Indexação e Tratamento de Dados Heterogêneos: Variedade

Mecanismos de Busca

Conteúdo do Bloco

Indexação

Recuperação de Informação

Solr

Elasticsearch

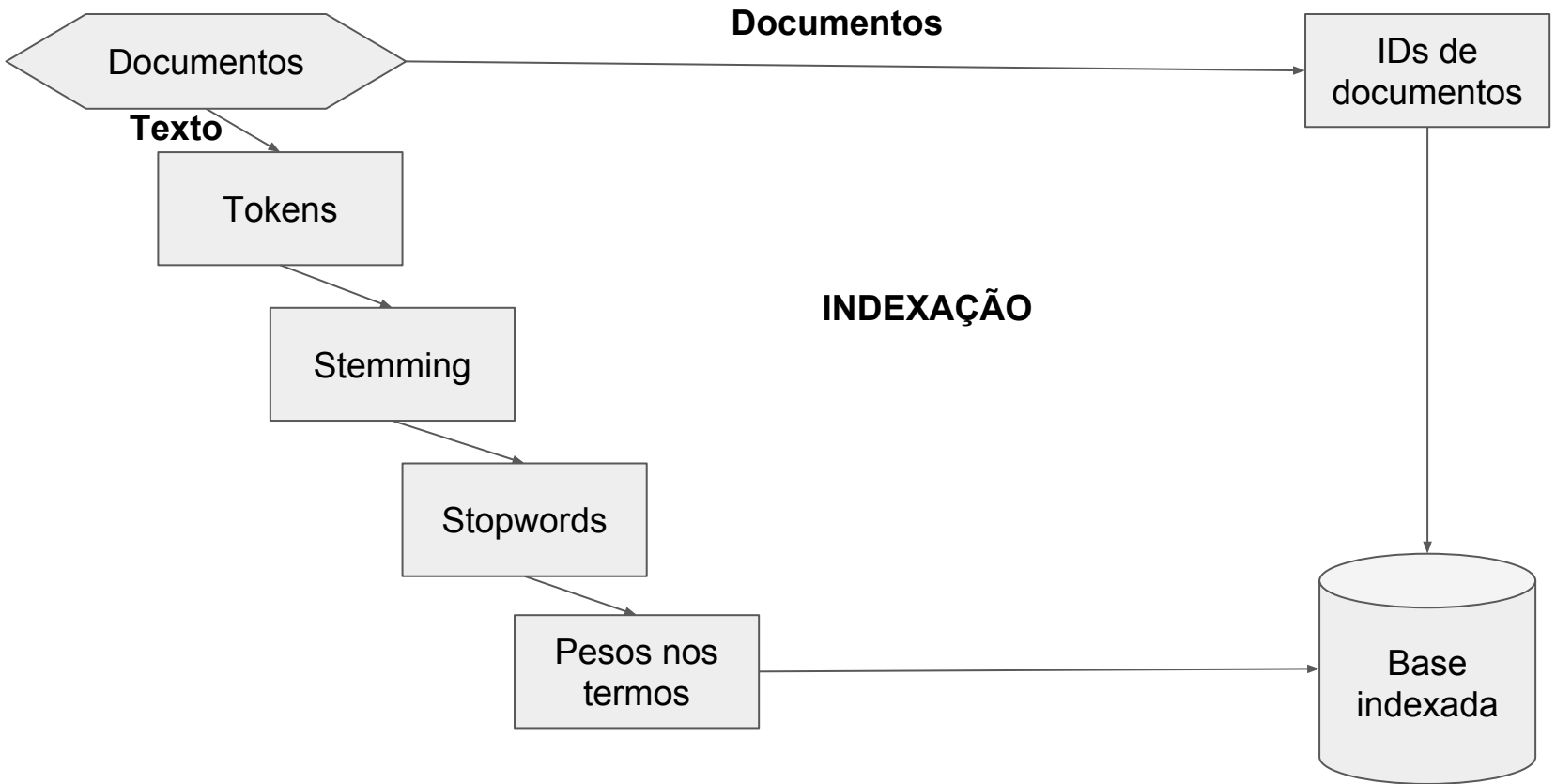
Aula 2

Métricas em information retrieval

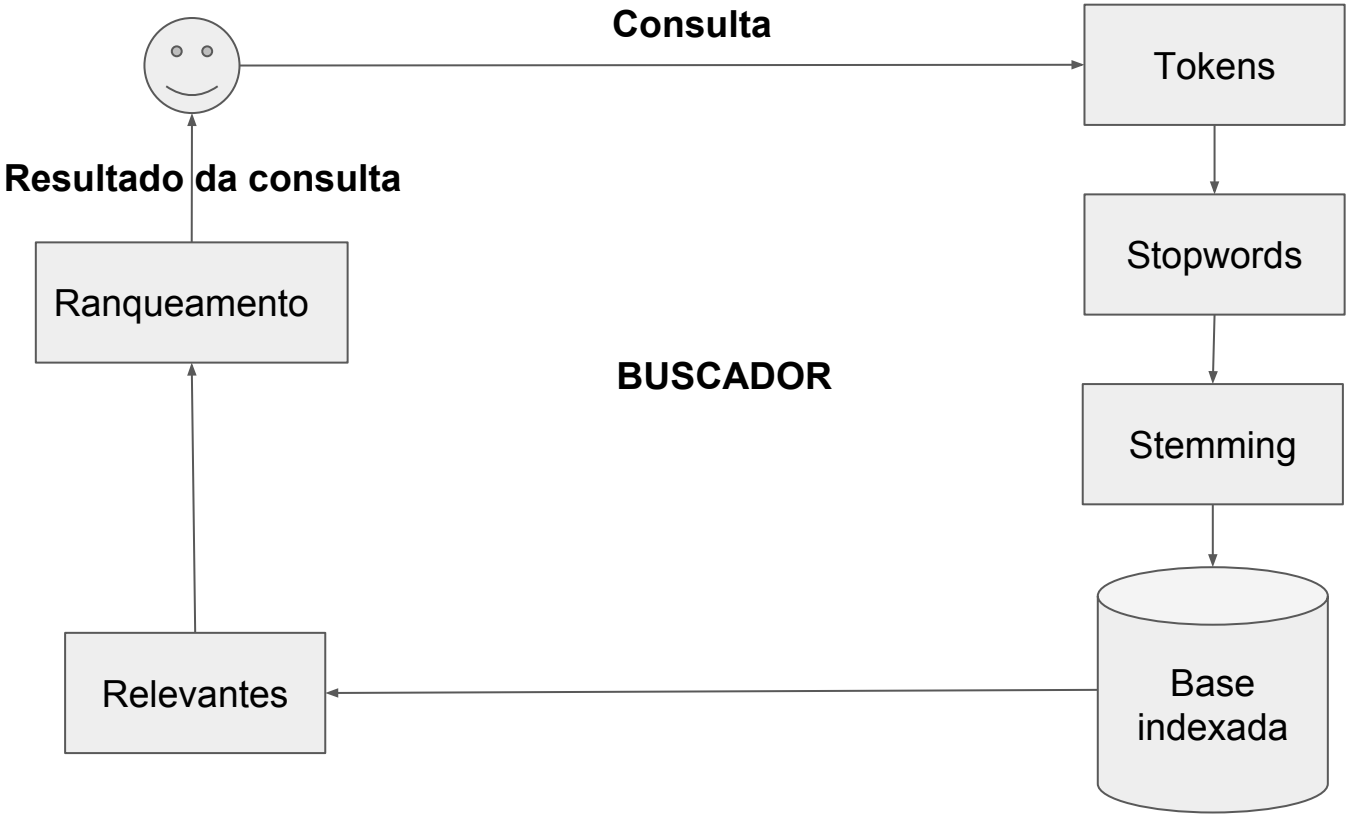
Um pouco mais sobre modelo vetorial

Brincando com NLTK

Relembrando



Relembrando



Relembrando

Similaridade entre consulta e documento no modelo vetorial

- Cosseno

Algoritmos de similaridade diferentes interferem no resultado final

- Existem outras métricas de similaridade?

Outras técnicas de similaridade

Coeficiente de Jaccard $s_{AB} = \frac{|A \cap B|}{|A \cup B|}$

Qual a distância entre dois objetos? Qual a similaridade entre eles?

