Busca Ecommerce

Criação do indice invertido

Ainda sobre o projeto 1

Construir o Wrapper

Criação do índice invertido

Utilização de Stemming e Elemininação de STOPWORDS utilizando a frequência dos termos

Compressão dos postings por intervalo

	Data	Size (KB)
0	List Documents	345
1	Index Compressed	172
2	Index Uncompressed	204

Dificuldade

Classificador não se saiu muito bem em páginas não relevantes que tinham produtos relevantes recomendados no final da página, fazendo-o classificar como relevante

Processamento de Consulta

Dificuldades

```
mock do array invertido
segundo Ramom o indice sera composto por TERMO, DOCUMENTO, FREQUENCIA
index = OrderedDict([
    ('Blusa',[('DOCUMENTO1',2),('DOCUMENTO3',3)]),
    ('Comprida',[('DOCUMENTO3',3),('DOCUMENTO4',1)]),
    ('Longa',[('DOCUMENTO4',1),('DOCUMENTO1',1)]),
    ('Cropped',[('DOCUMENTO4',1),('DOCUMENTO3',4)]),
    ('Azul',[('DOCUMENTO3',2),('DOCUMENTO1',3)]),
    ('Verde',[('DOCUMENTO4',1),('DOCUMENTO3',1)]),
])
```

Como trabalho é separado em 2 partes, foi necessário trabalhar com índice invertido de exemplo. O que causou um problema inicial forte já que não saberíamos ao certo como o dicionário iria ser no final

Dificuldades

usa	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
branca	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0

Construção da matrix

Facilidades

```
def consulting(text):
    My_stopwords = set(stopwords.words('portuguese'))
    palavras = word_tokenize(text)
    lista_Termos = [i for i in palavras if i not in My_stopwords]
    return lista_Termos

# exemplo de consulta do usuario
display(consulting('Blusas Longas de cor branca'))
```

```
['Blusas', 'Longas', 'cor', 'branca']
```

Stopwords rapidamente excluidas

Desenvolvimento

usa	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
branca	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
c	2	1	1	1	1	2	1	1	1	1	...	0	0	0	0	0	1	1	

Consultas por palavras (filtrando desde o começo quais termos da consulta estão disponíveis no índice invertido, buscando os documentos que as palavras estão presentes e construindo uma matriz)

Desenvolvimento

```
]);
index = json.load(open('./compress.json'), object_pairs_hook=OrderedDict)
consultaOriginal = "Blusas brancas e Longa Verde com Azul Azul"
len(index.items())
for term in index.items():
    print(term[0])
```

Carregando o indice invertido comprimido com palavras provindas de descrição, titulo e outros textos do site.

No total temos 1247 palavras únicas

Desenvolvimento

Utilizando Document-at-Time
Rankeando via Modelo de Espaço de Vetores
Com TFIDF Sem TFIDF (consiste em olhar a presença do termo no documento)

Rankeamento sem TFIDF

```
# exemplo da pesquisa sem usar TFIDF
pesquisa = postingTerms("Quero Blusa Comprida bem Longa, Azul com cropped Verde", index)
listaNomes = documentNames(pesquisa,index)
matrix = matrixFrequencia(pesquisa,listaNomes,index)
scoreThis(matrix)
```

13 rows × 330 columns

```
[(122, 1.0),
 (1785, 1.0),
 (1572, 1.0),
 (585, 1.0),
 (1041, 1.0),
 (665, 1.0),
 (1581, 1.0),
 (1163, 1.0),
 (1171, 1.0),
 (86, 1.0),
 (298, 1.0),
 (1746, 1.0),
 (1708, 1.0),
 (758, 1.0),
```

