequencia.most\_common()"

],

"execution\_count": null,

"outputs": [

{

"output\_type": "execute\_result",

"data": {

"text/plain": [

"[(',', 3),\n",

" ('o', 3),\n",

" ('a', 2),\n",

" ('New', 2),\n",

" ('Com', 1),\n",

" ('um', 1),\n",

" ('passe', 1),\n",

" ('de', 1),\n",

" ('Eli', 1),\n",

" ('Manning', 1),\n",

" ('para', 1),\n",

" ('Plaxico', 1),\n",

" ('Burress', 1),\n",

" ('39', 1),\n",

" ('segundos', 1),\n",

" ('do', 1),\n",

" ('fim', 1),\n",

" ('York', 1),\n",

" ('Giants', 1),\n",

" ('anotou', 1),\n",

" ('touchdown', 1),\n",

" ('decisivo', 1),\n",

" ('e', 1),\n",

" ('derrubou', 1),\n",

" ('favorito', 1),\n",

" ('England', 1),\n",

" ('Patriots', 1),\n",

" ('por', 1),\n",

" ('17', 1),\n",

" ('14', 1),\n",

" ('neste', 1),\n",

" ('domingo', 1),\n",

" ('em', 1),\n",

" ('Glendale', 1),\n",

" ('no', 1),\n",

" ('Super', 1),\n",

" ('Bowl', 1),\n",

" ('XLII', 1),\n",

" ('.', 1)]"

]

},

"metadata": {},

"execution\_count": 22

}

]

},

{

"cell\_type": "code",

"metadata": {

"colab": {

"base\_uri": "https://localhost:8080/"

},

"id": "HL3l9yQnGWj7",

"outputId": "83494e8f-76ab-4807-f302-f54f25997873"

},

"source": [

"from nltk.tokenize import RegexpTokenizer\n",

"texto = \"Com um passe de Eli Manning para Plaxico Burress a 39 segundos do fim, o New York Giants anotou o touchdown decisivo e derrubou o favorito New England Patriots por 17 a 14 neste domingo, em Glendale, no Super Bowl XLII.\"\n",

"\n",

"tokenizer = RegexpTokenizer(r'\\w+')\n",

"tokens = tokenizer.tokenize(texto)\n",

"\n",

"frequencia = nltk.FreqDist(tokens)\n",

"frequencia.most\_common()"

],

"execution\_count": null,

"outputs": [

{

"output\_type": "execute\_result",

"data": {

"text/plain": [

"[('o', 3),\n",

" ('a', 2),\n",

" ('New', 2),\n",

" ('Com', 1),\n",

" ('um', 1),\n",

" ('passe', 1),\n",

" ('de', 1),\n",

" ('Eli', 1),\n",

" ('Manning', 1),\n",

" ('para', 1),\n",

" ('Plaxico', 1),\n",

" ('Burress', 1),\n",

" ('39', 1),\n",

" ('segundos', 1),\n",

" ('do', 1),\n",

" ('fim', 1),\n",

" ('York', 1),\n",

" ('Giants', 1),\n",

" ('anotou', 1),\n",

" ('touchdown', 1),\n",

" ('decisivo', 1),\n",

" ('e', 1),\n",

" ('derrubou', 1),\n",

" ('favorito', 1),\n",

" ('England', 1),\n",

" ('Patriots', 1),\n",

" ('por', 1),\n",

" ('17', 1),\n",

" ('14', 1),\n",

" ('neste', 1),\n",

" ('domingo', 1),\n",

" ('em', 1),\n",

" ('Glendale', 1),\n",

" ('no', 1),\n",

" ('Super', 1),\n",

" ('Bowl', 1),\n",

" ('XLII', 1)]"

]

},

"metadata": {},

"execution\_count": 23

}

]

},

{

"cell\_type": "code",

"metadata": {

"colab": {

"base\_uri": "https://localhost:8080/"

},

"id": "QYomAwvfHTHF",

"outputId": "73761fdf-fa9c-4c56-e0af-ebef8698ae84"

},

"source": [

"corpus = open('/content/drive/MyDrive/recursos/corpus\_teste.txt').read()\n",

"print(corpus)"

