Faculdade XPe

RELATÓRIO

PROJETO APLICADO

PÓS-GRADUAÇÃO

XP Educação

Relatório do Projeto Aplicado

Identificação Automática de Notícias Negativas

Hélio Ricardo de Souza Pimentel

Orientador(a): Pedro Guerra

2023

HÉLIO RICARDO DE SOUZA PIMENTEL

XP EDUCAÇÃO

RELATÓRIO DO PROJETO APLICADO

Identificação Automática de Notícias Negativas

Relatório de Projeto Aplicado desenvolvido para fins de conclusão do curso MBA em Machine Learning.

Orientador (a): Pedro Guerra

São Paulo

2023

Sumário

1. CANVAS do Projeto Aplicado

1.1 Desafio

1.1.1 Análise de Contexto

Quando uma empresa segue as melhores práticas de *compliance*, precisa verificar se seus fornecedores têm débito tributário, trabalhista ou previdenciário; denúncias de trabalho escravo, crime ambiental ou improbidade administrativa; acordos de leniência; termos de ajustamento de conduta; restrições em organismos como o Conselho de Segurança das Nações Unidas; etc.

Muitas destas verificações são feitas em sites governamentais: a Certidão de Débitos Relativos a Créditos Tributários Federais e à Dívida Ativa da União, por exemplo, pode ser gerada no site da Receita Federal; o Cadastro de Empregadores que Submeteram Trabalhadores a Condições Análogas à de Escravo está disponível no site do Conselho Nacional de Justiça.

Não existe nem poderia existir, no entanto, um site oficial que garanta a boa reputação de um fornecedor. O que se faz, neste caso, é verificar se o fornecedor foi criticado em alguma reportagem nos principais sites de notícias.

Pesquisar as notícias relacionadas a uma empresa no Google, definindo que os resultados devem ser de um ano determinado é simples. Clicar nos dez primeiros links e ler as dez notícias para classificá-las como positivas ou negativas também. O problema é o custo: num caso real, a ***identificação de notícias negativas*** de quase 650 empresas exigiu a leitura de quase 30.000 reportagens.

O objetivo deste projeto é automatizar completamente este trabalho — tendo como entrada uma planilha Excel com os CNPJs e os nomes das empresas e, como saída, uma planilha Excel com o CNPJ, o ano, o link de uma reportagem e a classificação dela como positiva ou negativa.

Matriz CSD

Certezas	Suposições	Dúvidas
A ***identificação de notícias negativas*** é uma necessidade do *compliance* de várias empresas.	A análise de sentimento automatizada com inteligência artificial terá qualidade semelhante ou superior à que é feita manualmente. As pessoas que fazem manualmente têm pouquíssimo tempo para ler e compreender efetivamente o texto.	Como tratar a identificação errada? (Notícias que podem até ter um teor negativo, mas são positivas para a empresa pesquisada.)
O processamento não precisa ser rápido. O custo é que precisa ser baixo.	A ***identificação de notícias negativas*** poderá ser oferecida em larga escala. Com a automatização, a capacidade de oferecer o serviço não dependerá mais da quantidade de mão de obra disponível.	Qual o melhor algoritmo de inteligência artificial para este problema?
		O modelo de inteligência artificial precisará ser treinado?

Matriz POEMS

Pessoas	Objetos	Ambiente	Mensagem	Serviços
Profissionais que fazem a pesquisa manualmente e digitam as informações numa planilha Excel.	Micro com acesso à internet.	O trabalho manual é feito remotamente.	Planilha com CNPJs e nomes das empresas.	Um sistema externo gera a planilha de entrada e lê a planilha de saída.
			Planilha com CNPJ, ano, link de reportagem e classificação positiva ou negativa.	

1.1.2 Personas

Neste projeto, temos apenas um tipo de persona: alguém tipicamente com ensino médio ou cursando ensino superior para pesquisar as notícias relacionadas a uma empresa no Google e classificá-las como positivas ou negativas.

A geração da planilha de entrada (com os CNPJs e os nomes das empresas) e o processamento da planilha de saída (com o CNPJ, o ano, o link de uma reportagem e a classificação) não estão no escopo deste projeto.

1.1.3 Benefícios e Justificativas

Os benefícios deste projeto são:

a) Redução dos custos;
b) Melhoria da qualidade (a conferir);
c) Aumento das receitas (oferecendo a ***identificação de notícias negativas*** em larga escala).

Os itens A e C são bastante óbvios porque o processo manual será automatizado. O item B será avaliado no momento oportuno.

1.1.4 Hipóteses

As principais hipóteses que temos são as seguintes:

Observação	Hipótese
O trabalho é todo feito manualmente.	A equipe (talvez toda) será substituída pelo novo sistema.
Diferentes profissionais têm desempenhos diferentes na avaliação das notícias, que é sempre subjetiva.	O sistema gerará resultados mais padronizados.
O excesso de trabalho diminui a qualidade do resultado.	O sistema será escalonável.

1.2 Solução

1.2.1 Objetivo SMART

O objetivo SMART é: automatizar todo processo obtendo a mesma qualidade ou superior (para isso, vamos comparar o resultado gerado manualmente com o resultado gerado automaticamente).

Sprint 1	Sprint 2	Sprint 3
Desenvolver um sistema que leia a planilha de entrada e, para cada empresa, baixe 10 notícias de 2014, 10 notícias de 2015, até 10 notícias de 2023.	Testar as soluções de análise de sentimento disponíveis e escolher uma — da limpeza dos dados (como o tratamento das stop words) à classificação de positiva ou negativa.	Validar os resultados.

1.2.2 Premissas e Restrições

Sprint 1	Sprint 2
O programa da Sprint 1 será desenvolvido em C#. A linguagem é boa para a tarefa que será desenvolvida e tenho mais fluência nela do que em Python.	O programa da Sprint 2 será desenvolvido em Python por causa dos recursos necessários, típicos de inteligência artificial.

Risco	Impacto	Ações Preventivas	Ações Corretivas
O processamento da Sprint 1 (com centenas de CNPJs) é lento. Pode demorar alguns dias.	Sem o processamento completo da Sprint 1, os testes da Sprint 2 serão prejudicados.	Começar o processamento o quanto antes, mesmo que seja por etapas.	Usar um número limitado de empresas, se for necessário.
Resultados ruins depois do processamento da Sprint 2.	O processamento automatizado não substituiria o processamento manual.		Fazer o treinamento do modelo.

1.2.3 Backlog de Produto

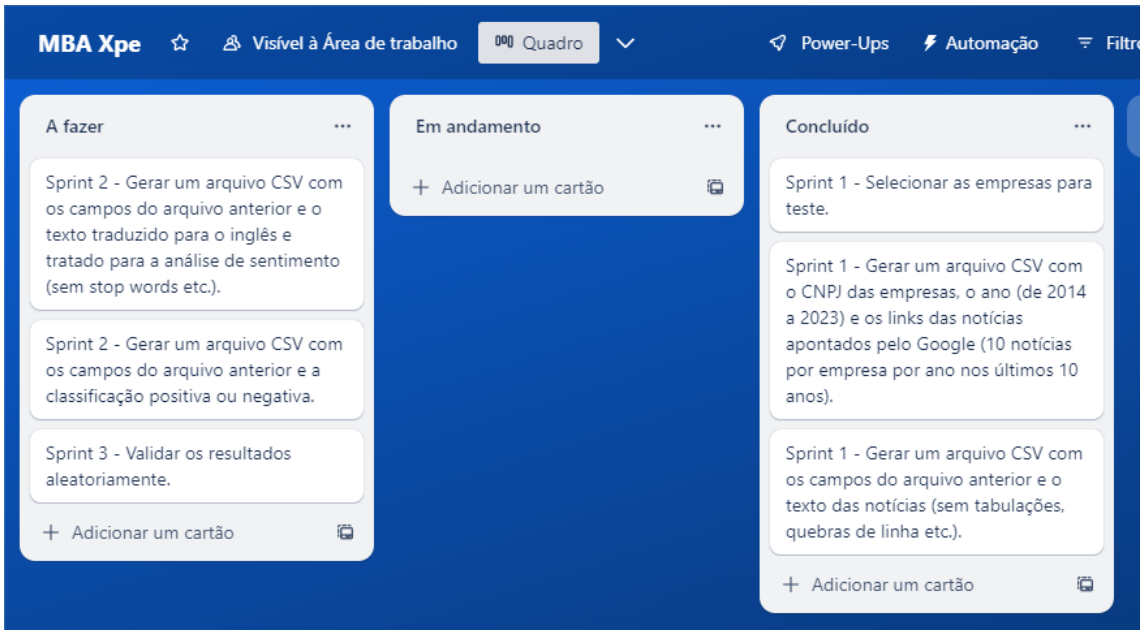
Sprint	Funcionalidade
1	Selecionar as empresas para teste.
1	Gerar um arquivo CSV com o CNPJ das empresas, o ano (de 2014 a 2023) e os links das notícias apontados pelo Google (10 notícias por empresa por ano nos últimos 10 anos).
1	Gerar um arquivo CSV com os campos do arquivo anterior e o texto das notícias (sem tabulações, quebras de linha etc.).
2	Gerar um arquivo CSV com os campos do arquivo anterior e o texto traduzido para o inglês e tratado para a análise de sentimento (sem stop words etc.).
2	Gerar um arquivo CSV com os campos do arquivo anterior e a classificação positiva ou negativa.
3	Validar os resultados aleatoriamente.

2. Área de Experimentação

2.1 Sprint 1

2.1.1 Solução

- Evidência do planejamento:



- Evidência da execução de cada requisito:

1 - Selecionar as empresas para teste.

As empresas selecionadas fazem parte de uma reportagem do portal R7 com as "Empresas Listadas na B3" (https://investidorsardinha.r7.com/geral/todos-os-cnpj-das-empresas-listadas-na-b3-a-bolsa-de-valores-brasileira).

CNPJ	Nome
00.000.000/0001-91	BCO BRASIL S.A.
00.000.208/0001-00	BRB BCO DE BRASILIA S.A.
00.001.180/0001-26	CENTRAIS ELET BRAS S.A. - ELETROBRAS
00.070.698/0001-11	CIA ENERGETICA DE BRASILIA
00.272.185/0001-93	CIMS S.A.
00.336.701/0001-04	TELEC BRASILEIRAS S.A. TELEBRAS
00.359.742/0001-08	ATOM EMPREENDIMENTOS E PARTICIPAÇÕES S.A.
00.383.281/0001-09	BNDES PARTICIPACOES S.A. - BNDESPAR
00.416.968/0001-01	BANCO INTER S.A.
00.622.416/0001-41	STATKRAFT ENERGIAS RENOVAVEIS S.A.

