Veegle

Recuperação de Informação

Crawler

Crawler

- 9 domínios capturados
- +1000 páginas em 6 domínios
- 3 domínios esgotados

- http://www.tudogostoso.com.br/
- http://gordelicias.biz/
- http://receitasdecomidas.com.br/
- https://www.tudoreceitas.com/
- http://presuntovegetariano.com.br/
- http://receitasdeminuto.com/
- https://www.receitasdemae.com.br/
- http://receitas.ig.com.br/
- https://comidasebebidas.uol.com.br/

Crawler

- Busca em largura
- Busca guiada e horizontal (Heurística)
- Busca sem emular javascript x Busca emulando javascript
- Busca em paralelo
  - 1 Thread para cada instância do crawler
- 2 perguntas para o Harvest Ratio
  - É uma receita?
  - É vegana?

Crawler

- Harvest Ratio BFS | É receita?
  - 67% | http://www.tudogostoso.com.br/
  - 58% | https://www.tudoreceitas.com/
  - 88% | http://receitasdecomidas.com.br/
  - 33% | http://gordelicias.biz/
  - 95% | https://www.receitasdemae.com.br/
  - 81% | http://receitas.ig.com.br/
  - 70% | http://receitasdeminuto.com/
  - 86% | https://comidasebebidas.uol.com.br/
  - 71% | http://presuntovegetariano.com.br/

Crawler

- Harvest Ratio Heurística | É receita?
  - 100% | http://www.tudogostoso.com.br/
  - 99% | https://www.tudoreceitas.com/
  - 100% | http://receitasdecomidas.com.br/
  - 100% | http://gordelicias.biz/
  - 100% | https://www.receitasdemae.com.br/
  - 100% | http://receitas.ig.com.br/
  - 99% | http://receitasdeminuto.com/
  - 100% | https://comidasebebidas.uol.com.br/
  - 96% | http://presuntovegetariano.com.br/

Crawler

- Harvest Ratio | É receita vegana?
  - 45% | http://www.tudogostoso.com.br/
  - 17% | https://www.tudoreceitas.com/
  - 9% | http://receitasdecomidas.com.br/
  - 8% | http://gordelicias.biz/
  - 4% | https://www.receitasdemae.com.br/
  - 4% | http://receitas.ig.com.br/
  - 2% | http://receitasdeminuto.com/
  - 17% | https://comidasebebidas.uol.com.br/
  - 100% | http://presuntovegetariano.com.br/

Crawler

- Estrutura do crawler dividido em:
  - Crawler
    - Requests Handler que gerencia as requisições com requests ou phantomjs
    - Parser Handler que passa o resultado do html para o parser possibilitando query

Crawler

- Crawler genérico para cada site
- Parâmetros
  - Nome de identificação
  - Url do domínio
  - Nome do agente
  - Função de próxima url
  - Função de extração de âncoras
  - Função de nomeação de página salva

Classifier

Pré Processamento

- Remoção do html através do beautiful soup
- Extração de texto das seguintes tags
  - p
  - title
  - h1 -> h6
- Remoção de números
- Testes com remoção de stopwords e stemming

Feature Extraction

- Pipeline
- Bag of Words (CountVectorizer())
- TF-IDF (TfidfTransformer)

Classificadores escolhidos

- SGClassifier -> Stochastic Gradient Descent
- MultinomialNB -> Naive Bayes
- DecisionTreeClassifier -> Árvore de Decisão
- LogisticRegression -> Logistic Regression
- MLPClassifier -> Multilayer Perceptron Classifier
- SVC -> Support Vector Classification

Treinamento

- Divisão do conjunto de treinamento em treino (70%) e teste (30%), aleatório
- 15 repetições do procedimento
- Medidas de desempenho:
- Accuracy
- Precision
- Recall
- Tempo de treinamento

Acurácia

	SGC	Naive Bayes	DecisionTree	LogisticRegression	MLP	SVC
0	0.771186	0.847458	0.711864	0.796610	0.838983	0.762712
1	0.762712	0.771186	0.737288	0.762712	0.779661	0.754237
2	0.669492	0.728814	0.652542	0.677966	0.703390	0.652542
3	0.661017	0.754237	0.720339	0.661017	0.779661	0.644068
4	0.728814	0.796610	0.652542	0.771186	0.788136	0.737288
5	0.754237	0.779661	0.652542	0.754237	0.754237	0.720339
6	0.669492	0.728814	0.618644	0.694915	0.762712	0.644068
7	0.677966	0.728814	0.584746	0.686441	0.737288	0.661017
8	0.711864	0.711864	0.711864	0.711864	0.737288	0.669492
9	0.694915	0.805085	0.652542	0.703390	0.771186	0.669492
10	0.669492	0.754237	0.703390	0.677966	0.771186	0.644068
11	0.576271	0.771186	0.694915	0.644068	0.728814	0.576271
12	0.720339	0.728814	0.677966	0.720339	0.762712	0.686441
13	0.677966	0.694915	0.669492	0.694915	0.779661	0.661017
14	0.669492	0.720339	0.669492	0.669492	0.703390	0.652542