],

"execution\_count": null,

"outputs": [

{

"output\_type": "stream",

"name": "stdout",

"text": [

"Giants batem os Patriots no Super Bowl XLII\n",

"Azarões acabam com a invencibilidade de New England e ficam com o título da temporada\n",

"04/02/2008 - 01h07m - Atualizado em 04/02/2008 - 09h49m\n",

"\n",

"Com um passe de Eli Manning para Plaxico Burress a 39 segundos do fim, o New York Giants anotou o touchdown decisivo e derrubou o favorito New England Patriots por 17 a 14 neste domingo, em Glendale, no Super Bowl XLII. O resultado, uma das maiores zebras da história do Super Bowl, acabou com a temporada perfeita de Tom Brady e companhia, que esperavam fazer história ao levantar o troféu da NFL sem sofrer uma derrota no ano. \n",

"\n",

"A vitória dos Giants, porém, também ficará para a história. Pela primeira vez, irmãos quarterbacks triunfam no Super Bowl em temporadas consecutivas. No ano passado, Peyton Manning, irmão de Eli, chegou ao título máximo da NFL pelo Indianapolis Colts.\n",

"\n",

"A partida\n",

"\n",

"Os Giants começaram com a posse de bola, e mostraram logo que iriam alongar ao máximo suas posses de bola. Misturando corridas com Brandon Jacobs e passes curtos, o time de Nova York chegou à red zone logo na primeira campanha. O avanço, no entanto, parou na linha de 17 jardas e Lawrence Tynes converteu o field goal de 32 jardas para abrir o placar.\n",

"\n",

"Eli Manning e companhia ficaram 9m54s com a bola, mas o ataque dos Patriots não entrou em campo frio. Logo no retorno do kickoff, o running back Laurence Maroney avançou 43 jardas, deixando Tom Brady em boa posição. Com passes curtos, os Patriots chegaram à linha de 17 jardas e, graças a uma penalidade (interferência de passe) do linebacker Antonio Pierce, alcançaram a linha de uma jarda. Maroney avançou pelo chão e anotou o primeiro touchdown do jogo.\n",

"\n",

"Os Giants pareciam rumo à virada na campanha seguinte. Manning achou Amani Toomer para um avanço de 38 jardas, e o time de Nova York entrou novamente na red zone. Com a bola na linha de 14 jardas dos Patriots, os Giants sofreram um revés. Manning passou para Steve Smith, que soltou a bola. Ellis Hobbs aproveitou, tomou a posse para os Patriots, e avançou 23 jardas. \n",

"\n",

"A defesa de Nova York manteve o jogo equilibrado. Com dois sacks seguidos, os Giants forçaram o punt e recuperaram a bola. Mas a campanha seguinte provou ser outra decepção para Nova York. O time chegou à linha de 25 jardas, mas Manning sofreu um sack e cometeu um fumble, e o ataque voltou para a linha de 39 jardas, não conseguindo pontuar mais uma vez.\n",

"\n",

"Os Patriots tiveram uma última chance de marcar antes do intervalo, mas, a 22 segundos do fim do segundo período, Brady foi novamente sacado. Desta vez, ele cometeu o fumble e os Giants tomaram a posse de bola. Manning tentou um passe longo, de 50 jardas, nos últimos segundos, mas não teve sucesso. \n",

"\n",

"O jogo continuou amarrado no terceiro quarto, com as defesas levando a melhor sobre os ataques. A única chance de pontuar do período foi dos Patriots, que chegaram à linha de 31 jardas dos Giants. O técnico Bill Bellichick, porém, optou por uma quarta descida em vez de um field goal. Brady tentou um passe para Jabar Gaffney, mas não conseguiu completar.\n",

"\n",

"O último período começou arrasador para os Giants. na primeira jogada, Manning achou o tight end Kevin Boss, para um incrível avanço de 45 jardas, que deixou o time na linha de 35 dos Patriots. Outro lançamento, desta vez para Steve Smith, marcou o avanço até a linha de 12 jardas. Duas jogadas depois, David Tyree pegou um passe de cinco jardas na end zone para anotar o touchdown e virar o jogo.\n",