- D1 = {aula, mba, bigdata, infnet}
- D2 = {aula, analise, sistemas, infnet}
  - S(d1, d2) = {aula, infnet} / {aula, mba, bigdata, analise, sistemas, infnet} = ⅓
  - Coeficiente de Jaccard = ⅓ (similaridade)
  - Distância de Jaccard = ⅔ (não similaridade)

Coeficiente de Pearson

Distância Euclidiana

Métricas

Como saber se um sistema de busca e recuperação da informação está se comportando bem?

- Quão preciso é o meu buscador?
- Quantos documentos relevantes meu buscador exibe como resultado?
- Existe uma hierarquia de relevância dentro dos documentos relevantes, o buscador exibe o resultado na ordem correta?

Métricas

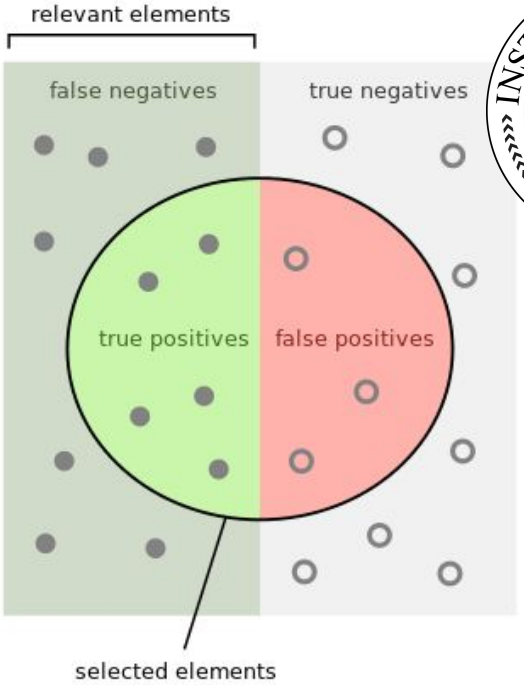
Precisão e revocação (precision and recall)

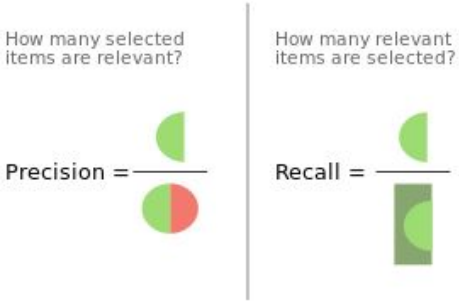
- Precision = relevantes / todos documentos
  - Quão útil foi o resultado?
- Recall = relevantes / todos relevantes
  - Quão completo foi o resultado?

Precision @ K

- K documentos relevantes na primeira página de resultado da consulta

F1 score = 2 * ((precision+recall) / (precision*recall))