00.743.065/0001-27	LITEL PARTICIPACOES S.A.
00.776.574/0001-56	B2W – COMPANHIA DIGITAL
00.864.214/0001-06	ENERGISA S.A.
00.924.429/0001-75	FERROVIA CENTRO-ATLANTICA S.A.
00.938.574/0001-05	CONC RIO-TERESOPOLIS S.A.
01.027.058/0001-91	CIELO S.A.
01.083.200/0001-18	NEOENERGIA S.A.
01.104.937/0001-70	ELETROBRÁS PARTICIPAÇÕES S.A. – ELETROPAR
01.107.327/0001-20	BBM LOGISTICA S.A.
01.417.222/0001-77	MRS LOGISTICA S.A.
01.545.826/0001-07	GAFISA S.A.
01.548.981/0001-79	INVESTIMENTOS BEMGE S.A.
01.599.101/0001-93	SEQUOIA LOGISTICA E TRANSPORTES S.A
01.838.723/0001-27	BRF S.A.
01.896.779/0001-38	CSU CARDSYSTEM S.A.
01.938.783/0001-11	CIA PARTICIPACOES ALIANCA DA BAHIA
01.957.772/0001-89	SUL 116 PARTICIPACOES S.A.
01.971.614/0001-83	EMPRESA NAC COM REDITO PART S.A.ENCORPAR
02.016.440/0001-62	RGE SUL DISTRIBUIDORA DE ENERGIA S.A.
02.062.747/0001-08	SUDESTE S.A.
02.105.040/0001-23	CIBRASEC – COMPANHIA BRASILEIRA DE SECURITIZACAO
02.149.205/0001-69	PORTO SEGURO S.A.
02.162.616/0001-94	UPTICK PARTICIPACOES S.A.
02.193.750/0001-52	GPC PARTICIPACOES S.A.
02.217.319/0001-07	ALEF S.A.
02.291.077/0001-93	PRODUTORES ENERGET.DE MANSO S.A.- PROMAN
02.302.100/0001-06	EDP SÃO PAULO DISTRIBUIÇÃO DE ENERGIA S.A.
02.302.101/0001-42	EMAE – EMPRESA METROP.AGUAS ENERGIA S.A.
02.318.346/0001-68	OPPORTUNITY ENERGIA E PARTICIPACOES S.A.
02.328.280/0001-97	ELEKTRO REDES S.A.
02.351.144/0001-18	TEGMA GESTAO LOGISTICA S.A.
02.351.877/0001-52	LOCAWEB SERVIÇOS DE INTERNET S.A.
02.357.251/0001-53	LIFEMED INDUSTRIAL EQUIP. DE ART. MÉD. HOSP. S.A.
02.365.069/0001-44	PADTEC HOLDING S.A.
02.387.241/0001-60	RUMO S.A.
02.397.080/0001-96	ITAPEBI GERACAO DE ENERGIA S.A.
02.415.408/0001-50	CONC ROD.OESTE SP VIAOESTE S.A
02.429.144/0001-93	CPFL ENERGIA S.A.
02.451.848/0001-62	CONC SIST ANHANG-BANDEIRANT S.A. AUTOBAN
02.474.103/0001-19	ENGIE BRASIL ENERGIA S.A.
02.502.844/0001-66	RUMO MALHA PAULISTA S.A.
02.509.186/0001-34	TRIÂNGULO DO SOL AUTO-ESTRADAS S.A.
02.509.491/0001-26	CONC ECOVIAS IMIGRANTES S.A.
02.558.115/0001-21	TIM S.A.
02.558.157/0001-62	TELEFÔNICA BRASIL S.A
02.635.522/0001-95	JALLES MACHADO S.A.

02.643.896/0001-52	REAL AI PIC SEC DE CREDITOS IMOBILIARIO S.A.
02.664.042/0001-52	TERMINAL GARAGEM MENEZES CORTES S.A.
02.724.983/0001-34	SANESALTO SANEAMENTO S.A.
02.736.470/0001-43	PATRIA CIA SECURITIZADORA DE CRED IMOB
02.762.113/0001-50	BRAZILIAN FINANCE E REAL ESTATE S.A.
02.762.115/0001-49	MMX MINERACAO E METALICOS S.A.
02.762.121/0001-04	SANTOS BRASIL PARTICIPACOES S.A.
02.762.124/0001-30	BETAPART PARTICIPACOES S.A.
02.773.542/0001-22	RB CAPITAL COMPANHIA DE SECURITIZAÇÃO
02.783.423/0001-50	ALTERE SECURITIZADORA S.A.
02.796.775/0001-40	GAMA PARTICIPACOES S.A.
02.800.026/0001-40	COGNA EDUCAÇÃO S.A.
02.846.056/0001-97	CCR S.A.
02.860.694/0001-62	T4F ENTRETENIMENTO S.A.
02.916.265/0001-60	JBS S.A.
02.932.074/0001-91	HYPERA S.A.
02.950.811/0001-89	PDG REALTY S.A. EMPREEND E PARTICIPACOES
02.992.449/0001-09	PROMPT PARTICIPACOES S.A.
02.998.301/0001-81	RIO PARANAPANEMA ENERGIA S.A.
02.998.611/0001-04	CTEEP – CIA TRANSMISSÃO ENERGIA ELÉTRICA PAULISTA
03.014.553/0001-91	TPI – TRIUNFO PARTICIP. E INVEST. S.A.
03.025.305/0001-46	RODOVIAS DAS COLINAS S.A.
03.220.438/0001-73	EQUATORIAL ENERGIA S.A.
03.303.999/0001-36	DTCOM – DIRECT TO COMPANY S.A.
03.378.521/0001-75	LIGHT S.A.
03.467.321/0001-99	ENERGISA MATO GROSSO-DISTRIBUIDORA DE ENERGIA S/A
03.758.318/0001-24	INVESTIMENTOS E PARTICIP. EM INFRA S.A. – INVEPAR
03.767.538/0001-14	BRAZILIAN SECURITIES CIA SECURITIZACAO
03.795.050/0001-09	TERMOPERNAMBUCO S.A.
03.847.461/0001-92	BRADESPAR S.A.
03.853.896/0001-40	MARFRIG GLOBAL FOODS S.A.
03.953.509/0001-47	CPFL GERACAO DE ENERGIA S.A.
03.983.431/0001-03	EDP – ENERGIAS DO BRASIL S.A.
04.030.182/0001-02	CABINDA PARTICIPACOES S.A.
04.031.213/0001-31	CACONDE PARTICIPACOES S.A.
04.032.433/0001-80	ATMA PARTICIPAÇÕES S.A.
04.065.791/0001-99	SINQIA S.A.
04.128.563/0001-10	AES TIETE ENERGIA SA
04.149.454/0001-80	ECORODOVIAS INFRAESTRUTURA E LOGÍSTICA S.A.
04.172.213/0001-51	CIA PIRATININGA DE FORCA E LUZ
04.423.567/0001-21	ENEVA S.A
04.437.534/0001-30	UNIDAS S.A.
04.752.991/0001-10	BIOMM S.A.
04.821.041/0001-08	METALFRIO SOLUTIONS S.A.
04.895.728/0001-80	EQUATORIAL PARA DISTRIBUIDORA DE ENERGIA S.A.
04.902.979/0001-44	BCO AMAZONIA S.A.

04.913.711/0001-08	BCO ESTADO DO PARA S.A.
04.986.320/0001-13	SER EDUCACIONAL S.A.
05.197.443/0001-38	HAPVIDA PARTICIPACOES E INVESTIMENTOS SA
05.336.882/0001-84	CACHOEIRA PAULISTA TRANSMISSORA ENERGIA S.A.
05.495.546/0001-84	LITELA PARTICIPAÇÕES S.A.
05.721.735/0001-28	WILSON SONS LTD.
05.730.375/0001-20	SMILES FIDELIDADE S.A.
05.799.312/0001-20	TERRA SANTA AGRO S.A.
05.878.397/0001-32	ALIANSCE SONAE SHOPPING CENTERS S.A.
06.047.087/0001-39	REDE DOR SÃO LUIZ S.A
06.137.677/0001-52	BRPR 56 SECURITIZADORA CRED IMOB S.A.
06.164.253/0001-87	GOL LINHAS AEREAS INTELIGENTES S.A.
06.272.793/0001-84	EQUATORIAL MARANHÃO DISTRIBUIDORA DE ENERGIA S.A.
06.626.253/0001-51	EMPREENDIMENTOS PAGUE MENOS S.A.
06.948.969/0001-75	LINX S.A.
06.977.745/0001-91	BR MALLS PARTICIPACOES S.A.
06.977.751/0001-49	BR PROPERTIES S.A.
06.981.180/0001-16	CEMIG DISTRIBUICAO S.A.
06.981.381/0001-13	CTC – CENTRO DE TECNOLOGIA CANAVIEIRA S.A.
07.047.251/0001-70	CIA ENERGETICA DO CEARA – COELCE
07.119.838/0001-48	BRAZIL REALTY CIA SECURIT. CRÉD. IMOBILIÁRIOS
07.206.816/0001-15	M.DIAS BRANCO S.A. IND COM DE ALIMENTOS
07.237.373/0001-20	BCO NORDESTE DO BRASIL S.A.
07.415.333/0001-20	SIMPAR S.A.
07.437.016/0001-05	CINESYSTEM S.A.
07.526.557/0001-00	AMBEV S.A.
07.587.384/0001-30	GAIA SECURITIZADORA S.A.
07.594.978/0001-79	SMARTFIT ESCOLA DE GINÁSTICA E DANÇA S.A.
07.628.528/0001-59	BRASILAGRO – CIA BRAS DE PROP AGRICOLAS
07.689.002/0001-89	EMBRAER S.A.
07.718.269/0001-57	SPRINGS GLOBAL PARTICIPACOES S.A.
07.816.890/0001-53	MULTIPLAN – EMPREEND IMOBILIARIOS S.A.
07.820.907/0001-46	CR2 EMPREENDIMENTOS IMOBILIARIOS S.A.
07.857.093/0001-14	AURA MINERALS INC.
07.857.850/0001-50	GP INVESTMENTS. LTD.
07.859.971/0001-30	TRANSMISSORA ALIANÇA DE ENERGIA ELÉTRICA S.A.
07.882.930/0001-65	MITRE REALTY EMPREENDIMENTOS E PARTICIPAÇÕES S.A.
08.065.557/0001-12	TECNISA S.A.
08.070.508/0001-78	RAIZEN ENERGIA S.A.
08.078.847/0001-09	LPS BRASIL – CONSULTORIA DE IMOVEIS S.A.
08.159.965/0001-33	IGUA SANEAMENTO S.A.
08.294.224/0001-65	JHSF PARTICIPACOES S.A.
08.312.229/0001-73	EZ TEC EMPREEND. E PARTICIPACOES S.A.
08.324.196/0001-81	CIA ENERGETICA DO RIO GDE NORTE – COSERN
08.343.492/0001-20	MRV ENGENHARIA E PARTICIPACOES S.A.
08.364.948/0001-38	ALUPAR INVESTIMENTO S/A

08.402.943/0001-52	GUARARAPES CONFECCOES S.A.
08.439.659/0001-50	CPFL ENERGIAS RENOVÁVEIS S.A.
08.467.115/0001-00	CIA ESTADUAL DE DISTRIB ENER ELET-CEEE-D
08.534.605/0001-74	RENOVA ENERGIA S.A.
08.560.444/0001-93	CIA CELG DE PARTICIPACOES – CELGPAR
08.574.411/0001-00	PRATICA KLIMAQUIP INDUSTRIA E COMERCIO SA
08.613.550/0001-98	BRASIL BROKERS PARTICIPACOES S.A.
08.764.621/0001-53	GENERAL SHOPPING E OUTLETS DO BRASIL S.A.
08.795.211/0001-70	MAESTRO LOCADORA DE VEICULOS S.A.
08.801.621/0001-86	CYRELA COMMERCIAL PROPERT S.A. EMPR PART
08.807.432/0001-10	YDUQS PARTICIPACOES S.A.
08.811.643/0001-27	TRISUL S.A.
08.873.873/0001-10	ECORODOVIAS CONCESSÕES E SERVIÇOS S.A.
08.926.302/0001-05	DOMMO ENERGIA S.A.
09.041.168/0001-10	LOG COMMERCIAL PROPERTIES
09.042.817/0001-05	BEMOBI MOBILE TECH S.A.
09.083.175/0001-84	Mosaico Tecnologia ao Consumidor S.A.
09.112.685/0001-32	OSX BRASIL S.A.
09.114.805/0001-30	OCEANPACT SERVICOS MARITIMOS S.A.
09.149.503/0001-06	OMEGA GERAÇÃO S.A.
09.288.252/0001-32	ANIMA HOLDING S.A.
09.295.063/0001-97	TECHNOS S.A.
09.305.994/0001-29	AZUL S.A.
09.346.601/0001-25	B3 S.A. – BRASIL. BOLSA. BALCÃO
09.347.516/0001-81	ELETROMIDIA S.A.
09.391.823/0001-60	SANTO ANTONIO ENERGIA S.A.
09.538.973/0001-53	PDG COMPANHIA SECURITIZADORA
09.611.768/0001-76	INTER CONSTRUTORA E INCORPORADORA S.A.
10.139.870/0001-08	NEOGRID PARTICIPACOES S.A.
10.215.988/0001-60	CIA LOCAÇÃO DAS AMÉRICAS
10.285.590/0001-08	GRUPO DE MODA SOMA S.A.
10.338.320/0001-00	AFLUENTE TRANSMISSÃO DE ENERGIA ELÉTRICA S/A
10.502.676/0001-37	TERMELÉTRICA PERNAMBUCO III S.A.
10.531.501/0001-58	CONC AUTO RAPOSO TAVARES S.A.
10.629.105/0001-68	PETRO RIO S.A.
10.647.979/0001-48	CONC ROTA DAS BANDEIRAS S.A.
10.678.505/0001-63	CONC RODOVIAS DO TIETÊ S.A.
10.753.164/0001-43	ECO SECURITIZADORA DIREITOS CRED AGRONEGÓCIO S.A.
10.760.260/0001-19	CVC BRASIL OPERADORA E AGÊNCIA DE VIAGENS S.A.
10.835.932/0001-08	CIA ENERGETICA DE PERNAMBUCO – CELPE
10.841.050/0001-55	CONC ROD AYRTON SENNA E CARV PINTO S.A.-ECOPISTAS
10.851.805/0001-00	FLEX GESTÃO DE RELACIONAMENTOS S.A.
11.421.994/0001-36	ORIZON VALORIZACAO DE RESIDUOS S.A.
11.669.021/0001-10	ENAUTA PARTICIPAÇÕES S.A.
11.721.921/0001-60	ALPER CONSULTORIA E CORRETORA DE SEGUROS S.A.
11.725.176/0001-27	BOA VISTA SERVIÇOS S.A.