"\n",

"Na hora da decisão, o ataque dos Patriots voltou a funcionar. Com uma série de passes curtos e variados, Brady achou Wes Welker, Randy Moss e Kevin Faulk seguidas vezes até chegar à red zone. A 2m45s do fim, o quarterback conectou mais uma vez com Moss, que se desmarcou e ficou livre na lateral direita da end zone.\n",

"\n",

"Quando os fãs de New England já comemoravam a vitória, o inesperado aconteceu. Em uma jogada incrível, Eli Manning se soltou de dois marcadores que o seguravam pela camisa e, na corrida, lançou para Amani Toomer. O wide receiver, bem marcado, saltou e conseguiu a fazer recepção para um avanço de 32 jardas, deixando os Giants na linha de 24 de New England.\n",

"\n",

"Quatro jogadas depois, a 39 segundos do fim, Manning achou Plaxico Burress na end zone para conseguir o touchdown do título.\n"

]

}

]

},

{

"cell\_type": "code",

"metadata": {

"colab": {

"base\_uri": "https://localhost:8080/"

},

"id": "C8fB03n6H0oS",

"outputId": "5fc1ce22-ccb4-48f0-ba01-4cf3d25b52e6"

},

"source": [

"from nltk.tokenize import RegexpTokenizer\n",

"\n",

"tokenizer = RegexpTokenizer(r'\\w+')\n",

"tokens = tokenizer.tokenize(corpus)\n",

"\n",

"frequencia = nltk.FreqDist(tokens)\n",

"frequencia.most\_common()"

