PROCESSAMENTO DE LINGUAGEM NATURAL NOS DIAS ATUAIS

FEITO POR GYO123

MARÇO DE 2022

DEFINIÇÃO

1

O processamento de linguagem natural ou mineração de textos é uma ampliação da mineração de dados.

2

Ela pode ser definida como um processo de extração de informações desconhecidas e que podem ser úteis de textos escritos em linguagem natural.

3

É um sub-campo que foca mais em análise exploratória de dados e é popularmente conhecida como aprendizado não supervisionado.

CONCEITOS

CORPUS

TOKENIZATION

LEMMATIZING (LEMMA)

WORDCLOUD

CORPUS

é um conjunto de textos escritos e registros orais em uma determinada lingua e que serve como base para uma análise ou o estudo de um fenômeno

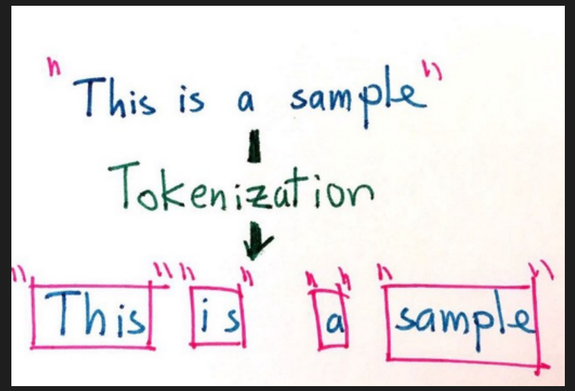
O COMPARA é um corpus eletrônico paralelo[i] cuja estrutura foi inspirada no ENPC (English-Norwegian Parallel Corpus, Johansson et al. 1999). Os textos que constituem o corpus são originais em língua inglesa e portuguesa e as suas traduções para português e inglês. O corpus é extensível (podendo conter um número ilimitado de textos), é de acesso gratuito através da Web, e foi desenvolvido para ser útil tanto para pessoas com pouca ou nenhuma experiência prévia na utilização de corpora, como para utilizadores experimentados.

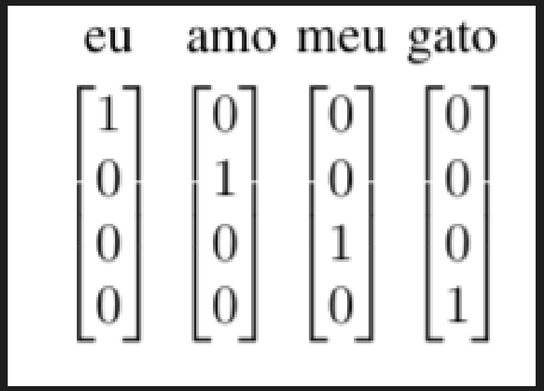
Determinou-se que o corpus seria extensível por duas razões de ordem prática. Primeiro, porque isso permitiria torná-lo operacional logo de início e ainda com poucos textos, uma vez que não existia (e ainda não parece existir) nenhum outro corpus paralelo de livre acesso que inclua o português. Segundo, porque assim poderiam ser os próprios utilizadores do corpus a indicar o melhor caminho para a sua expansão. Esta escolha também acabou por facilitar a correção de alguns problemas inicias e a incorporação de novas funcionalidades, já que o corpus ainda era pequeno e havia menos alterações a fazer.