11.992.680/0001-93	QUALICORP CONSULTORIA E CORRETORA DE SEGUROS S.A.
12.049.631/0001-84	MOURA DUBEUX ENGENHARIA S/A
12.130.744/0001-00	TRUE SECURITIZADORA S.A.
12.139.922/0001-63	OCTANTE SECURITIZADORA S.A.
12.181.987/0001-77	MELNICK DESENVOLVIMENTO IMOBILIÁRIO S.A.
12.261.588/0001-16	POLO CAPITAL SECURITIZADORA S.A
12.320.349/0001-90	OURINVEST SECURITIZADORA SA
12.489.315/0001-23	FERREIRA GOMES ENERGIA S.A.
12.528.708/0001-07	AERIS IND. E COM. DE EQUIP. GERACAO DE ENERGIA S/A
12.648.266/0001-24	AMBIPAR PARTICIPACOES E EMPREENDIMENTOS S/A
12.648.327/0001-53	HIDROVIAS DO BRASIL S.A.
12.919.786/0001-24	TCP TERMINAL DE CONTEINERES DE PARANAGUA SA
13.009.717/0001-46	BCO ESTADO DE SERGIPE S.A. – BANESE
13.217.485/0001-11	CENTAURO S.A.
13.574.594/0001-96	BK BRASIL OPERAÇÃO E ASSESSORIA A RESTAURANTES SA
14.110.585/0001-07	MÉLIUZ S.A.
14.388.334/0001-99	PARANA BCO S.A.
14.776.142/0001-50	WESTWING COMERCIO VAREJISTA S.A.
14.785.152/0001-51	HBR REALTY EMPREENDIMENTOS IMOBILIARIOS S/A
14.807.945/0001-24	SANSUY S.A. INDUSTRIA DE PLASTICOS
14.876.090/0001-93	GAIA AGRO SECURITIZADORA S.A.
14.998.371/0001-19	J. MACEDO S.A.
15.073.274/0001-88	PPLA PARTICIPATIONS LTD.
15.101.405/0001-93	CORREA RIBEIRO S.A. COMERCIO E INDUSTRIA
15.115.504/0001-24	TRONOX PIGMENTOS DO BRASIL S.A.
15.139.629/0001-94	CIA ELETRICIDADE EST. DA BAHIA – COELBA
15.141.799/0001-03	CIA FERRO LIGAS DA BAHIA – FERBASA
15.144.017/0001-90	CIA SEGUROS ALIANCA DA BAHIA
15.413.826/0001-50	ENERGISA MATO GROSSO DO SUL – DIST DE ENERGIA S.A.
15.494.541/0001-90	SALUS INFRAESTRUTURA PORTUARIA SA
15.527.906/0001-36	BIOSEV S.A.
15.578.569/0001-06	CONC DO AEROPORTO INTERNACIONAL DE GUARULHOS S.A.
16.404.287/0001-55	SUZANO S.A.
16.590.234/0001-76	AREZZO INDÚSTRIA E COMÉRCIO S.A.
16.614.075/0001-00	DIRECIONAL ENGENHARIA S.A.
16.670.085/0001-55	LOCALIZA RENT A CAR S.A.
16.811.931/0001-00	ALPHAVILLE S.A.
16.922.038/0001-51	ENJOEI.COM.BR ATIVIDADES DE INTERNET S.A.
17.155.730/0001-64	CIA ENERGETICA DE MINAS GERAIS – CEMIG
17.161.241/0001-15	MINASMAQUINAS S.A.
17.167.396/0001-69	ALFA HOLDINGS S.A.
17.167.412/0001-13	FINANCEIRA ALFA S.A.- CRED FINANC E INVS
17.184.037/0001-10	BCO MERCANTIL DO BRASIL S.A.
17.193.806/0001-46	CONSORCIO ALFA DE ADMINISTRACAO S.A.
17.245.234/0001-00	CIA FIACAO TECIDOS CEDRO CACHOEIRA
17.281.106/0001-03	CIA SANEAMENTO DE MINAS GERAIS-COPASA MG

17.314.329/0001-20	INTERNATIONAL MEAL COMPANY ALIMENTACAO S.A.
17.344.597/0001-94	BB SEGURIDADE PARTICIPAÇÕES S.A.
17.346.997/0001-39	COSAN LOGISTICA S.A.
18.328.118/0001-09	PET CENTER COMERCIO E PARTICIPACOES S.A.
18.494.485/0001-82	PORTO SUDESTE V.M. S.A.
18.593.815/0001-97	PRINER SERVIÇOS INDUSTRIAIS S.A.
19.296.342/0001-29	MGI – MINAS GERAIS PARTICIPAÇÕES S.A.
19.378.769/0001-76	INSTITUTO HERMES PARDINI S.A.
19.526.748/0001-50	CIA INDUSTRIAL CATAGUASES
19.853.511/0001-84	NOTRE DAME INTERMEDICA PARTICIPACOES SA
20.258.278/0001-70	OURO FINO SAUDE ANIMAL PARTICIPACOES S.A.
21.255.567/0001-89	CIA TECIDOS SANTANENSE
21.314.559/0001-66	MOVIDA PARTICIPACOES SA
22.266.175/0001-88	FERTILIZANTES HERINGER S.A.
22.677.520/0001-76	CIA TECIDOS NORTE DE MINAS COTEMINAS
23.373.000/0001-32	VAMOS LOCAÇÃO DE CAMINHÕES. MÁQUINAS E EQUIP. S.A.
24.230.275/0001-80	PLANO & PLANO DESENVOLVIMENTO IMOBILIÁRIO S.A.
24.962.466/0001-36	RUMO MALHA NORTE S.A.
24.990.777/0001-09	GRUPO MATEUS S.A.
25.005.683/0001-09	VERT COMPANHIA SECURITIZADORA
26.462.693/0001-28	LAVVI EMPREENDIMENTOS IMOBILIÁRIOS S.A.
26.659.061/0001-59	MPM CORPÓREOS S.A.
26.735.020/0001-02	FOCUS ENERGIA HOLDING PARTICIPAÇÕES S.A
27.093.558/0001-15	MILLS ESTRUTURAS E SERVIÇOS DE ENGENHARIA S.A.
28.127.603/0001-78	BANESTES S.A. – BCO EST ESPIRITO SANTO
28.152.650/0001-71	EDP ESPIRITO SANTO DISTRIBUIÇÃO DE ENERGIA S.A.
28.195.667/0001-06	BCO ABC BRASIL S.A.
29.780.061/0001-09	SAO CARLOS EMPREEND E PARTICIPACOES S.A.
29.950.060/0001-57	NORTEC QUÍMICA S.A.
29.978.814/0001-87	SUL AMERICA S.A.
30.306.294/0001-45	BCO BTG PACTUAL S.A.
30.540.991/0001-66	HAGA S.A. INDUSTRIA E COMERCIO
31.553.627/0001-01	Mobly S.A.
32.785.497/0001-97	NATURA &CO HOLDING S.A.
33.000.167/0001-01	PETROLEO BRASILEIRO S.A. PETROBRAS
33.014.556/0001-96	LOJAS AMERICANAS S.A.
33.035.536/0001-00	JOAO FORTES ENGENHARIA S.A.
33.040.601/0001-87	MERCANTIL BRASIL FINANC S.A. C.F.I.
33.041.260/0652-90	VIA VAREJO S.A.
33.042.730/0001-04	CIA SIDERURGICA NACIONAL
33.050.071/0001-58	AMPLA ENERGIA E SERVICOS S.A.
33.050.196/0001-88	CIA PAULISTA DE FORCA E LUZ
33.102.476/0001-92	MONTEIRO ARANHA S.A.
33.111.246/0001-90	TECNOSOLO ENGENHARIA S.A.
33.113.309/0001-47	VALID SOLUÇÕES S.A.
33.200.049/0001-47	HOTEIS OTHON S.A.

33.228.024/0001-51	WLM PART. E COMÉRCIO DE MÁQUINAS E VEÍCULOS S.A.
33.256.439/0001-39	ULTRAPAR PARTICIPACOES S.A.
33.376.989/0001-91	IRB – BRASIL RESSEGUROS S.A.
33.386.210/0001-19	SONDOTECNICA ENGENHARIA SOLOS S.A.
33.412.081/0001-96	REFINARIA DE PETROLEOS MANGUINHOS S.A.
33.467.572/0001-34	TEKNO S.A. – INDUSTRIA E COMERCIO
33.592.510/0001-54	VALE S.A.
33.611.500/0001-19	GERDAU S.A.
33.839.910/0001-11	VIVARA PARTICIPAÇOES S.A
33.938.119/0001-69	CIA DISTRIB DE GAS DO RIO DE JANEIRO-CEG
33.958.695/0001-78	UNIPAR CARBOCLORO S.A.
34.169.557/0001-72	BCO MERCANTIL DE INVESTIMENTOS S.A.
34.274.233/0001-02	PETROBRAS DISTRIBUIDORA S/A
35.791.391/0001-94	QUALITY SOFTWARE S.A.
36.542.025/0001-64	BRQ SOLUCOES EM INFORMATICA S.A.
42.150.391/0001-70	BRASKEM S.A.
42.278.291/0001-24	LOG-IN LOGISTICA INTERMODAL S.A.
42.278.473/0001-03	WIZ SOLUÇÕES E CORRETAGEM DE SEGUROS S.A.
42.331.462/0001-31	BATTISTELLA ADM PARTICIPACOES S.A.
42.771.949/0001-35	CENTRO DE IMAGEM DIAGNOSTICOS S.A.
43.185.362/0001-07	IGB ELETRÔNICA S/A
43.470.988/0001-65	EVEN CONSTRUTORA E INCORPORADORA S.A.
43.776.517/0001-80	CIA SANEAMENTO BASICO EST SAO PAULO
45.242.914/0001-05	CEA MODAS S.A.
45.453.214/0001-51	PROFARMA DISTRIB PROD FARMACEUTICOS S.A.
45.987.245/0001-92	BAHEMA EDUCAÇÃO S.A.
47.508.411/0001-56	CIA BRASILEIRA DE DISTRIBUICAO
47.509.120/0001-82	BRADESCO LEASING S.A. ARREND MERCANTIL
47.960.950/0001-21	MAGAZINE LUIZA S.A.
49.263.189/0001-02	HELBOR EMPREENDIMENTOS S.A.
49.669.856/0001-43	RESTOQUE COMÉRCIO E CONFECÇÕES DE ROUPAS S.A.
50.564.053/0001-03	BOMBRIL S.A.
50.746.577/0001-15	COSAN S.A.
50.926.955/0001-42	VULCABRAS/AZALEIA S.A.
51.128.999/0001-90	NUTRIPLANT INDUSTRIA E COMERCIO S.A.
51.218.147/0001-93	IGUATEMI EMPRESA DE SHOPPING CENTERS S.A
51.466.860/0001-56	SAO MARTINHO S.A.
51.928.174/0001-50	PLASCAR PARTICIPACOES INDUSTRIAIS S.A.
52.548.435/0001-79	JSL S.A.
53.113.791/0001-22	TOTVS S.A.
56.643.018/0001-66	EUCATEX S.A. INDUSTRIA E COMERCIO
56.720.428/0001-63	INDUSTRIAS ROMI S.A.
56.992.423/0001-90	BICICLETAS MONARK S.A.
58.119.199/0001-51	ODONTOPREV S.A.
59.105.999/0001-86	WHIRLPOOL S.A.
59.285.411/0001-13	BCO PAN S.A.