],

"execution\_count": null,

"outputs": [

{

"output\_type": "execute\_result",

"data": {

"text/plain": [

"[('de', 34),\n",

" ('o', 26),\n",

" ('a', 23),\n",

" ('e', 21),\n",

" ('para', 16),\n",

" ('jardas', 15),\n",

" ('do', 12),\n",

" ('na', 12),\n",

" ('Giants', 11),\n",

" ('um', 11),\n",

" ('os', 10),\n",

" ('Patriots', 10),\n",

" ('Manning', 10),\n",

" ('uma', 10),\n",

" ('linha', 10),\n",

" ('com', 8),\n",

" ('no', 7),\n",

" ('O', 7),\n",

" ('que', 7),\n",

" ('dos', 7),\n",

" ('bola', 7),\n",

" ('da', 6),\n",

" ('em', 6),\n",

" ('vez', 6),\n",

" ('à', 6),\n",

" ('zone', 6),\n",

" ('New', 5),\n",

" ('Com', 5),\n",

" ('passe', 5),\n",

" ('York', 5),\n",

" ('Brady', 5),\n",

" ('A', 5),\n",

" ('avanço', 5),\n",

" ('mas', 5),\n",

" ('Super', 4),\n",

" ('Bowl', 4),\n",

" ('England', 4),\n",

" ('Eli', 4),\n",

" ('segundos', 4),\n",

" ('fim', 4),\n",

" ('touchdown', 4),\n",

" ('time', 4),\n",

" ('Nova', 4),\n",

" ('não', 4),\n",

" ('jogo', 4),\n",

" ('achou', 4),\n",

" ('end', 4),\n",

" ('título', 3),\n",

" ('39', 3),\n",

" ('17', 3),\n",

" ('história', 3),\n",

" ('ao', 3),\n",

" ('primeira', 3),\n",

" ('chegou', 3),\n",

" ('Os', 3),\n",

" ('posse', 3),\n",

" ('passes', 3),\n",

" ('curtos', 3),\n",

" ('red', 3),\n",

" ('campanha', 3),\n",

" ('ataque', 3),\n",

" ('avançou', 3),\n",

" ('período', 3),\n",

" ('XLII', 2),\n",

" ('temporada', 2),\n",

" ('04', 2),\n",

" ('02', 2),\n",

" ('2008', 2),\n",

" ('Plaxico', 2),\n",

" ('Burress', 2),\n",

" ('anotou', 2),\n",

" ('por', 2),\n",

" ('14', 2),\n",

" ('Tom', 2),\n",

" ('companhia', 2),\n",

" ('fazer', 2),\n",

" ('NFL', 2),\n",

" ('ano', 2),\n",

" ('vitória', 2),\n",

" ('porém', 2),\n",

" ('máximo', 2),\n",

" ('pelo', 2),\n",

" ('logo', 2),\n",

" ('field', 2),\n",

" ('goal', 2),\n",

" ('32', 2),\n",

" ('entrou', 2),\n",

" ('Maroney', 2),\n",

" ('deixando', 2),\n",

" ('chegaram', 2),\n",

" ('seguinte', 2),\n",

" ('Amani', 2),\n",

" ('Toomer', 2),\n",

" ('novamente', 2),\n",

" ('Steve', 2),\n",

" ('Smith', 2),\n",

" ('soltou', 2),\n",

" ('dois', 2),\n",

" ('cometeu', 2),\n",

" ('fumble', 2),\n",

" ('voltou', 2),\n",

" ('pontuar', 2),\n",

" ('mais', 2),\n",

" ('chance', 2),\n",

" ('foi', 2),\n",

" ('tentou', 2),\n",

" ('conseguiu', 2),\n",

" ('jogada', 2),\n",

" ('Kevin', 2),\n",

" ('incrível', 2),\n",

" ('até', 2),\n",

" ('jogadas', 2),\n",

" ('depois', 2),\n",

" ('Moss', 2),\n",

" ('se', 2),\n",

" ('batem', 1),\n",

" ('Azarões', 1),\n",

" ('acabam', 1),\n",

" ('invencibilidade', 1),\n",

" ('ficam', 1),\n",

" ('01h07m', 1),\n",

" ('Atualizado', 1),\n",

" ('09h49m', 1),\n",

" ('decisivo', 1),\n",

" ('derrubou', 1),\n",

" ('favorito', 1),\n",

" ('neste', 1),\n",

" ('domingo', 1),\n",

" ('Glendale', 1),\n",

" ('resultado', 1),\n",

" ('das', 1),\n",

" ('maiores', 1),\n",

" ('zebras', 1),\n",

" ('acabou', 1),\n",

" ('perfeita', 1),\n",

" ('esperavam', 1),\n",

" ('levantar', 1),\n",

" ('troféu', 1),\n",

" ('sem', 1),\n",

" ('sofrer', 1),\n",

" ('derrota', 1),\n",

" ('também', 1),\n",

" ('ficará', 1),\n",

" ('Pela', 1),\n",

" ('irmãos', 1),\n",

" ('quarterbacks', 1),\n",

" ('triunfam', 1),\n",

" ('temporadas', 1),\n",

" ('consecutivas', 1),\n",

" ('No', 1),\n",

" ('passado', 1),\n",

" ('Peyton', 1),\n",

" ('irmão', 1),\n",

" ('Indianapolis', 1),\n",

" ('Colts', 1),\n",

" ('partida', 1),\n",

" ('começaram', 1),\n",

" ('mostraram', 1),\n",

" ('iriam', 1),\n",

" ('alongar', 1),\n",

" ('suas', 1),\n",

" ('posses', 1),\n",

" ('Misturando', 1),\n",