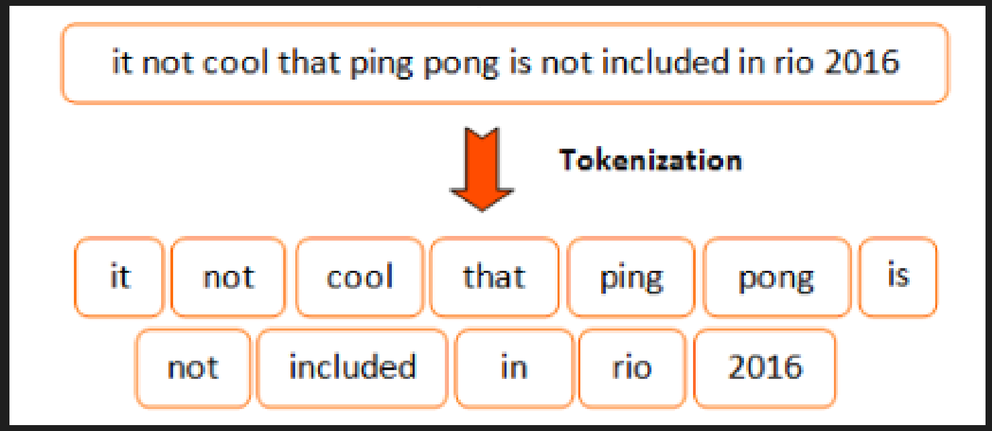
Quisemos que o público-alvo do COMPARA abrangesse todos os que estudam, investigam ou trabalham com o português e o inglês, entre os quais encontram-se falantes nativos de português a aprender inglês, falantes nativos de inglês a aprender português, professores e autores de materiais didáticos de português e inglês língua estrangeira, tradutores, professores e estudantes de tradução, lexicógrafos, engenheiros da linguagem, lingüistas interessados no estudo da tradução e investigadores na área da literatura comparada. Todos estes grupos são utilizadores potenciais do COMPARA. Tivemos, por isso, especial preocupação em assegurar que uma gama vasta de utilizadores pudesse facilmente servir-se do COMPARA, e que o corpus não se limitasse a ser útil apenas às pessoas já habituadas a trabalhar com corpora. O nosso objectivo foi desenvolver um sistema que não afastasse – ou assustasse – quem nunca tivesse lidado com um corpus informatizado.

TOKENIZATION

É o processo de separar a sentença em suas partes: palavras, pontos, símbolos etc.