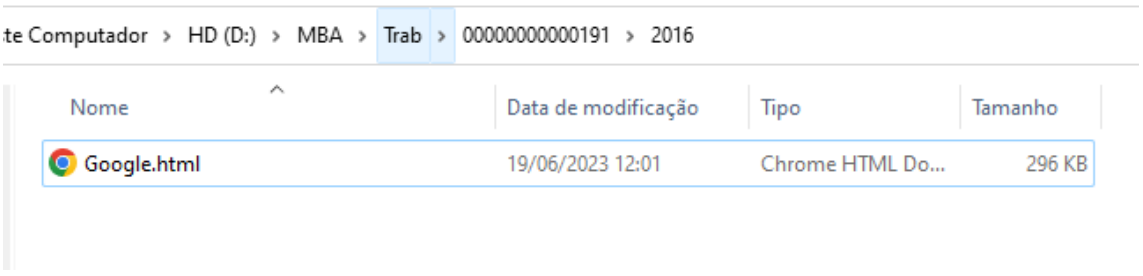
59.418.806/0001-47	TRACK & FIELD CO S.A.
59.789.545/0001-71	POLPAR S.A.
60.398.369/0004-79	PARANAPANEMA S.A.
60.444.437/0001-46	LIGHT SERVICOS DE ELETRICIDADE S.A.
60.476.884/0001-87	MAHLE-METAL LEVE S.A.
60.500.139/0001-26	SARAIVA LIVREIROS S.A. - EM RECUPERAÇÃO JUDICIAL
60.537.263/0001-66	ALLPARK EMPREENDIMENTOS PARTICIPACOES SERVICOS S.A
60.543.816/0001-93	JEREISSATI PARTICIPACOES S.A.
60.637.238/0001-54	INDUSTRIAS J B DUARTE S.A.
60.651.809/0001-05	SUZANO HOLDING S.A.
60.730.348/0001-66	CIA MELHORAMENTOS DE SAO PAULO
60.746.948/0001-12	BCO BRADESCO S.A.
60.770.336/0001-65	BCO ALFA DE INVESTIMENTO S.A.
60.840.055/0001-31	FLEURY S.A.
60.851.615/0001-53	BARDELLA S.A. INDUSTRIAS MECANICAS
60.872.504/0001-23	ITAU UNIBANCO HOLDING S.A.
60.884.319/0001-59	NORDON INDUSTRIAS METALURGICAS S.A.
60.894.730/0001-05	USINAS SID DE MINAS GERAIS S.A.-USIMINAS
60.933.603/0001-78	CESP - CIA ENERGETICA DE SAO PAULO
61.022.042/0001-18	CONSTRUTORA ADOLPHO LINDENBERG S.A.
61.024.352/0001-71	BCO INDUSVAL S.A.
61.065.298/0001-02	MANGELS INDUSTRIAL S.A.
61.065.751/0001-80	ROSSI RESIDENCIAL S.A.
61.079.117/0001-05	ALPARGATAS S.A.
61.082.004/0001-50	MANUFATURA DE BRINQUEDOS ESTRELA S.A.
61.088.894/0001-08	CAMBUCI S.A.
61.092.037/0001-81	ETERNIT S.A.
61.156.113/0001-75	IOCHPE MAXION S.A.
61.156.931/0001-78	SIDERURGICA J. L. ALIPERTI S.A.
61.186.680/0001-74	BANCO BMG S.A.
61.189.288/0001-89	LOJAS MARISA S.A.
61.351.532/0001-68	AZEVEDO E TRAVASSOS S.A.
61.374.161/0001-30	BAUMER S.A.
61.486.650/0001-83	DIAGNOSTICOS DA AMERICA S.A.
61.532.644/0001-15	ITAUSA S.A.
61.584.140/0001-49	REDE ENERGIA PARTICIPAÇÕES S.A.
61.585.865/0001-51	RAIA DROGASIL S.A.
61.695.227/0001-93	ELETROPAULO METROP. ELET. SAO PAULO S.A.
61.856.571/0001-17	CIA GAS DE SAO PAULO - COMGAS
62.002.886/0001-60	SAO PAULO TURISMO S.A.
62.144.175/0001-20	BCO PINE S.A.
62.984.091/0001-02	CRUZEIRO DO SUL EDUCACIONAL S.A.
64.904.295/0001-03	CAMIL ALIMENTOS S.A.
65.654.303/0001-73	DIBENS LEASING S.A. - ARREND.MERCANTIL
67.010.660/0001-24	RNI NEGÓCIOS IMOBILIÁRIOS S.A.
67.260.377/0001-14	MINERVA S.A.

67.571.414/0001-41	VIVER INCORPORADORA E CONSTRUTORA S.A.
71.208.516/0001-74	ALGAR TELECOM S/A
71.476.527/0001-35	CONSTRUTORA TENDA S.A.
71.673.990/0001-77	NATURA COSMETICOS S.A.
73.178.600/0001-18	CYRELA BRAZIL REALTY S.A.EMPREEND E PART
75.315.333/0001-09	ATACADÃO S.A.
75.609.123/0001-23	OURO VERDE LOCACAO E SERVICO S.A.
76.483.817/0001-20	CIA PARANAENSE DE ENERGIA – COPEL
76.484.013/0001-45	CIA SANEAMENTO DO PARANA – SANEPAR
76.535.764/0001-43	OI S.A.
76.627.504/0001-06	INEPAR S.A. INDUSTRIA E CONSTRUCOES
78.876.950/0001-71	CIA HERING
80.227.184/0001-66	METALGRAFICA IGUACU S.A.
67.571.414/0001-41	VIVER INCORPORADORA E CONSTRUTORA S.A.
71.208.516/0001-74	ALGAR TELECOM S/A
71.476.527/0001-35	CONSTRUTORA TENDA S.A.
71.673.990/0001-77	NATURA COSMETICOS S.A.
73.178.600/0001-18	CYRELA BRAZIL REALTY S.A.EMPREEND E PART
81.243.735/0001-48	POSITIVO TECNOLOGIA S.A.
82.508.433/0001-17	CIA CATARINENSE DE AGUAS E SANEAM.-CASAN
82.636.986/0001-55	TEKA-TECELAGEM KUEHNRICH S.A.
82.640.558/0001-04	KARSTEN S.A.
82.643.537/0001-34	ELECTRO ACO ALTONA S.A.
82.901.000/0001-27	INTELBRAS S.A. IND DE TELEC ELETRONICA BRASILEIRA
82.982.075/0001-80	TEXTIL RENAUXVIEW S.A.
83.475.913/0001-91	PBG S/A
83.878.892/0001-55	CENTRAIS ELET DE SANTA CATARINA S.A.
84.429.695/0001-11	WEG S.A.
84.683.374/0001-49	TUPY S.A.
84.683.408/0001-03	DOHLER S.A.
84.683.671/0001-94	WETZEL S.A.
84.693.183/0001-68	SCHULZ S.A.
85.778.074/0001-06	METALURGICA RIOSULENSE S.A.
86.375.425/0001-09	METISA METALURGICA TIMBOENSE S.A.
86.550.951/0001-50	POMIFRUTAS S/A
87.456.562/0001-22	JOSAPAR-JOAQUIM OLIVEIRA S.A. – PARTICIP
87.762.563/0001-03	CIA HABITASUL DE PARTICIPACOES
88.610.126/0001-29	FRAS-LE S.A.
88.610.191/0001-54	MUNDIAL S.A. – PRODUTOS DE CONSUMO
88.611.835/0001-29	MARCOPOLO S.A.
88.613.658/0001-10	PETTENATI S.A. INDUSTRIA TEXTIL
89.086.144/0001-16	RANDON S.A. IMPLEMENTOS E PARTICIPACOES
89.096.457/0001-55	SLC AGRICOLA S.A.
89.463.822/0001-12	LUPATECH S.A.
89.637.490/0001-45	KLABIN S.A.
89.850.341/0001-60	GRENDENE S.A.

90.076.886/0001-40	MINUPAR PARTICIPACOES S.A.
90.400.888/0001-42	BANCO SANTANDER (BRASIL) S.A.
90.441.460/0001-48	UNICASA INDÚSTRIA DE MÓVEIS S.A.
91.333.666/0001-17	RECRUSUL S.A.
91.495.499/0001-00	STARA S.A. – INDÚSTRIA DE IMPLEMENTOS AGRÍCOLAS
91.669.747/0001-92	FINANSINOS S.A.- CREDITO FINANC E INVEST
91.983.056/0001-69	KEPLER WEBER S.A.
92.012.467/0001-70	GRAZZIOTIN S.A.
92.660.570/0001-26	TREVISA INVESTIMENTOS S.A.
92.665.611/0001-77	DIMED S.A. DISTRIBUIDORA DE MEDICAMENTOS
92.690.783/0001-09	METALURGICA GERDAU S.A.
92.693.019/0001-89	PANATLANTICA S.A.
92.702.067/0001-96	BCO ESTADO DO RIO GRANDE DO SUL S.A.
92.715.812/0001-31	CIA ESTADUAL GER.TRANS.ENER.ELET-CEEE-GT
92.749.225/0001-63	HERCULES S.A. FABRICA DE TALHERES
92.754.738/0001-62	LOJAS RENNER S.A.
92.781.335/0001-02	TAURUS ARMAS S.A.
92.791.243/0001-03	IRANI PAPEL E EMBALAGEM S.A.
93.828.986/0001-73	CEMEPE INVESTIMENTOS S.A.
95.426.862/0001-97	EXCELSIOR ALIMENTOS S.A.
96.418.264/0218-02	LOJAS QUERO-QUERO S/A
97.191.902/0001-94	CONSERVAS ODERICH S.A.
97.837.181/0001-47	DEXCO S.A

2 - Gerar um arquivo CSV com o CNPJ das empresas, o ano (de 2014 a 2023) e os links das notícias apontados pelo Google (10 notícias por empresa por ano nos últimos 10 anos).