" ('corridas', 1),\n",

" ('Brandon', 1),\n",

" ('Jacobs', 1),\n",

" ('entanto', 1),\n",

" ('parou', 1),\n",

" ('Lawrence', 1),\n",

" ('Tynes', 1),\n",

" ('converteu', 1),\n",

" ('abrir', 1),\n",

" ('placar', 1),\n",

" ('ficaram', 1),\n",

" ('9m54s', 1),\n",

" ('campo', 1),\n",

" ('frio', 1),\n",

" ('Logo', 1),\n",

" ('retorno', 1),\n",

" ('kickoff', 1),\n",

" ('running', 1),\n",

" ('back', 1),\n",

" ('Laurence', 1),\n",

" ('43', 1),\n",

" ('boa', 1),\n",

" ('posição', 1),\n",

" ('graças', 1),\n",

" ('penalidade', 1),\n",

" ('interferência', 1),\n",

" ('linebacker', 1),\n",

" ('Antonio', 1),\n",

" ('Pierce', 1),\n",

" ('alcançaram', 1),\n",

" ('jarda', 1),\n",

" ('chão', 1),\n",

" ('primeiro', 1),\n",

" ('pareciam', 1),\n",

" ('rumo', 1),\n",

" ('virada', 1),\n",

" ('38', 1),\n",

" ('sofreram', 1),\n",

" ('revés', 1),\n",

" ('passou', 1),\n",

" ('Ellis', 1),\n",

" ('Hobbs', 1),\n",

" ('aproveitou', 1),\n",

" ('tomou', 1),\n",

" ('23', 1),\n",

" ('defesa', 1),\n",

" ('manteve', 1),\n",

" ('equilibrado', 1),\n",

" ('sacks', 1),\n",

" ('seguidos', 1),\n",

" ('forçaram', 1),\n",

" ('punt', 1),\n",

" ('recuperaram', 1),\n",

" ('Mas', 1),\n",

" ('provou', 1),\n",

" ('ser', 1),\n",

" ('outra', 1),\n",

" ('decepção', 1),\n",

" ('25', 1),\n",

" ('sofreu', 1),\n",

" ('sack', 1),\n",

" ('conseguindo', 1),\n",

" ('tiveram', 1),\n",

" ('última', 1),\n",

" ('marcar', 1),\n",

" ('antes', 1),\n",

" ('intervalo', 1),\n",

" ('22', 1),\n",

" ('segundo', 1),\n",

" ('sacado', 1),\n",

" ('Desta', 1),\n",

" ('ele', 1),\n",

" ('tomaram', 1),\n",

" ('longo', 1),\n",

" ('50', 1),\n",

" ('nos', 1),\n",

" ('últimos', 1),\n",

" ('teve', 1),\n",

" ('sucesso', 1),\n",

" ('continuou', 1),\n",

" ('amarrado', 1),\n",

" ('terceiro', 1),\n",

" ('quarto', 1),\n",

" ('as', 1),\n",

" ('defesas', 1),\n",

" ('levando', 1),\n",

" ('melhor', 1),\n",

" ('sobre', 1),\n",

" ('ataques', 1),\n",

" ('única', 1),\n",

" ('31', 1),\n",

" ('técnico', 1),\n",

" ('Bill', 1),\n",

" ('Bellichick', 1),\n",

" ('optou', 1),\n",

" ('quarta', 1),\n",

" ('descida', 1),\n",

" ('Jabar', 1),\n",

" ('Gaffney', 1),\n",

" ('completar', 1),\n",

" ('último', 1),\n",

" ('começou', 1),\n",

" ('arrasador', 1),\n",

" ('tight', 1),\n",

" ('Boss', 1),\n",

" ('45', 1),\n",

" ('deixou', 1),\n",

" ('35', 1),\n",

" ('Outro', 1),\n",

" ('lançamento', 1),\n",

" ('desta', 1),\n",

" ('marcou', 1),\n",

" ('12', 1),\n",

" ('Duas', 1),\n",

" ('David', 1),\n",

" ('Tyree', 1),\n",

" ('pegou', 1),\n",

" ('cinco', 1),\n",

" ('anotar', 1),\n",

" ('virar', 1),\n",

" ('Na', 1),\n",

" ('hora', 1),\n",

" ('decisão', 1),\n",

" ('funcionar', 1),\n",

" ('série', 1),\n",

" ('variados', 1),\n",

" ('Wes', 1),\n",

" ('Welker', 1),\n",

" ('Randy', 1),\n",

" ('Faulk', 1),\n",

" ('seguidas', 1),\n",

" ('vezes', 1),\n",

" ('chegar', 1),\n",

" ('2m45s', 1),\n",

" ('quarterback', 1),\n",

" ('conectou', 1),\n",

" ('desmarcou', 1),\n",

" ('ficou', 1),\n",

" ('livre', 1),\n",

" ('lateral', 1),\n",

" ('direita', 1),\n",

" ('Quando', 1),\n",

" ('fãs', 1),\n",

" ('já', 1),\n",

" ('comemoravam', 1),\n",

" ('inesperado', 1),\n",

" ('aconteceu', 1),\n",

" ('Em', 1),\n",

" ('marcadores', 1),\n",

" ('seguravam', 1),\n",

" ('pela', 1),\n",

" ('camisa', 1),\n",

" ('corrida', 1),\n",

" ('lançou', 1),\n",

" ('wide', 1),\n",

" ('receiver', 1),\n",

" ('bem', 1),\n",

" ('marcado', 1),\n",

" ('saltou', 1),\n",

" ('recepção', 1),\n",

" ('24', 1),\n",

" ('Quatro', 1),\n",

" ('conseguir', 1)]"