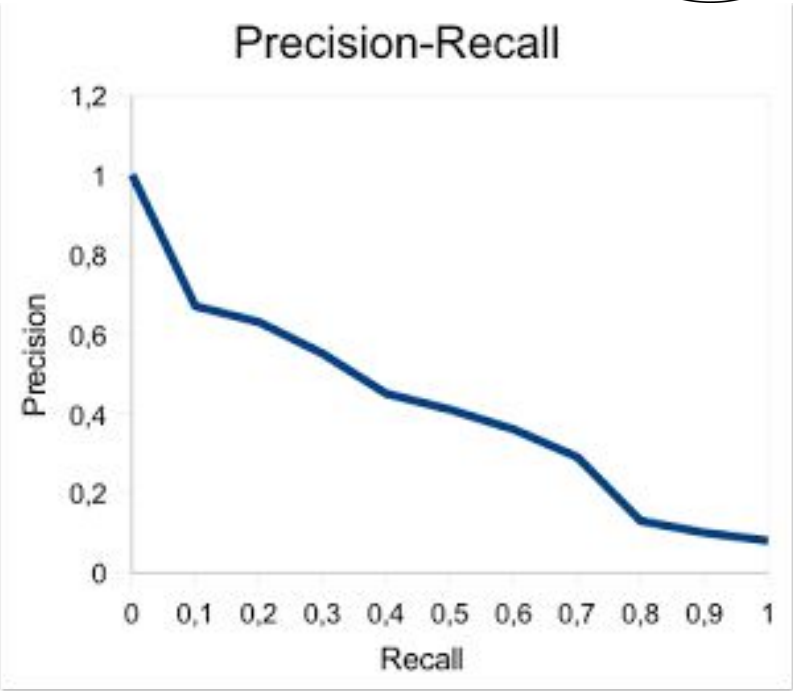




Métricas

Interpolação de 11 pontos de recall e precision

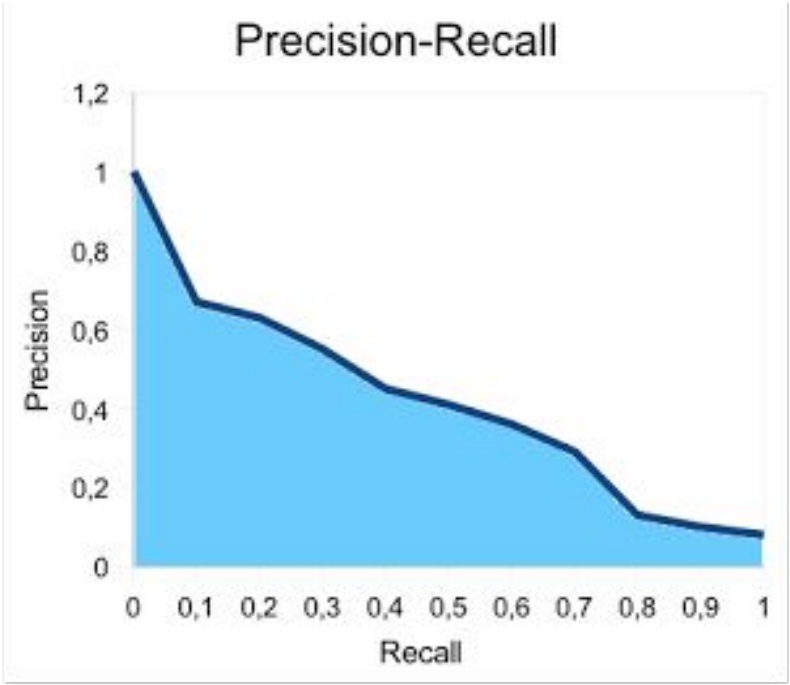
- Precision e recall são inversamente proporcionais
- Quanto mais documentos retornados: Maior fica o recall e menor fica o precision



Métricas

Mean Average Precision (MAP)

- Precisão média de cada consulta / total de consultas



Métricas

DCG (Discounted Cumulative Gain)

- Ganho cumulativo descontado
  - Páginas relevantes que aparecem com ranqueamento baixo são penalizadas

Rank	Relevância	Ganho descontado	DCG
1	2	2/1	2
2	0	0/2	2
3	3	3/3	3
4	2	2/4	3,5

Métricas

NDCG (Normalized DCG)

- NDCG = DCG/IDCG
  - IDCG = Ideal discounted cumulative gain

Rank	Relevância	Ganho descontado	**DCG**	Ganho ideal descontado	**IDCG**	**NDCG**
1	2	2/1	2	3/1	3	0,67
2	0	0/2	2	2/2	4	0,5
3	3	3/3	3	2/3	4,67	0,64
4	2	2/4	3,5	0/4	4,67	0,75