Na verdade, em vez de gerar o arquivo CSV, foi gerada uma estrutura com um diretório para o CNPJ, um diretório para o ano e um arquivo chamado Google.html com os links (a extensão .html foi usada para facilitar a visualização dos arquivos).




```
00000000000191 2016 BCO BRASIL
https://agenciabrasil.ebc.com.br/economia/noticia/2016-11/bb-fara-reestruturacao-fechando-402-agencias-e-incentivando-aposentadoria
https://g1.globo.com/economia/noticia/2016/11/banco-do-brasil-anuncia-fechamento-de-agencias-e-plano-de-aposentadoria.html
http://www1.folha.uol.com.br/mercado/2016/11/1834120-maioria-das-agencias-fechadas-pelo-bb-esta-em-sp-confira-a-lista-completa.shtml
https://agenciabrasil.ebc.com.br/economia/noticia/2016-10/banco-do-brasil-e-caixa-estao-elevando-taxas-de-juros
https://epocanegocios.globo.com/Empresa/noticia/2016/10/epoca-negocios-banco-do-brasil-e-caixa-ja-tem-juros-mais-altos-que-os-de-bancos-privados.html
https://www.estadao.com.br/economia/bb-e-caixa-ja-tem-juros-mais-altos-que-os-de-bancos-privados/
https://agenciabrasil.ebc.com.br/economia/noticia/2016-12/mais-de-94-mil-funcionarios-aderem-ao-plano-de-aposentadoria-do-bb
https://agenciabrasil.ebc.com.br/economia/noticia/2016-10/apos-31-dias-de-greve-bancarios-retornam-ao-trabalho-hoje
https://agenciabrasil.ebc.com.br/economia/noticia/2016-07/aposentados-e-pensionistas-que-recebem-pelo-banco-do-brasil-ganham-opcao-d
https://g1.globo.com/economia/negocios/noticia/bb-sera-banco-digital-com-ponto-de-atendimento-fisico-diz-vice-presidente.ghtml
```

- Evidência dos resultados:

3 - Gerar um arquivo CSV com os campos do arquivo anterior e o texto das notícias (sem tabulações, quebras de linha etc.).

	A	B	C	D	E	F	G	H	I	J	K
1	Cnpj	Ano	Noticia								
2	191	2016	BB fará reestruturação, com fechamento de agências e incentivo à aposentadoria | Agênci								
3	191	2016	Economia - Banco do Brasil anuncia fechamento de agências e plano de aposentadoriaMEI								
4	191	2016	Maioria das agências fechadas pelo Banco do Brasil está em SP, veja lista - 21/11/2016 - Me								
5	191	2016	Banco do Brasil e Caixa elevam taxas de juros | Agência Brasil MENU EBC Notícias TV Play F								
6	191	2016	Banco do Brasil e Caixa já têm juros mais altos que os de bancos privados - Época Negócios								
7	191	2016	BB e Caixa já têm juros mais altos que os de bancos privados - Estadão Estadão/EconomiaM								
8	191	2016	Mais de 9,4 mil funcionários aderem ao plano de aposentadoria do BB | Agência Brasil ME								
9	191	2016	Após 31 dias de greve, bancários retornam ao trabalho hoje | Agência Brasil MENU EBC No								
10	191	2016	Aposentados e pensionistas que recebem pelo BB têm nova opção de saque | Agência Bra								
11	191	2016	Cliente do BB pode pegar senha para atendimento da agência no celular | Economia | G1 E								
12	191	2017	Ex-presidente do Banco do Brasil é preso na Operação Lava-Jato - Politica - Estado de Mina								
13	191	2017	Quem é Aldemir Bendine, ex-presidente do BB e da Petrobras preso na Lava-Jato | GZHM								
14	191	2017	Ex-presidente da Petrobras e do BB é preso em nova fase da Lava Jato Ex-presidente da Pe								
15	191	2017	Túnel de quadrilha causou rachaduras em sala de cofre do Banco do Brasil em SP | São Pau								
16	191	2017	Vídeo: Veja o túnel de R$ 4 milhões que seria usado para roubar R$ 1 bilhão do Banco do B								
17	191	2017	Quadrilha investiu R$ 4 milhões para cavar túnel que seria usado em roubo a banco, diz de								

Junto com este arquivo, foram postados o arquivo Excel citado acima e um vídeo do programa C# que pesquisa as notícias no Google. O C# tem uma biblioteca chamada HtmlAgilityPack que facilita a “limpeza” do HTML para separar estilos, scripts etc. do texto principal de um site.

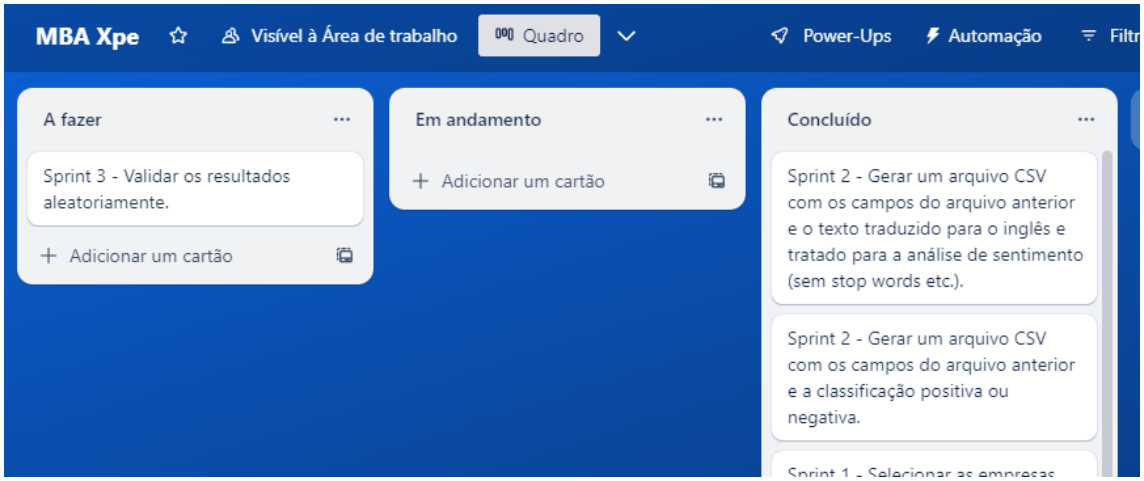
2.1.2 Lições Aprendidas

O nome da empresa é bastante relevante. Seria melhor preparar este dado antes. Exemplo: Em vez de pesquisas BCO BRASIL S.A., BRB BCO DE BRASILIA S.A. ou CENTRAIS ELET BRAS S.A. – ELETROBRAS (nomes abreviados pelo portal R7), seria melhor pesquisar “Banco do Brasil” (com aspas), BRB e Eletrobras.

2.2 Sprint 2

2.2.1 Solução

- Evidência do planejamento:



- Evidência da execução de cada requisito:

1 - Gerar um arquivo CSV com os campos do arquivo anterior e o texto traduzido para o inglês e tratado para a análise de sentimento (sem stop words etc.).

Depois de vários testes, a tradução gratuita se mostrou inviável. O TextBlob / Translate (que estava sendo utilizado) tem limitações quanto ao tamanho do texto e à quantidade de traduções realizadas em determinado período. Estas limitações não são divulgadas em detalhes, mas não foi possivel traduzir milhares de notícias num único dia.

De todo modo, as bibliotecas HtmlAgilityPack (no C#) e NLTK (no Python, com os recursos stopwords e vader_lexicon) funcionaram razoavelmente.

	A	B	C	
1	Cnpj	Ano	Url	Noticia
2	00.000.000/0001-91	2014	https://g1.globo.com/jornal-nacio	Jornal Nacional - Polícia da Itália prende Pizzolato
3	00.000.000/0001-91	2014	https://noticias.r7.com/brasil/pizz	Pizzolato foi preso com passaporte do irmão mort
4	00.000.000/0001-91	2014	https://agenciabrasil.ebc.com.br/i	Polícia italiana diz que Pizzolato foi preso para ext
5	00.000.000/0001-91	2014	https://g1.globo.com/economia/n	G1 - Banco do Brasil e Cielo fazem acordo para trai
6	00.000.000/0001-91	2014	https://epocanegocios.globo.com/	Empresa da Cielo e do Banco do Brasil começa cor
7	00.000.000/0001-91	2014	https://globoesporte.globo.com/v	Banco do Brasil suspende patrocínio CBV por caus
8	00.000.000/0001-91	2014	https://g1.globo.com/economia/n	G1 - Banco do Brasil inaugura 1 agência na China -
9	00.000.000/0001-91	2014	https://brasil.elpais.com/brasil/20	Como esse cara saiu do Brasil sem que ninguém s
10	00.000.000/0001-91	2015	https://agenciabrasil.ebc.com.br/e	Alexandre Abreu substituirá Bendine na presidên
11	00.000.000/0001-91	2015	https://g1.globo.com/economia/n	Economia - Alexandre Abreu é o novo presidente
12	00.000.000/0001-91	2015	https://epocanegocios.globo.com/	Banco do Brasil confirma Alexandre Abreu para pr
13	00.000.000/0001-91	2015	https://g1.globo.com/politica/mer	G1 - Justiça da Itália autoriza extradição de Henriq
14	00.000.000/0001-91	2015	https://g1.globo.com/economia/n	Economia - Concentração aumenta e 5 bancos já d
15	00.000.000/0001-91	2015	https://correiodoestado.com.br/e	Banco do Brasil espera liberar R$ 1 bilhão em finai
16	00.000.000/0001-91	2015	https://g1.globo.com/sao-paulo/n	G1 - Ex-vice-presidente do BB é preso em operaçã
17	00.000.000/0001-91	2015	https://ge.globo.com/volei/noticia	CBV e Banco do Brasil assinam aditivo e patrocínio
18	00.000.000/0001-91	2015	https://g1.globo.com/bom-dia-bra	Bom Dia Brasil - Justiça da Itália decide extraditar

- Evidência dos resultados:

2 - Gerar um arquivo CSV com os campos do arquivo anterior e a classificação positiva ou negativa.

O arquivo foi gerado. A nota foi multiplicada por 100 e arredondada para facilitar a visualização.

Cnpj	Razão Social	Nome Ajustado	Ano	Nota (x10)	Url
00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2014	60	https://g1.globo.com/jornal-nac
00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2014	99	https://noticias.r7.com/brasil/pi
00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2014	100	https://agenciabrasil.ebc.com.b
00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2014	10	https://g1.globo.com/economia
00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2014	0	https://epocanegocios.globo.co
00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2014	-3	https://globoesporte.globo.com
00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2014	59	https://g1.globo.com/economia
00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2014	-57	https://brasil.elpais.com/brasil/
00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2015	100	https://agenciabrasil.ebc.com.b
00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2015	34	https://g1.globo.com/economia
00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2015	0	https://epocanegocios.globo.co
00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2015	91	https://g1.globo.com/politica/m
00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2015	-5	https://g1.globo.com/economia
00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2015	87	https://correiodoestado.com.br
00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2015	-60	https://g1.globo.com/sao-paulo
00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2015	0	https://ge.globo.com/volei/noti

Uma análise preliminar de 50 registros com notas negativas mostrou que quase metade (23) dos apontamentos estavam corretos. Na sprint 3, teremos uma análise mais completa a respeito destes resultados.

OBS: Aprendemos na sprint 1 que o nome da empresa era bastante relevante. Os nomes foram “ajustados” manualmente. Na planilha acima, a cerquilha representa as aspas (que não apareceriam no Excel).