]

},

"metadata": {},

"execution\_count": 29

}

]

},

{

"cell\_type": "code",

"metadata": {

"id": "jx-YnEGnIsVb"

},

"source": [

"from nltk.tokenize import RegexpTokenizer\n",

"\n",

"tokenizer = RegexpTokenizer(r'[a-zA-Z]\\w\*')\n",

"tokens = tokenizer.tokenize(corpus)\n",

"\n",

"nova\_lista = []\n",

"\n",

"for token in tokens:\n",

" nova\_lista.append(token.lower())\n",

"\n",

"frequencia = nltk.FreqDist(nova\_lista)\n",

"frequencia.most\_common()"

],

"execution\_count": null,

"outputs": []

},

{

"cell\_type": "code",

"metadata": {

"id": "TMDOSjJhMPih"

},

"source": [

"stopwords = nltk.corpus.stopwords.words('portuguese')"

],

"execution\_count": null,

"outputs": []

},

{

"cell\_type": "code",

"metadata": {

"id": "uzzoj8scNBeS"

},

"source": [

"from nltk.tokenize import RegexpTokenizer\n",

"\n",

"tokenizer = RegexpTokenizer(r'[a-zA-Z]\\w\*')\n",

"tokens = tokenizer.tokenize(corpus)\n",

"\n",

"nova\_lista = []\n",

"\n",

"for token in tokens:\n",

" if token.lower() not in stopwords:\n",

" nova\_lista.append(token.lower())\n",

"\n",

"frequencia = nltk.FreqDist(nova\_lista)\n",

"frequencia.most\_common()"

],

"execution\_count": null,

"outputs": []

},

{

"cell\_type": "code",

"metadata": {

"colab": {

"base\_uri": "https://localhost:8080/"

},

"id": "xVV5GpUJPG4o",

"outputId": "0cc44970-4634-4da1-ac7f-a43cffae1bc2"

},

"source": [

"from nltk.tokenize import RegexpTokenizer\n",

"\n",

"tokenizer = RegexpTokenizer(r'[a-zA-Z]\\w\*')\n",

"tokens = tokenizer.tokenize(corpus)\n",

"\n",

"nova\_lista = []\n",

"\n",

"#for token in tokens:\n",

"# if token.lower() not in stopwords:\n",

"# nova\_lista.append(token.lower())\n",

"\n",

"nova\_lista = [token.lower() for token in tokens if token.lower() not in stopwords]\n",

"\n",

"frequencia = nltk.FreqDist(nova\_lista)\n",

"frequencia.most\_common()"