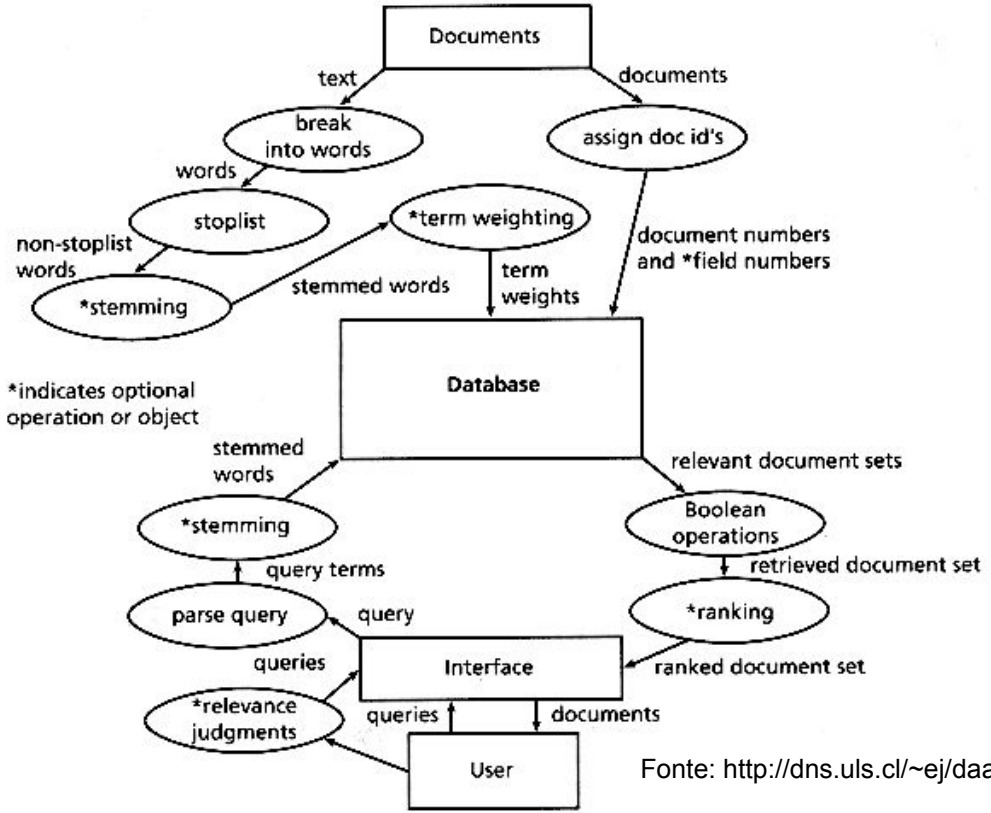
Part of speech tag

Serve para diminuir ambiguidades

Melhora pré-processamento

Word	Tag
heat	verb (noun)
water	noun (verb)
in	prep (noun, adv)
a	det (noun)
large	adj (noun)
vessel	noun

Recuperação da Informação

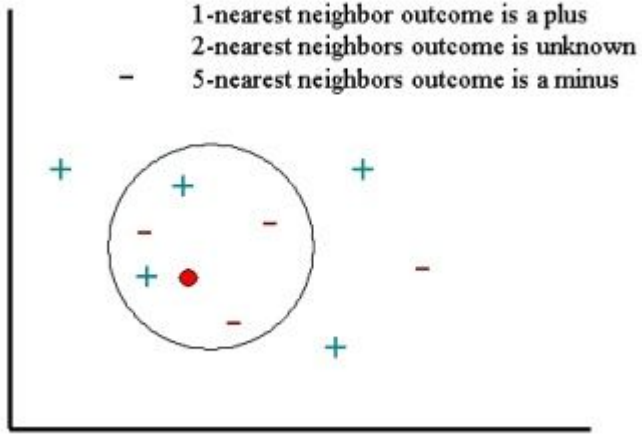


Fonte: http://dns.uls.cl/~ej/daa_08/Algoritmos/books/book5/chap01.htm

Recuperação da Informação

Algoritmos comumente usados em Information Retrieval

- K-NN (K nearest neighbor)
  - Por exemplo, usado para classificar documentos dentro de uma categoria
- Outros métodos serão descritos no decorrer do curso