Cnpj	Razão Social	Nome Ajustado	Ano	Nota (x10)	OK	Url
30.306.294/0001-45	BANCO BTG PACTUAL S.A.	#BTG PACTUAL#	2018	-100		https://g1.globo.com/fato-ou-fake/noticia/
92.715.812/0001-31	COMPANHIA ESTADUAL DE TRANSI	#CEEE-T#	2022	-100	OK	https://www.cut.org.br/noticias/rs-oito-me
92.715.812/0001-31	COMPANHIA ESTADUAL DE TRANSI	#CEEE-T#	2022	-100	OK	https://www.cut.org.br/noticias/apos-baixa
01.548.981/0001-79	INVESTIMENTOS BEMGE S/A	#BEMGE#	2014	-99		https://vermelho.org.br/2014/07/29/cristo-
09.114.805/0001-30	OCEANPACT SERVICOS MARITIMOS	#OCEANPACT SERVICOS MARITIMC	2018	-99		https://www.reuters.com/article/us-brazil-
00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2017	-90	OK	https://brasil.elpais.com/brasil/2017/07/27
00.416.968/0001-01	BANCO INTER S.A	#BANCO INTER#	2017	-90		https://tecnoblog.net/noticias/2017/05/05/
06.981.180/0001-16	CEMIG DISTRIBUICAO S.A	#CEMIG#	2016	-90	OK	https://www.em.com.br/app/noticia/gerais
06.981.180/0001-16	CEMIG DISTRIBUICAO S.A	#CEMIG#	2019	-90	OK	https://www.em.com.br/app/noticia/econo
07.689.002/0001-89	EMBRAER S.A.	#EMBRAER#	2015	-90	OK	https://www.em.com.br/app/noticia/inter
01.548.981/0001-79	INVESTIMENTOS BEMGE S/A	#BEMGE#	2016	-80	OK	https://g1.globo.com/minas-gerais/noticia/
03.758.318/0001-24	INVESTIMENTOS E PARTICIPACOES	#INVEPAR#	2015	-80	OK	https://brasil.elpais.com/brasil/2015/11/06
07.857.850/0001-50	GP INVESTMENTS, LTD.	#GP INVESTMENTS#	2018	-80		https://www.em.com.br/app/noticia/econo
13.009.717/0001-46	BANCO DO ESTADO DE SERGIPE S/A	#BANCO DO ESTADO DE SERGIPE#	2016	-80		https://g1.globo.com/se/sergipe/noticia/20
61.088.894/0001-08	CAMBUCI S/A	#CAMBUCI SA#	2020	-80		https://jeonline.com.br/noticia/21194/patr
02.328.280/0001-97	ELEKTRO REDES S.A.	#ELEKTRO#	2020	-70		https://opopularmm.com.br/elektro-reforc
06.981.381/0001-13	CTC - CENTRO DE TECNOLOGIA CAN	#CENTRO DE TECNOLOGIA CANAVI	2018	-70		https://sobancarios.com.br/08/2018/fazend

2.2.2 Lições Aprendidas

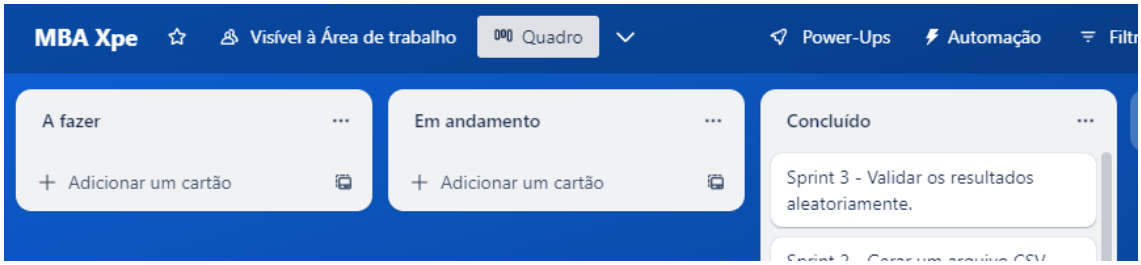
A biblioteca NLTK foi a que trabalhou melhor com os textos relativamente grandes das notícias tanto para o tratamento das stopwords, quanto para a tokenização e para a análise de sentimento propriamente dita sem precisar de treinamento. Outras bibliotecas testadas foram, como TextBlob e Gensim.

Alguns sites que não são de notícias devem ser evitados, como cut.org.br. Na sprint 3, faremos a análise de um processamento restrito aos portais uol.com.br e globo.com (incluindo Folha de S. Paulo, O Antagonista, G1, Época etc.).

2.3 Sprint 3

2.3.1 Solução

- Evidência do planejamento:



- Evidência da execução de cada requisito:

A validação dos resultados foi feita com dois reprocessamentos:

- ⇨ O primeiro restrito aos portais uol.com.br e globo.com (incluindo Folha de S. Paulo, O Antagonista, G1, Época etc.);

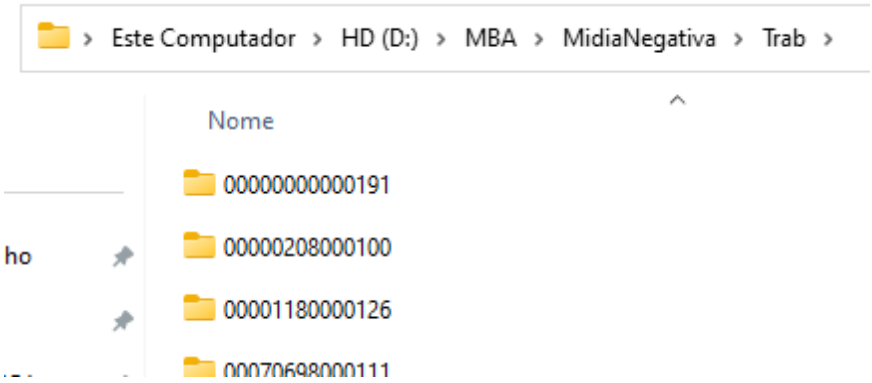
- ⇨ O segundo buscando as palavras “notícia negativa” nos sites uol.com.br, globo.com, bbc.com e dw.com (resultados em português).

O segundo reprocessamento gerou 1/3 das notícias quando comparado com o primeiro (3.259 em vez de 9.599) e quase metade de notícias negativas (258 em vez de 561).

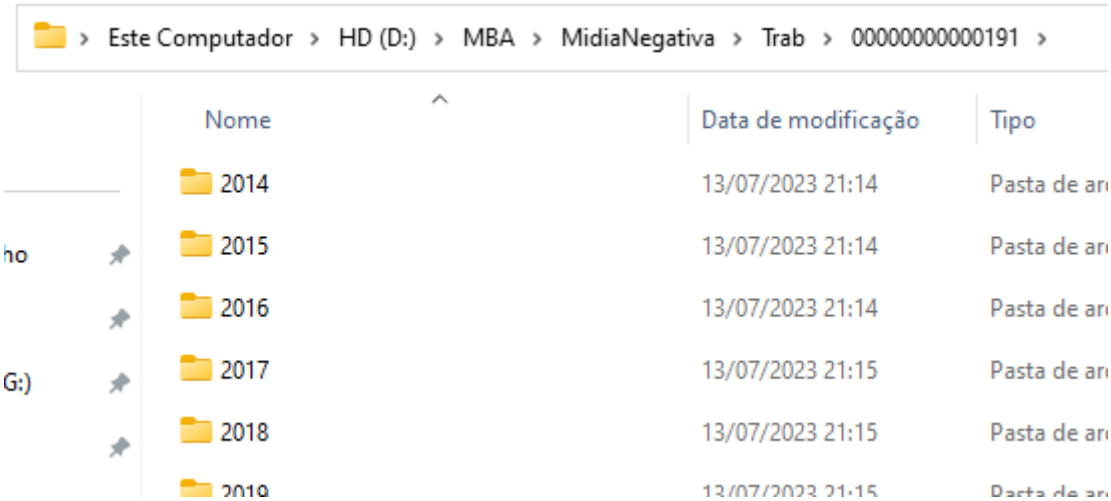
Na validação dos resultados, optou-se por ignorar alguns sites que trazem resultados ruins, como brasilescola.uol.com.br e valorinveste.globo.com/mercados/renda-variavel/bolsas-e-indices e outros.

O Brasil Escola é um site de educação e, no texto em que fala da Petrobrás, trata da criação da empresa por Getúlio Vargas. O link citado do Valor Investe trata principalmente da cotação das empresas nas bolsas de valores. Quando os preços das commodities caem, por exemplo, muitas empresas brasileiras seguem a tendência, do agronegócio aos minérios e energia.

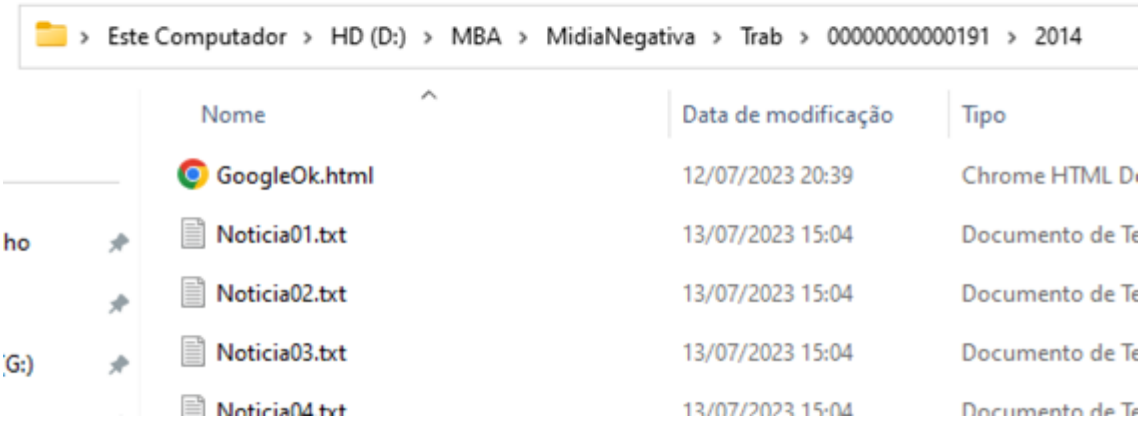
Diretório de trabalho:



Subdiretórios dos anos:


Este Computador › HD (D:) › MBA › MidiaNegativa › Trab › 00000000000191 ›
Nome
Data de modificação
Tipo
2014
13/07/2023 21:14
Pasta de ar
2015
13/07/2023 21:14
Pasta de ar
2016
13/07/2023 21:14
Pasta de ar
2017
13/07/2023 21:15
Pasta de ar
2018
13/07/2023 21:15
Pasta de ar
ho
G:)

Notícias (links e textos):



Links:

```
00000000000191 2014 BANCO+DO+BRASIL
https://g1.globo.com/politica/mensalao/noticia/2014/02/policia-federal-diz-que-prendeu-henrique-pizzolato-na-italia.html
https://g1.globo.com/politica/mensalao/noticia/2014/02/henrique-pizzolato-era-procurado-pela-interpol-relembre-o-caso.html
https://extra.globo.com/economia-e-financas/conheca-os-dez-mandamentos-do-cliente-para-obter-credito-em-banco-14435367.html
https://epoca.oglobo.globo.com/colunas-e-blogs/blog-do-fucs/noticia/2014/12/13-bizarrices-de-dilma-na-economia-das-quais-jama
https://oglobo.globo.com/economia/klabin-ofertara-bonus-no-exterior-para-refinanciar-endividamento-13133485
https://noticias.uol.com.br/politica/ultimas-noticias/2014/03/10/justica-italiana-pode-esperar-mais-de-2-anos-para-decidir-so
https://valor.globo.com/financas/noticia/2014/11/14/concentracao-bancaria-bate-recorde.ghtml
https://globoplay.globo.com/v/3678214/
https://epoca.globo.com/tempo/noticia/2014/02/o-bbanco-do-nordeste-perdeu-milhoesb-em-operacoes-consideradas-irregulares.html
https://epoca.globo.com/tempo/noticia/2014/05/b-evidencias-de-fraudeb-no-fundo-dos-correios-ligado-ao-pmdb.html
 <!doctype html><html itemscope="" itemtype="http://schema.org/SearchResultsPage" lang="pt-BR"><head><meta charset="UTF-8"><me
content="/images/branding/googleg/1x/googleg_standard_color_128dp.png" itemprop="image"><style>@font-face{font-family:'Google
weight:400;font-display:optional;src:url(//fonts.gstatic.com/s/googlesans/v14/4UaGrENHsxJlGDuGo1OIlL3Kwp5MKg.woff2)format('wo
0490-0491,U+04B0-04B1,U+2116;}@font-face{font-family:'Google Sans';font-style:normal;font-weight:400;font-
display:optional;src:url(//fonts.gstatic.com/s/googlesans/v14/4UaGrENHsxJlGDuGo1OIlL3Nwp5MKg.woff2)format('woff2');unicode-ra
```