],

"execution\_count": null,

"outputs": [

{

"output\_type": "execute\_result",

"data": {

"text/plain": [

"[('jardas', 15),\n",

" ('giants', 11),\n",

" ('patriots', 10),\n",

" ('manning', 10),\n",

" ('linha', 10),\n",

" ('bola', 7),\n",

" ('vez', 6),\n",

" ('zone', 6),\n",

" ('new', 5),\n",

" ('passe', 5),\n",

" ('york', 5),\n",

" ('brady', 5),\n",

" ('avanço', 5),\n",

" ('super', 4),\n",

" ('bowl', 4),\n",

" ('england', 4),\n",

" ('eli', 4),\n",

" ('segundos', 4),\n",

" ('fim', 4),\n",

" ('touchdown', 4),\n",

" ('time', 4),\n",

" ('nova', 4),\n",

" ('jogo', 4),\n",

" ('achou', 4),\n",

" ('end', 4),\n",

" ('título', 3),\n",

" ('história', 3),\n",

" ('primeira', 3),\n",

" ('chegou', 3),\n",

" ('posse', 3),\n",

" ('logo', 3),\n",

" ('passes', 3),\n",

" ('curtos', 3),\n",

" ('red', 3),\n",

" ('campanha', 3),\n",

" ('ataque', 3),\n",

" ('avançou', 3),\n",

" ('período', 3),\n",

" ('xlii', 2),\n",

" ('temporada', 2),\n",

" ('plaxico', 2),\n",

" ('burress', 2),\n",

" ('anotou', 2),\n",

" ('tom', 2),\n",

" ('companhia', 2),\n",

" ('fazer', 2),\n",

" ('nfl', 2),\n",

" ('ano', 2),\n",

" ('vitória', 2),\n",

" ('porém', 2),\n",

" ('máximo', 2),\n",

" ('field', 2),\n",

" ('goal', 2),\n",

" ('entrou', 2),\n",

" ('maroney', 2),\n",

" ('deixando', 2),\n",

" ('chegaram', 2),\n",

" ('seguinte', 2),\n",

" ('amani', 2),\n",

" ('toomer', 2),\n",

" ('novamente', 2),\n",

" ('steve', 2),\n",

" ('smith', 2),\n",

" ('soltou', 2),\n",

" ('dois', 2),\n",

" ('cometeu', 2),\n",

" ('fumble', 2),\n",

" ('voltou', 2),\n",

" ('pontuar', 2),\n",

" ('chance', 2),\n",

" ('desta', 2),\n",

" ('tentou', 2),\n",

" ('conseguiu', 2),\n",

" ('jogada', 2),\n",

" ('kevin', 2),\n",

" ('incrível', 2),\n",

" ('jogadas', 2),\n",

" ('moss', 2),\n",

" ('batem', 1),\n",

" ('azarões', 1),\n",

" ('acabam', 1),\n",

" ('invencibilidade', 1),\n",

" ('ficam', 1),\n",

" ('h07m', 1),\n",

" ('atualizado', 1),\n",

" ('h49m', 1),\n",

" ('decisivo', 1),\n",

" ('derrubou', 1),\n",

" ('favorito', 1),\n",

" ('neste', 1),\n",

" ('domingo', 1),\n",

" ('glendale', 1),\n",

" ('resultado', 1),\n",

" ('maiores', 1),\n",

" ('zebras', 1),\n",

" ('acabou', 1),\n",

" ('perfeita', 1),\n",

" ('esperavam', 1),\n",

" ('levantar', 1),\n",

" ('troféu', 1),\n",

" ('sofrer', 1),\n",

" ('derrota', 1),\n",

" ('ficará', 1),\n",

" ('irmãos', 1),\n",

" ('quarterbacks', 1),\n",

" ('triunfam', 1),\n",

" ('temporadas', 1),\n",

" ('consecutivas', 1),\n",

" ('passado', 1),\n",

" ('peyton', 1),\n",

" ('irmão', 1),\n",

" ('indianapolis', 1),\n",

" ('colts', 1),\n",

" ('partida', 1),\n",

" ('começaram', 1),\n",

" ('mostraram', 1),\n",

" ('iriam', 1),\n",

" ('alongar', 1),\n",

" ('posses', 1),\n",

" ('misturando', 1),\n",

" ('corridas', 1),\n",

" ('brandon', 1),\n",

" ('jacobs', 1),\n",

" ('entanto', 1),\n",

" ('parou', 1),\n",

" ('lawrence', 1),\n",

" ('tynes', 1),\n",

" ('converteu', 1),\n",