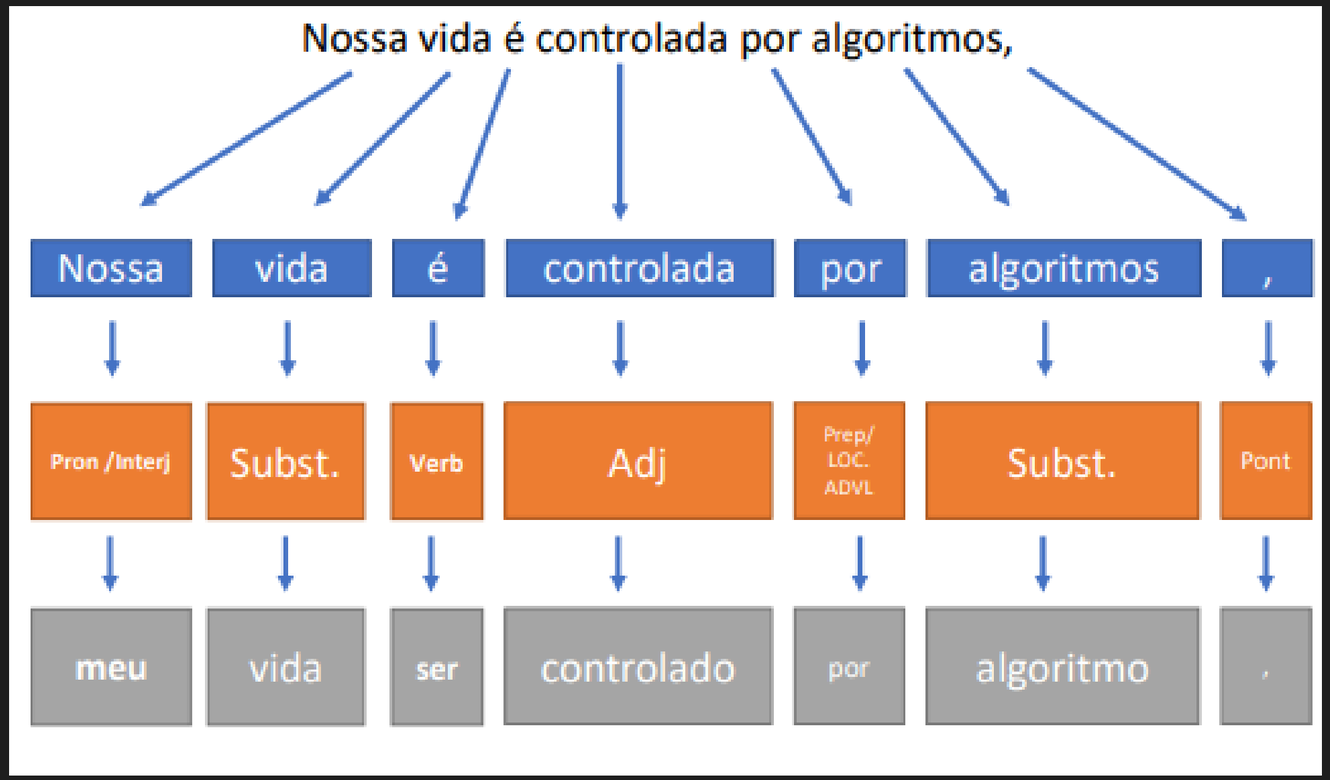






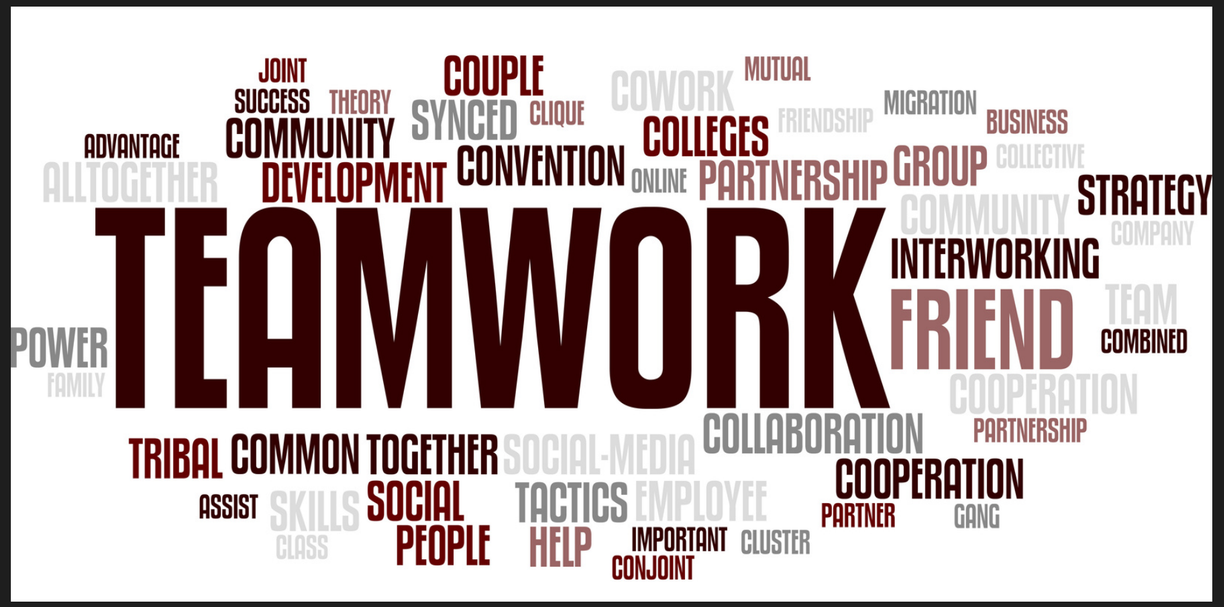
LEMMATIZING

É o processo de agrupar as formas flexionadas de uma palavra para que possam ser analisadas como um único item, identificado pelo lema da palavra, ou forma de dicionário



WORDCLOUD

Nuvem de palavras (word cloud) é um gráfico digital que mostra o grau de frequência das palavras em um texto. Quanto mais a palavra é utilizada, mais chamativa é a representação dessa palavra no gráfico.



APLICAÇÕES

O processamento de linguagem natural possui aplicações em diversas áreas científicas e comerciais. A seguir Veja alguns exemplos:

CORREÇÃO ORTOGRÁFICA

MINERAÇÃO NO TWITTER

WORDCLOUD

Correção Ortográfica

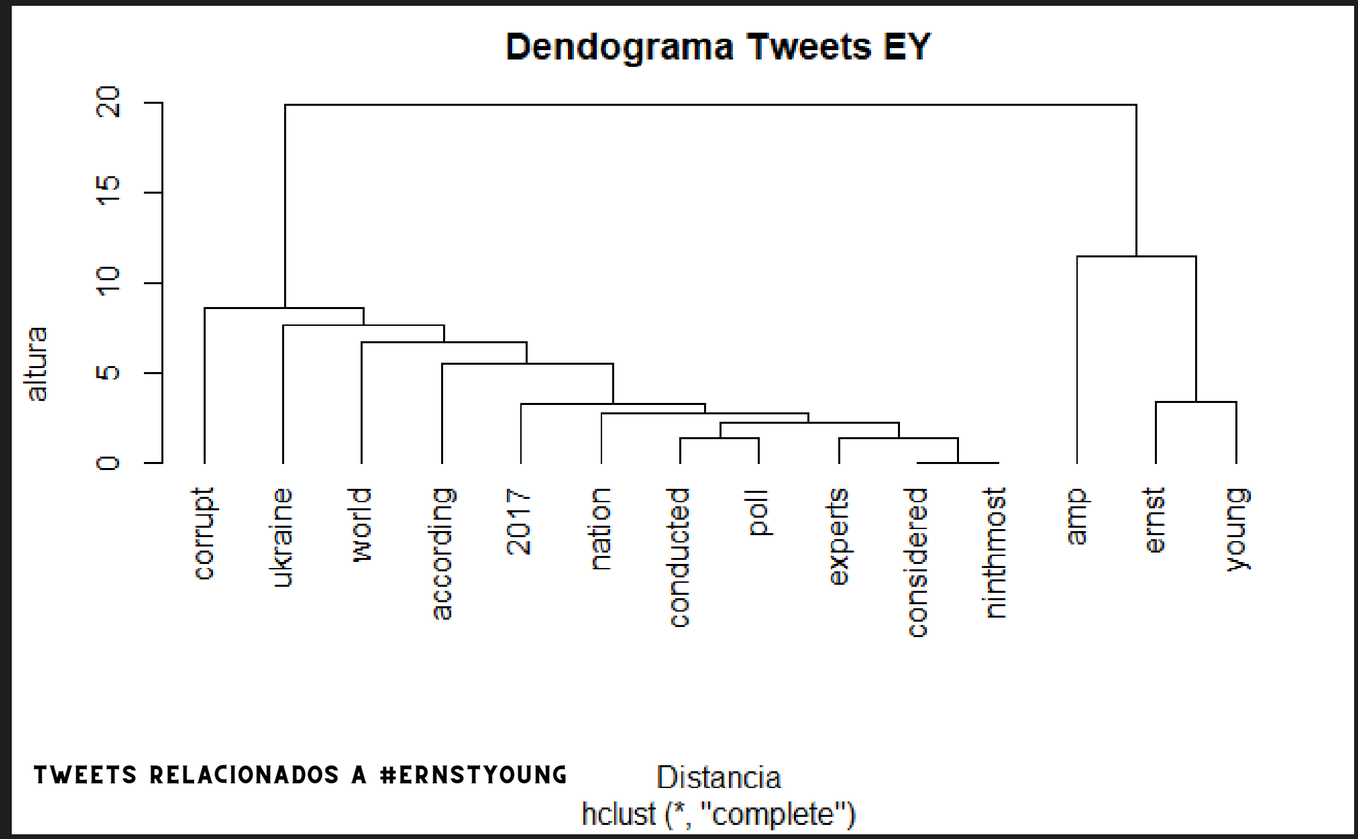
É uma aplicação muito utilizada onde o usuário pode conferir se há erros ortográficos em textos e/ou sites. São baseados em algoritmo que compara as palavras extraídas com aquelas definidas em seu dicionário.

O principa objetivo da nossa área e Consultori é promover mudanças e melhorias de processos ou reduções de riscs, otimizando os resultados das organizaçõs. Para isso, o time de Consulting utiliza as melhores metodologias de gestão de negócios, combinadas co recursos tecnológicos que são indispensáveis para melhori de performance atualmente. Gestão de finanças, clientes, TI, cadeia de suprimentos, riss, inovação, estratégia, gestão da mudança, cibersegurança, robotização: são temas como esses que nossos profissionais de Consulting transformam em soluções e oportunidades para nossos clients. **(grifo nosso)**

FONTE: SOBRE A EY I EY - #BEYELLOW

Mineração no Twitter

O twitter é uma grande fonte de informações e por meio da biblioteca rtweet da linguagem de programação R, você pode classificar e filtrar tweets especificos conforme anexo e exemplo ao lado feito na linguagem R.



WORDCLOUD

Feito na linguagem Python, o gráfico ao lado demonstra as palavras mais utilizadas no Relatório Anual da EY 2020.

Referências

- Silge And Robinson, Julia and David. Text Mining with R, 1º ed.. Boston: O'REILLY. 2017
- Lovato, Gustavo. Aplicação de Mineração de Textos na análise de produções textuais, disponível em www.repositorio.ucs.br/xmlui/handle/11338/1516 Caxias do Sul 2015,
- Ladeira, Ana Paula. Processamento de linguagem natural: caracterização da produção científica / Ana Paula Ladeira. Minas Gerais - 2010
- Mendes, Charles. Text Mining Conceitos e Práticas usando a linguagem R. Youtube, 06/07/21.