- Resultados mais simples e com menor precisão

Rankeamento com TFIDF

```
# exemplo da pesquisa usando TFIDF
pesquisa = postingTerms("Blusa", index)
listaNomes = documentNames(pesquisa,index)
matrix = matrixFrequenciaTFIDF(pesquisa,listaNomes,index)
scoreThisTDIDF(matrix)
```

```
[('000000010000000', 1.0),
 (460, 0.9733285267845753),
 (1738, 0.9733285267845753),
 (1613, 0.9733285267845753),
 (1611, 0.9733285267845753),
 (1621, 0.9733285267845753),
 (843, 0.9733285267845753),
 (220, 0.9733285267845753),
 (1757, 0.9733285267845753),
 (385, 0.9733285267845753),
 (1705, 0.9733285267845753),
 (991, 0.9733285267845753),
 (7, 0.9733285267845753),
 (1743, 0.9733285267845753),
```

- Como se utiliza da frequência resultados muito mais precisos

Desenvolvimento

8	8	Blusa com Abotoamento Frontal	R 149,00	OFF-WHITE/VERDE	PP,P,M	O shap aliado d...
10	10	Regata com Estampa de Folhas Feminina	R 169,00	LARANJA/MARINHO	PP,P,M,G,GG	A blus folhag repag.
12	12	Top Boxy Jeans com Estampa de Folhas	R 149,00	OFF-WHITE/VERDE	PP,P,M,G,GG	A blus descol
14	14	Body Manga Curta de Malha Canelada	R 149,00	OFF-WHITE/VERDE	PP,P,M,G	A clás aparec de tec
16	16	Blusa Cropped Listrada Feminina	R 169,00	OFF-WHITE	PP,P,M,G,GG	A blus silk me
18	18	Camisa Manga Curta Jeans Masculina	R 149,00	Amarelo	PP,P,M,G	A insp marca blusa

- Carrego um documento com todas informações de um produto, bem como seu link e seus indices para poder criar um frontend de consulta.

Consultas finais

```
# exemplo da pesquisa usando TFIDF
pesquisa = postingTerms("Decote", index)
listaNomes = documentNames(pesquisa,index)
matrix = matrixFrequenciaTFIDF(pesquisa,listaNomes,index)
hello = scoreThisTDIDF(matrix)
# print(hello)
i = 1
contador = 0
ranking = []
while i < 20:
    valor = hello[i][0]
    if not pd.isnull(df.loc[valor].title):
        contador = contador+1
        print("Resposta ",contador,df.loc[valor].title,"Url ",df.loc[valor].url)
    i=i+1
```

```
Resposta  1 Vestido Longo Fenda e Amarraçã... Url  https://www.bellaseda.online/departamento/6722/01/blusa?idpage=2&viewtype=M&ordem=A&nrrows=12&filtros=136768
Resposta  2 Blusa com Estampa Floral e Amarração Url  https://www.damyller.com.br/blusa-com-estampa-floral-e-amarracao-9v0hf06/p
Resposta  3 BODY AMARRAÇÃO ANALICE Url  https://www.boutiquedassi.com.br/BODY-AMARRACAO-ANALICE
Resposta  4 Body Basic Canelado Off Url  https://www.achados96.com.br/bodies/body-basic-canelado-off/
Resposta  5 Cropped Básico Ju Markan Brand Url  https://www.lojavillevie.com.br/ju-markan/cropped-basico-ju-markan-brand
Resposta  6 BLUSA OMBRO PRINCESA AMÉLIA Url  https://www.boutiquedassi.com.br/BLUSA-OMBRO-PRINCESA-AMELIA
Resposta  7 Blusa Crepe Color Block Alças com Fivelas  Marinho Url  https://www.bluk.com.br/blusa-crepe-color-block-alcas-com-fivelas--marinho--bt60860719_331/p
Resposta  8 Blusa Malha Listras Frente com Transpasse Branco Url  https://www.bluk.com.br/blu
```

- Foi feito uma mini-interface para consultas
- Você escreve a consulta e ele filtra os termos busca na matrix os documentos e faz o rankeamento.
- Após isso ele da um match do documentos ranqueados com a tabela do produtos.
- Mostra o produto e links no resultado pesquisa
- 20 Melhores respostas segundo ranking TFIDF são mostradas