" ('abrir', 1),\n",

" ('placar', 1),\n",

" ('ficaram', 1),\n",

" ('m54s', 1),\n",

" ('campo', 1),\n",

" ('frio', 1),\n",

" ('retorno', 1),\n",

" ('kickoff', 1),\n",

" ('running', 1),\n",

" ('back', 1),\n",

" ('laurence', 1),\n",

" ('boa', 1),\n",

" ('posição', 1),\n",

" ('graças', 1),\n",

" ('penalidade', 1),\n",

" ('interferência', 1),\n",

" ('linebacker', 1),\n",

" ('antonio', 1),\n",

" ('pierce', 1),\n",

" ('alcançaram', 1),\n",

" ('jarda', 1),\n",

" ('chão', 1),\n",

" ('primeiro', 1),\n",

" ('pareciam', 1),\n",

" ('rumo', 1),\n",

" ('virada', 1),\n",

" ('sofreram', 1),\n",

" ('revés', 1),\n",

" ('passou', 1),\n",

" ('ellis', 1),\n",

" ('hobbs', 1),\n",

" ('aproveitou', 1),\n",

" ('tomou', 1),\n",

" ('defesa', 1),\n",

" ('manteve', 1),\n",

" ('equilibrado', 1),\n",

" ('sacks', 1),\n",

" ('seguidos', 1),\n",

" ('forçaram', 1),\n",

" ('punt', 1),\n",

" ('recuperaram', 1),\n",

" ('provou', 1),\n",

" ('ser', 1),\n",

" ('outra', 1),\n",

" ('decepção', 1),\n",

" ('sofreu', 1),\n",

" ('sack', 1),\n",

" ('conseguindo', 1),\n",

" ('ltima', 1),\n",

" ('marcar', 1),\n",

" ('antes', 1),\n",

" ('intervalo', 1),\n",

" ('segundo', 1),\n",

" ('sacado', 1),\n",

" ('tomaram', 1),\n",

" ('longo', 1),\n",

" ('ltimos', 1),\n",

" ('sucesso', 1),\n",

" ('continuou', 1),\n",

" ('amarrado', 1),\n",

" ('terceiro', 1),\n",

" ('quarto', 1),\n",

" ('defesas', 1),\n",

" ('levando', 1),\n",

" ('melhor', 1),\n",

" ('sobre', 1),\n",

" ('ataques', 1),\n",

" ('nica', 1),\n",

" ('técnico', 1),\n",

" ('bill', 1),\n",

" ('bellichick', 1),\n",

" ('optou', 1),\n",

" ('quarta', 1),\n",

" ('descida', 1),\n",

" ('jabar', 1),\n",

" ('gaffney', 1),\n",

" ('completar', 1),\n",

" ('ltimo', 1),\n",

" ('começou', 1),\n",

" ('arrasador', 1),\n",

" ('tight', 1),\n",

" ('boss', 1),\n",

" ('deixou', 1),\n",

" ('outro', 1),\n",

" ('lançamento', 1),\n",

" ('marcou', 1),\n",

" ('duas', 1),\n",

" ('david', 1),\n",

" ('tyree', 1),\n",

" ('pegou', 1),\n",

" ('cinco', 1),\n",

" ('anotar', 1),\n",

" ('virar', 1),\n",

" ('hora', 1),\n",

" ('decisão', 1),\n",

" ('funcionar', 1),\n",

" ('série', 1),\n",

" ('variados', 1),\n",

" ('wes', 1),\n",

" ('welker', 1),\n",

" ('randy', 1),\n",

" ('faulk', 1),\n",

" ('seguidas', 1),\n",

" ('vezes', 1),\n",

" ('chegar', 1),\n",

" ('m45s', 1),\n",

" ('quarterback', 1),\n",

" ('conectou', 1),\n",

" ('desmarcou', 1),\n",

" ('ficou', 1),\n",

" ('livre', 1),\n",

" ('lateral', 1),\n",

" ('direita', 1),\n",

" ('fãs', 1),\n",

" ('comemoravam', 1),\n",

" ('inesperado', 1),\n",

" ('aconteceu', 1),\n",

" ('marcadores', 1),\n",

" ('seguravam', 1),\n",

" ('camisa', 1),\n",

" ('corrida', 1),\n",

" ('lançou', 1),\n",

" ('wide', 1),\n",

" ('receiver', 1),\n",

" ('bem', 1),\n",

" ('marcado', 1),\n",

" ('saltou', 1),\n",

" ('recepção', 1),\n",

" ('quatro', 1),\n",

" ('conseguir', 1)]"

]

},

"metadata": {},

"execution\_count": 39

}

]

}

]

}