Uma das notícias:

```
https://g1.globo.com/politica/mensalao/noticia/2014/02/policia-federal-diz-que-prendeu-henrique-pizzolato-na-italia.html
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "https://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd"><html dir="ltr" xml:lar
xmlns="https://www.w3.org/1999/xhtml" xmlns:fb="https://www.facebook.com/2008/fbml" itemscope itemtype="https://schema.org/"><head><m
content="text/html;charset=UTF-8" /><title>G1 - Polícia Federal diz que Henrique Pizzolato foi preso na Itália - notícias em Julgamer
equiv="Content-Type" content="text/html;charset=UTF-8" /><meta content="G1 - Julgamento do mensalão" name="editoria" /><meta content=
name="dtnoticia" /><meta content="Polícia Federal diz que Henrique Pizzolato foi preso na Itália" name="title" /><meta content="Pizzo
do mensalão considerado foragido. Ele pegou 12 anos e 7 meses de prisão por peculato e outros crimes." name="description" /><meta cor
Catarina, Julgamento do mensalão" name="keywords" /><meta property="og:title" content="Polícia Federal diz que Henrique Pizzolato foi
property="og:type" content="article" /><meta property="article:section" content="Julgamento do mensalão" /><meta property="article:pu
content="2014-02-05T12:39:25"/><meta property="og:locale" content="pt_BR" /><meta property="og:url"
content="https://g1.globo.com/politica/mensalao/noticia/2014/02/policia-federal-diz-que-prendeu-henrique-pizzolato-na-italia.html" />
content="https://s2.glbimg.com/J9tn4zQwDWGxjB7bGJMxgDfdyKc=/1200x630/filters:max_age(3600)/s01.video.glbimg.com/deo/vi/64/66/3126664"
content="1200" /><meta property="og:image:height" content="630" /><meta property="og:site_name" content="Julgamento do mensalão" /><m
content="289255557788943" /><link rel="stylesheet" href="https://s.glbimg.com/jo/g1/static/live/COMPR/css/d8/c5f2fa7761d8.css" type="
rel="stylesheet" href="https://s.glbimg.com/jo/g1/static/live/COMPR/css/3e/04d979b75a3e.css" type="text/css" media="print" /><link re
href="https://s.glbimg.com/jo/g1/static/live/COMPR/css/85/0310ea48e785.css" type="text/css" media="screen" /><link type="text/css" re
href="https://comentarios.globo.com/static/widget/css/comentarios.v2.all.css" /><link type="text/css" rel="stylesheet" media="screen"
href="https://s.glbimg.com/jo/g1/static/live/fonts/typography.css" /><link rel="stylesheet" href="https://s.glbimg.com/jo/g1/static/l
type="text/css" media="screen" /><link type="text/css" rel="stylesheet" href="https://s.glbimg.com/bu/c/busca.padrao.suggest.css" /><
_ssi/dynamic/folder-style/640/" --><link type="text/css" rel="stylesheet" media="screen" href="https://s.glbimg.com/jo/g1/o/politica/
v20151123011400.css" /><link rel="shortcut icon" href="https://s.glbimg.com/jo/g1/static/live/portal/img/logos/favicon.png" /><script
SETTINGS = {}; SETTINGS.STATIC_URL = 'https://s.glbimg.com/jo/g1/static/live/'; </script><script type="text/javascript"
```

- Evidência dos resultados:

Analisando manualmente as notícias apontadas como negativas (tirando alguns sites como citado anteriormente), 61% dos textos eram realmente negativos.

	A	B	C	D	E	F	G	H	I	J	K
1	Cnpj	Razão Social	Nome Ajustado	Ano	Nota (x 100)	Negativa ou Erro ou Pesq/Site Ruim	Url		N (%) =	61%	
2	00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2014	-81	N	https://epoca.oglobo.globo.com/colunas				
3	00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2014	-69	S	https://globoplay.globo.com/v/3678214/				
4	00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2014	-46	N	https://epoca.globo.com/tempo/noticia/				
5	00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2015	-53	N	https://g1.globo.com/economia/noticia/				
6	00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2015	-25	E	https://g1.globo.com/sp/sao-carlos-regia				
7	00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2015	-18	N	https://g1.globo.com/tecnologia/blog/se				
8	00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2015	-25	N	https://www.bbc.com/portuguese/notic				
9	00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2015	-75	N	https://g1.globo.com/to/tocantins/notici				
10	00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2015	-59	N	https://g1.globo.com/economia/noticia/				
11	00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2015	-85	S	https://globoplay.globo.com/v/4018187/				
12	00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2017	-38	E	https://oglobo.globo.com/cultura/artes-				
13	00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2019	-88	N	https://www.bbc.com/portuguese/brasil				
14	00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2019	-10	S	https://extra.globo.com/noticias/brasil/t				
15	00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2020	-84	N	https://epocanegocios.globo.com/Brasil/				
16	00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2020	-13	N	https://portaldobitcoin.uol.com.br/agen				
17	00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2022	-69	E	https://www.bbc.com/portuguese/brasil				
18	00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2023	-8	N	https://www.bol.uol.com.br/noticias/20				
19	00.000.208/0001-00	BRB BANCO DE BRASILIA SA	#BRB#	2023	-77	E	https://jc.ne10.uol.com.br/cultura/2023/				
20	00.001.180/0001-26	CENTRAIS ELETRICAS BRASILEIRAS S	#ELETROBRAS#	2014	-97	N	https://m.folha.uol.com.br/mercado/201				

Analisando algumas notícias apontadas como positivas, 10% dos textos eram negativos.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Cnpj	Razão Social	Nome Ajustado	Ano	Nota (x 100)	Negativa?	Url						
2	00.000.000/0001-91	BANCO DO BRASIL SA	#BANCO DO BRASIL#	2014	97		https://g1.globo.com/politica/mensalao/noticia/2014/02/policia-fe						
3	00.000.208/0001-00	BRB BANCO DE BRASILIA SA	#BRB#	2015	89		https://g1.globo.com/distrito-federal/noticia/2015/04/empresario-						
4	00.001.180/0001-26	CENTRAIS ELETRICAS BRASILEIRAS S	#ELETROBRAS#	2014	71		https://g1.globo.com/am/amazonas/noticia/2014/06/foi-tudo-exag						
5	00.336.701/0001-04	TELECOMUNICACOES BRASILEIRAS	#TELEBRAS#	2014	73		https://gq.globo.com/Prazeres/Poder/noticia/2014/08/os-15-advog						
6	00.383.281/0001-09	BNDES PARTICIPACOES SA BNDESP	#BNDESPAR#	2014	71		https://valor.globo.com/agronegocios/noticia/2014/09/25/comeca-						
7	00.416.968/0001-01	BANCO INTER S.A	#BANCO INTER#	2017	99	N	https://g1.globo.com/economia/noticia/sp-coloca-notas-de-bancos						
8	00.776.574/0001-56	AMERICANAS S.A - EM RECUPERAC	#AMERICANAS SA#	2022	99		https://g1.globo.com/tecnologia/noticia/2022/02/25/site-do-shopt						
9	00.864.214/0001-06	ENERGISA S/A	#ENERGISA#	2018	99	N	https://g1.globo.com/economia/noticia/apos-cortar-nota-do-brasil						
10	01.083.200/0001-18	NEOENERGIA S.A	#NEOENERGIA#	2018	92	N	https://valor.globo.com/empresas/noticia/2018/01/22/distribuidor						
11	01.545.826/0001-07	GAFISA S/A.	#GAFISA#	2015	86		http://www1.folha.uol.com.br/mercado/2015/09/1678612-mudanca						
12	01.838.723/0001-27	BRF S.A.	#BRF#	2014	64		https://valor.globo.com/agronegocios/noticia/2014/09/10/receita-(						
13	01.896.779/0001-38	CSU DIGITAL S.A.	#CSU DIGITAL#	2021	82		https://valor.globo.com/financas/noticia/2021/09/24/com-apoio-do						
14	02.062.747/0001-08	SUDESTE S.A.	#SUDESTE SA#	2020	99		https://g1.globo.com/sp/campinas-regiao/terra-da-gente/noticia/2						

Considerando que o objetivo do trabalho é encontrar notícias “ruins”, podemos estimar os seguintes valores (desprezando os sites inadequados):

Estimativa	
Total de notícias	1.953
Notícias apontadas como ruins	
Notícias realmente ruins	93
Notícias realmente boas	60
Notícias apontadas como boas	
Notícias realmente ruins	180 (10%)
Notícias realmente boas	1.620 (90%)

Com a seguinte matriz de confusão e as seguintes métricas:

Matriz de Confusão	
Verdadeiro Positivo: 93	Falso Negativo: 180
Falso Positivo: 60	Verdadeiro Negativo: 1.620

Métricas	
Acurácia	88%
Precisão	61%
Revocação	34%
F1-score	44%

Considerando, novamente, que o objetivo do trabalho é encontrar notícias “ruins” e que não precisamos de todas, a métrica mais importante é a precisão (que utilizamos quando calculamos o número de notícias efetivamente ruins em relação ao número de notícias apontadas como ruins).

Assim, embora o trabalho de classificação das notícias continue exigindo a intervenção humana, ele fica muito facilitado quando a equipe que faz a classificação lendo apenas as notícias apontadas como “ruins”.

Sem o método proposto, a equipe precisa ler 7 notícias para encontrar 1 ruim (1.953 para encontrar 273). Com o método, precisa ler 5 para encontrar 3 ruins (153 para encontrar 93). E o sistema poderia triplicar o número de notícias pesquisadas para que o número de notícias ruins fosse igual ao do trabalho manual.

O resultado fica aquém do esperado (substituir totalmente a intervenção humana), mas, como veremos nas “considerações finais”, não de forma definitiva.

2.3.2 Lições Aprendidas

O método proposto ainda pode receber vários ajustes. Aqui vão alguns exemplos:

a) Melhor trabalhar com um número limitado de sites. Em vez do portal uol.com.br, por exemplo, apenas com folha.uol.com.br ou noticias.uol.com.br; em vez do portal globo.com, apenas com g1.globo.com ou oglobo.globo.com.
b) Melhor considerar somente os textos em que o nome da empresa aparece mais de uma vez. Isto aumenta a chance de que a empresa seja realmente o foco da notícia em questão.
c) O programa precisa confirmar se o nome da empresa está em letra maiúscula e não faz parte de expressões não relacionadas à pesquisa que está sendo feita. A pesquisa pela empresa Natura, por exemplo, trouxe vários textos com a expressão “in natura”.

d) Parece melhor considerar apenas os parágrafos próximos ao nome da empresa (isto não foi testado) para aumentar a chance de não considerarmos outros textos ou links para outras notícias.

e) Notícias sobre eventos específicos, como a pandemia do coronavírus, devem ser tradados separadamente. Várias notícias sobre o que as empresas estavam fazendo para colaborar com a sociedade foram classificadas incorretamente. O texto era negativo porque falava da pandemia, mas a ação das empresas era (muitas vezes) positiva, como quando a Ambev produziu álcool em gel.