Fonte: http://www.statsoft.com/textbook/k-nearest-neighbors

Praticando com o NLTK

Vamos testar o NLTK segundo o livro da ferramenta

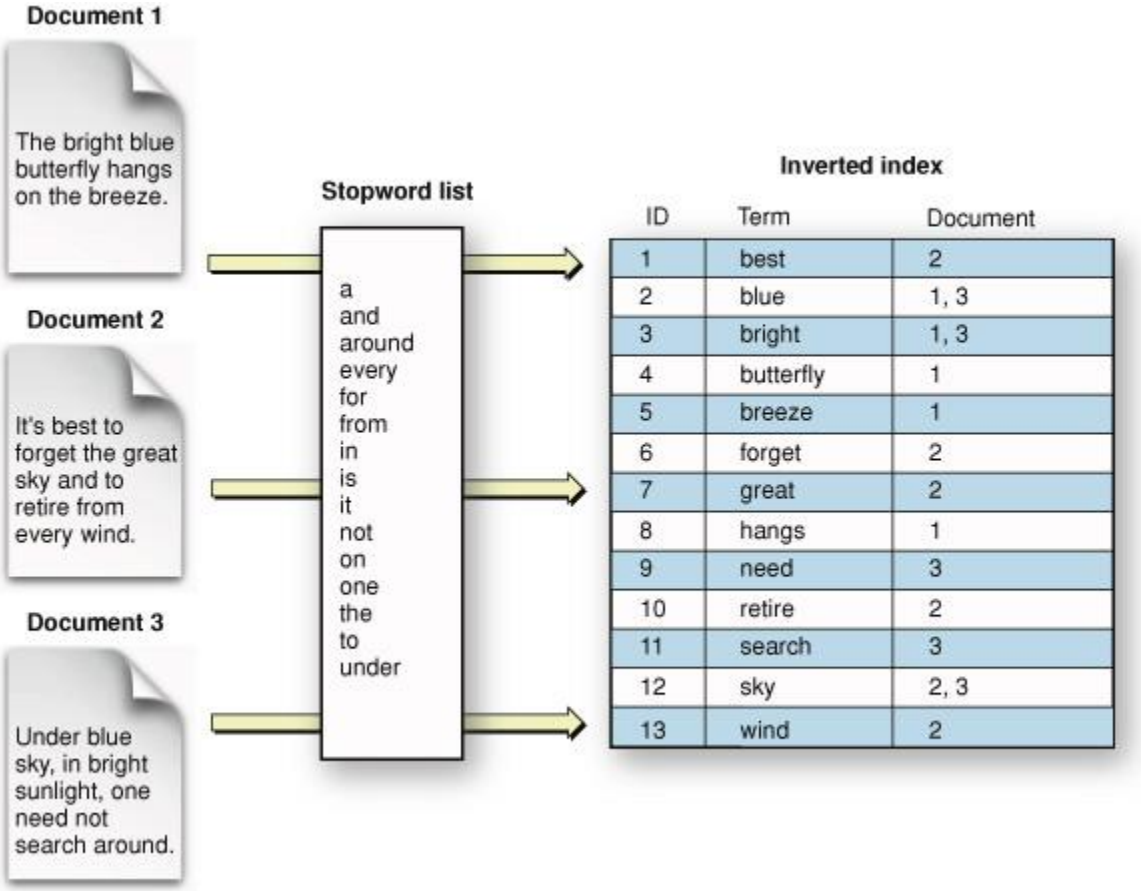
- http://www.nltk.org/book/
  - Capítulo 1
  - Capítulo 3
  - Capítulo 5
- Capítulos acima nortearão a implementação do segundo trabalho semanal
- Praticando enquanto construímos nosso próprio sistema de IR

Exercício

Criar uma lista invertida em python:

- Pegar 3 textos da imagem seguinte e inicializar cada um em um arquivo diferente
- Extrair o texto desses arquivos para dentro do python (3 variáveis ou array)
- Aplicar o tratamento aprendido na aula 1 e praticado na aula 2 para criação dos termos e remoção de stopwords
- Criar a lista de ocorrências de termos dentro dos moldes de uma lista invertida e salvar em arquivo
- Usar a figura seguinte como gabarito (aproximado)

Exercício


Document 1
The bright blue butterfly hangs on the breeze.
Document 2
It's best to forget the great sky and to retire from every wind.
Document 3
Under blue sky, in bright sunlight, one need not search around.
Stopword list
a
and
around
every
for
from
in
is
it
not
on
one
the
to
under
Inverted index
ID
Term
Document
1 best 2
2 blue 1, 3
3 bright 1, 3
4 butterfly 1
5 breeze 1
6 forget 2
7 great 2
8 hangs 1
9 need 3
10 retire 2
11 search 3
12 sky 2, 3
13 wind 2

Exercício

Sugestão para estrutura a ser salva em arquivo

- Chave é o termo, valor é um array de documento/ocorrências desses termos
  - blue;[‘1’, ‘3’]

Enviar arquivo (zipado) com código do exercício

- raul.ferreira@prof.infnet.edu.br
- Colocar no assunto do email “Trabalho 2 - Turma de quinta - Infnet”
- Colocar identificação no corpo do email (nome e sobrenome)

Trabalho individual. Trabalhos copiados = zero

Próxima aula

Recuperação da informação

Mais implementações em sala usando python

Continuar a construção de parte do nosso buscador