Acurácia - Média

SGC	0.694350
Naive Bayes	0.754802
DecisionTree	0.674011
LogisticRegression	0.708475
MLP	0.759887
SVC	0.675706

Tempo

	SGC	Naive Bayes	DecisionTree	LogisticRegression	MLP	SVC
0	0.081759	0.286944	0.093660	0.063092	7.116157	0.163318
1	0.069931	0.275889	0.112706	0.062579	5.925161	0.155812
2	0.060302	0.294048	0.112717	0.062450	6.847748	0.152462
3	0.058402	0.276021	0.103189	0.060779	5.896733	0.145871
4	0.058971	0.289460	0.099581	0.062125	7.554471	0.158279
5	0.061379	0.277892	0.096056	0.063656	7.479907	0.163200
6	0.060149	0.281625	0.099888	0.064145	5.698795	0.150875
7	0.058135	0.272419	0.095381	0.060957	5.756792	0.147292
8	0.061413	0.295742	0.124370	0.064232	5.909900	0.158156
9	0.060178	0.294827	0.098886	0.062536	6.645585	0.151836
10	0.060995	0.287066	0.100898	0.063138	5.189716	0.152113
11	0.060481	0.296326	0.102607	0.062735	5.468317	0.146322
12	0.060855	0.281965	0.116509	0.062435	6.151085	0.155290
13	0.058666	0.278347	0.089292	0.061252	5.875048	0.150271
14	0.059118	0.278324	0.107223	0.084239	6.162345	0.152010

Tempo - Média

SGC	0.062049
Naive Bayes	0.284460
DecisionTree	0.103531
LogisticRegression	0.064023
MLP	6.245184
SVC	0.153540

Precisão

	SGC	Naive Bayes	DecisionTree	LogisticRegression	MLP	SVC
0	0.733394	0.856351	0.740664	0.780508	0.841048	0.762712
1	0.724702	0.785578	0.759916	0.728540	0.770213	0.754237
2	0.649545	0.723520	0.633475	0.661278	0.689487	0.652542
3	0.642938	0.750912	0.718903	0.642938	0.777889	0.644068
4	0.671418	0.796610	0.684566	0.750210	0.782103	0.737288
5	0.742687	0.794085	0.710678	0.742687	0.747843	0.720339
6	0.683616	0.726152	0.616628	0.700130	0.760734	0.644068
7	0.674536	0.720873	0.597400	0.694719	0.727139	0.661017
8	0.730006	0.705096	0.702257	0.707022	0.728148	0.669492
9	0.687140	0.809609	0.641712	0.702480	0.763906	0.669492
10	0.668636	0.750912	0.701857	0.674768	0.766388	0.644068
11	0.576271	0.772274	0.697004	0.711918	0.739789	0.576271
12	0.801293	0.737532	0.673334	0.728159	0.753045	0.686441
13	0.674536	0.680342	0.671534	0.710940	0.794016	0.661017
14	0.649545	0.711781	0.667639	0.645920	0.689487	0.652542

Precisão - Média

```
print(np.mean(precisao_pd))
```

```
SGC                   0.687351
Naive Bayes           0.754775
DecisionTree          0.681171
LogisticRegression    0.705481
MLP                   0.755416
SVC                   0.675706
dtype: float64
```