3. Considerações Finais

3.1 Resultados

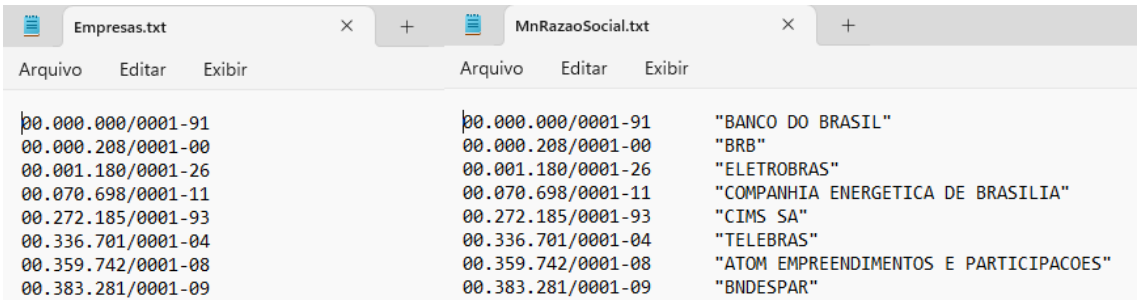
O objetivo desta pesquisa foi automatizar um trabalho que costuma ser realizado manualmente: a busca de notícias negativas relacionadas a determinadas empresas. Em vez de um operador digitar o nome de uma empresa no Google, clicar em "notícias", abrir pelo menos dez links para ler e avaliar, nossa proposta era que tudo fosse feito automaticamente.

A solução encontrada substitui grande parte do trabalho:

1) Um programa, em C#, lê um arquivo com diversos CNPJs e grava um arquivo com a razão social das empresas a partir de uma API do site do TRT 20.

Exemplo: https://www.trt20.jus.br/standalone/wsreceita.php?cpfSolicitante=00000000191&cpf=08028776000121 (o que muda, em cada pesquisa, é o CNPJ no final da ULR)

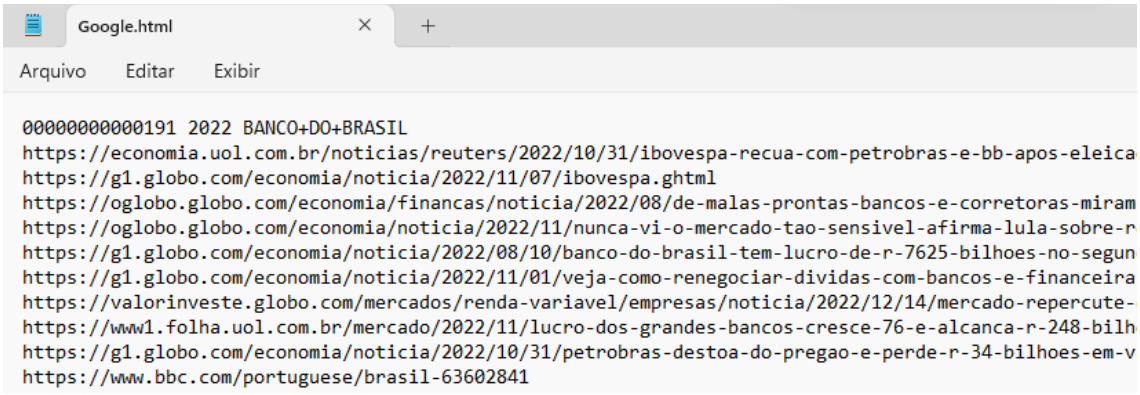
2) O arquivo com a razão social das empresas é modificado manualmente para que somente a parte principal da razão social seja usada na busca.



3) O programa C# procura os nomes das empresas entre as notícias do Google definindo, também, um período (ano a ano, nos últimos 10 anos) e os sites que devem ser considerados.

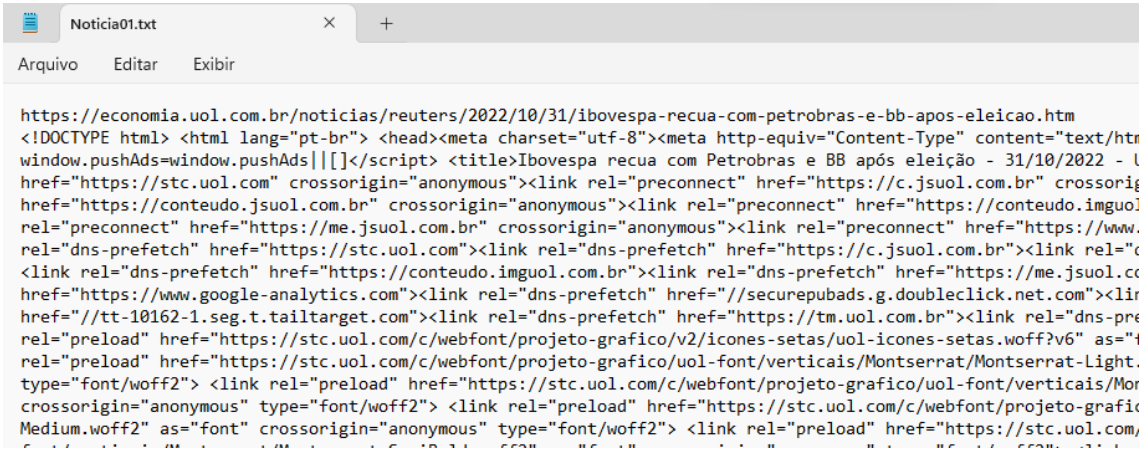
O exemplo abaixo traz as notícias em português, referentes ao BANCO DO BRASIL, no ano de 2022, nos portais globo.com, uol.com.br, bbc.com e dw.com:

https://www.google.com/search?tbm=nws&lr=lang_pt&tbs=cdr:1,cd_min:01/01/2022,cd_max:12/31/2022&q=%22BANCO+DO+BRASIL%22+site%3Aglobo.com+OR+site%3Auol.com.br+OR+site%3Abbc.com+OR+site%3Adw.com

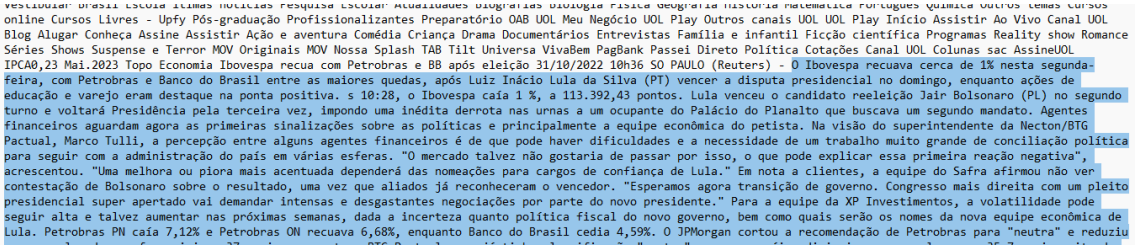


4) O programa C# abre cada link encontrado no item anterior, baixa a notícia e, com a biblioteca HtmlAgilityPack, tira o HTML, o CSS e o JavaScript.

Página da notícia:



Notícia sem HTML, CSS e JavaScript:


online Cursos Livres - Upfy Pós-graduação Profissionalizantes Preparatório OAB UOL Meu Negócio UOL Play Outros canais UOL UOL Play Início Assistir Ao Vivo Canal UOL
Blog Alugar Conheça Assine Assistir Ação e aventura Comédia Criança Drama Documentários Entrevistas Família e infantil Ficção científica Programas Reality show Romance
Séries Shows Suspense e Terror MOV Originais MOV Nossa Splash TAB Tilt Universa VivaBem PagBank Passei Direto Política Cotações Canal UOL Colunas sac AssineUOL
IPCA0,23 Mai.2023 Topo Economia Ibovespa recua com Petrobras e BB após eleição 31/10/2022 10h36 SO PAULO (Reuters) - O Ibovespa recuava cerca de 1% nesta segunda-
feira, com Petrobras e Banco do Brasil entre as maiores quedas, após Luiz Inácio Lula da Silva (PT) vencer a disputa presidencial no domingo, enquanto ações de
educação e varejo eram destaque na ponta positiva. s 10:28, o Ibovespa caía 1 %, a 113.392,43 pontos. Lula venceu o candidato reeleição Jair Bolsonaro (PL) no segundo
turno e voltará Presidência pela terceira vez, impondo uma inédita derrota nas urnas a um ocupante do Palácio do Planalto que buscava um segundo mandato. Agentes
financeiros aguardam agora as primeiras sinalizações sobre as políticas e principalmente a equipe econômica do petista. Na visão do superintendente da Necton/BTG
Pactual, Marco Tulli, a percepção entre alguns agentes financeiros é de que pode haver dificuldades e a necessidade de um trabalho muito grande de conciliação política
para seguir com a administração do país em várias esferas. "O mercado talvez não gostaria de passar por isso, o que pode explicar essa primeira reação negativa",
acrescentou. "Uma melhora ou piora mais acentuada dependerá das nomeações para cargos de confiança de Lula." Em nota a clientes, a equipe do Safra afirmou não ver
contestação de Bolsonaro sobre o resultado, uma vez que aliados já reconheceram o vencedor. "Esperamos agora transição de governo. Congresso mais direita com um pleito
presidencial super apertado vai demandar intensas e desgastantes negociações por parte do novo presidente." Para a equipe da XP Investimentos, a volatilidade pode
seguir alta e talvez aumentar nas próximas semanas, dada a incerteza quanto política fiscal do novo governo, bem como quais serão os nomes da nova equipe econômica de
Lula. Petrobras PN caía 7,12% e Petrobras ON recuava 6,68%, enquanto Banco do Brasil cedia 4,59%. O JPMorgan cortou a recomendação de Petrobras para "neutra" e reduziu

5) Um programa, em Python, classifica as notícias usando a biblioteca NLTK (vader_lexicon). O resultado é o arquivo mostrado abaixo: CNPJ, ano, link da notícia e uma nota que vai de -1 a +1, sendo -1 uma notícia muito negativa e +1 uma notícia muito positiva. (Só as negativas interessam.)

Embora os resultados tenham ficado abaixo das expectativas iniciais, a solução proposta consegue selecionar diversas notícias negativas com uma precisão de cerca de 60%. **Manualmente, um operador precisa ler 21 notícias para encontrar 3 negativas. Com a solução proposta, precisa ler 5 para encontrar as mesmas 3. O processo não foi completamente automatizado, mas reduz a "análise de sentimento" do operador em 75%.**

Considerando, ainda, que a solução proposta substitui a busca manual no Google selecionando o período e os sites pesquisados, o trabalho do operador diminui mais um pouco.

(Mais adiante, no item 3.3, citaremos algumas dificuldades encontradas e como aprimorar a solução proposta.)

3.2 Contribuições

Podemos dividir a solução proposta em duas partes: a que foi feita em C# e a que foi feita em Python.

A que foi feita em C# substitui bem os procedimentos que não são “inteligentes” e abre novas possibilidades com a escolha dos sites em que as notícias devem ser pesquisadas.

A parte que foi feita em Python serve como uma triagem inicial das notícias e, como já dissemos, reduz o trabalho em 75%.

3.3 Próximos passos

A solução proposta já traz um ganho significativo para quem pesquisa “mídia negativa”, mas pode ser aprimorada comparando a classificação feita pela inteligência artificial com a classificação feita por humanos.

1) Sites pesquisados: Sites educacionais ou focados no mercado financeiro não são bons para nossos propósitos. Analisando individualmente as métricas de cada site, a solução pode focar nos sites com notícias mais relevantes e análises de sentimento mais corretas.

2) Relação da empresa pesquisada com a notícia: Muitos textos citam a empresa pesquisada, mas não são exatamente sobre ela. Selecionando textos em que o nome da empresa apareça três ou mais vezes, é provável que os resultados sejam melhores.

3) Seleção do texto: Alguns textos são negativos do modo geral, mas positivos em relação à empresa pesquisada ou o contrário. Analisando apenas os parágrafos em que o nome da empresa aparece, este problema pode ser minimizado.

4) O item 3 faz com que os textos sejam mais curtos, o que pode facilitar a geração de resultados melhores mesmo com modelos relativamente simples para análise de sentimento.

5) Com um histórico razoável, podemos trabalhar com um modelo supervisionado para a classificação dos textos. O resultado, então, tende a ser significativamente melhor.