Recall

	SGC	Naive Bayes	DecisionTree	LogisticRegression	MLP	SVC
0	0.771186	0.847458	0.711864	0.796610	0.838983	1.0
1	0.762712	0.771186	0.737288	0.762712	0.779661	1.0
2	0.669492	0.728814	0.652542	0.677966	0.703390	1.0
3	0.661017	0.754237	0.720339	0.661017	0.779661	1.0
4	0.728814	0.796610	0.652542	0.771186	0.788136	1.0
5	0.754237	0.779661	0.652542	0.754237	0.754237	1.0
6	0.669492	0.728814	0.618644	0.694915	0.762712	1.0
7	0.677966	0.728814	0.584746	0.686441	0.737288	1.0
8	0.711864	0.711864	0.711864	0.711864	0.737288	1.0
9	0.694915	0.805085	0.652542	0.703390	0.771186	1.0
10	0.669492	0.754237	0.703390	0.677966	0.771186	1.0
11	1.000000	0.771186	0.694915	0.644068	0.728814	1.0
12	0.720339	0.728814	0.677966	0.720339	0.762712	1.0
13	0.677966	0.694915	0.669492	0.694915	0.779661	1.0
14	0.669492	0.720339	0.669492	0.669492	0.703390	1.0

Recall- Média

SGC	0.722599
Naive Bayes	0.754802
DecisionTree	0.674011
LogisticRegression	0.708475
MLP	0.759887
SVC	1.000000

Information Gain

- Using with max_df of 0.9
- Selecionando as 1000 palavras mais importantes com CountVectorizer

```
0.03442397499771957: 'receita',
0.034534954020725747: 'uma',
0.034569969048912005: 'queijo',
0.034644649472241829: 'tudogostoso',
0.035100474715353339: 'forno',
0.035157100572908204: 'algum',
0.035477915258373703: 'página',
0.036327348800868006: 'através',
0.039000517226016419: 'que',
0.039309127866855462: 'sem',
0.03936400850027888: 'manteiga',
0.039958669438062866: 'almoço',
0.041814951477100973: 'na',
0.042516604620792098: 'veganas',
0.048125647672915203: 'para',
0.055247084348922892: 'frango',
0.060987463543420828: 'com',
0.061658916104465646: 'receitas',
0.074096784975157148: 'vegano'}
```

Acurácia com Information Gain

	SGC	Naive Bayes	DecisionTree	LogisticRegression	MLP	SVC
0	0.677966	0.677966	0.669492	0.694915	0.711864	0.635593
1	0.635593	0.720339	0.601695	0.677966	0.694915	0.661017
2	0.661017	0.745763	0.661017	0.669492	0.728814	0.652542
3	0.686441	0.728814	0.618644	0.703390	0.754237	0.644068
4	0.737288	0.779661	0.720339	0.745763	0.805085	0.703390
5	0.669492	0.711864	0.686441	0.669492	0.720339	0.652542
6	0.703390	0.711864	0.644068	0.711864	0.737288	0.669492
7	0.677966	0.703390	0.652542	0.720339	0.694915	0.669492
8	0.677966	0.745763	0.703390	0.694915	0.805085	0.661017
9	0.644068	0.796610	0.627119	0.652542	0.805085	0.593220
10	0.737288	0.805085	0.661017	0.762712	0.830508	0.711864
11	0.745763	0.728814	0.720339	0.762712	0.762712	0.703390
12	0.745763	0.754237	0.694915	0.754237	0.745763	0.703390
13	0.677966	0.728814	0.576271	0.677966	0.711864	0.652542
14	0.669492	0.771186	0.644068	0.703390	0.737288	0.669492

Acurácia com Information Gain - Média

SGC	0.689831
Naive Bayes	0.740678
DecisionTree	0.658757
LogisticRegression	0.706780
MLP	0.749718
SVC	0.665537

Tempo com Information Gain

	SGC	Naive Bayes	DecisionTree	LogisticRegression	MLP	SVC
0	0.085007	0.279718	0.118544	0.064304	6.070609	0.153494
1	0.060738	0.295795	0.101674	0.061634	6.061091	0.149521
2	0.059413	0.276343	0.090178	0.061493	5.534863	0.149536
3	0.060333	0.281294	0.092060	0.061951	6.006109	0.149856
4	0.061603	0.288482	0.128693	0.063151	6.015458	0.157222
5	0.059606	0.274176	0.103018	0.061838	5.930658	0.151713
6	0.059107	0.280988	0.095521	0.061366	5.885871	0.148424
7	0.060083	0.288757	0.110696	0.062036	6.113147	0.152964
8	0.059129	0.271849	0.083115	0.060814	5.717072	0.146601
9	0.059438	0.287933	0.099498	0.061567	5.750058	0.145534
10	0.059327	0.294890	0.119408	0.061920	5.769093	0.156012
11	0.060816	0.294151	0.098184	0.062525	5.751090	0.155607
12	0.060047	0.284466	0.098328	0.063826	5.727402	0.155469
13	0.058846	0.274780	0.118125	0.061512	6.353389	0.149281
14	0.059896	0.285499	0.104361	0.061964	6.055654	0.152733

Tempo com Information Gain - Média

SGC	0.061559
Naive Bayes	0.283941
DecisionTree	0.104093
LogisticRegression	0.062127
MLP	5.916104
SVC	0.151598

Extractor

Extractor

- Abordagem individual
  - Localização das informações pré-identificadas por análise
  - Extração guiada
- Abordagem genérica
  - Busca em DFS por expressão regular
  - Extração não guiada

- Resultados
  - N = 1366
  - C e E = 1260
  - Recall de 92%
  - Precisão de 100%
  - F-Measure de 96%

Indexador

Ranqueador

Ranqueador

- Tratamento da consulta
  - Conversão para lower case
  - Remoção de stopwords e pontuação
  - Stemming
- Composição da consulta
  - Todos os campos
  - Name
  - Ingredients
  - Steps
- Leitura do posting
  - Document-at-a-time

Ranqueador

- Modelo vetorial
- Pesos atribuídos
  - Booleano
  - TF-IDF
- Correlação de Ranking
  - Kendal Tau
- Avaliação dos 10 primeiros resultados
- Consultas
  - [‘Coxinha de jaca’, ‘Hambúrguer de lentilha’, ‘suco detox’, batata recheada’, ‘leite de soja’]

Consultas

- Boolean X TF-ID
- 1 -> **0.97** X 0.93
- 2 -> **1.00** X 0.82
- 3 -> **0.99** X 0.90
- 4 -> **0.93** X 0.85
- 5 -> **0.92** X 0.90

Interface

Interface

- Implementação em React.js
- Design similar ao do Google
- Busca simples e avançada

PESQUISAR

Título

Ingredientes

Instruções

PESQUISAR

Veegle

coxinha de jaca,

PESQUISAR

Receita de Coxinha vegana de jaca - passo a passo

Passos a seguir para fazer esta receita: 1 Para preparar esta coxinha sem carne comece por cozinhar a jaca. Para isso coloque-a na panela de pressao, cubra com água e acrescente um fio de azeite ou de óleo (para que ela nao fique muito grudenta). Deixe cozinhar por cerca de 1 hora depois de pegar pressao. No fim desse tempo pique um garfo para conferir que está macia. Dica: Use uma jaca verde, do tipo dura, para evitar o gosto de jaca na coxinha. 2 Quando sua jaca estiver macia,...

Coxinha de Jaca - 06/04/2016 - UOL Universa

Modo de preparo Recheio Corte a jaca em pedaços pequenos, com as maos e a faca umedecidas. Lave-os para tirar um pouco da gosma da jaca. Coloque ospedaços em uma panela de pressao, cobrindo com água. Deixe cozinhar e, quando a panela pegar pressao, deixe mais 15 minutos e desligue. Escorra toda a água e espere esfriar para começar a desfiar. Nao encha muito...

Receita de Coxinha vegana de jaca - passo a passo

Passos a seguir para fazer esta receita: 1 Para preparar esta coxinha sem carne comece por cozinhar a jaca . Para isso coloque-a na panela de pressao, cubra com água e acrescente um fio de azeite ou de óleo (para que ela nao fique muito grudenta). Deixe cozinhar por cerca de 1 hora depois de pegar pressao. No fim desse tempo pique um garfo para conferir que está macia. Dica: ...

Coxinha de "Carne de Jaca" - Presunto Vegetariano

Modo de preparo Antes de tudo, prepare o recheio. Em uma panela em fogo médio, coloque um fio de óleo e refogue o óleo, a cebola e o alho até ficarem levemente dourados. Acrescente a "carne de jaca", tempere com sal e com temperos a gosto, eu usei orégano, cominho, limao e açafrao. Refogue por cerca de 3 minutos, adicione o extrato de tomate, as azeitonas e refogue mais um minutinho. Corrija o sal se necessário e desligue. Acrescente a salsinha e cebolinha picadas, misture...

Coxinha Vegana - Presunto Vegetariano

Modo de preparo Antes de tudo, prepare o recheio. Em uma panela em fogo médio coloque um fio de óleo e refogue o alho e a cebola até ficarem levemente dourados. Acrescente o palmito picado, as azeitonas picadas, tempere com uma pitada de sal e misture. Adicione o extrato de tomate, misture e acrescente o amido de milho misturado com a água e mexa novamente. Adicione a salsinha e a cebolinha, o orégano desidratado, e se gostar tempere com pimenta do reino...

https://veegle-web.herokuapp.com/