Recuperação e Mineração de Texto sobre as Compras Governamentais

Autor: Rafael Odon de Alencar

Email: odon.rafael@gmail.com

Data: 15/11/2018

Introdução

O presente trabalho busca exercitar técnicas de recuperação de informação e de mineração de texto através do desenvolvimento de um sistema que coleta, extrai, processa e analisa sob determinadas óticas o conteúdo textual descritivo de uma amostra das compras do Governo Federal. Os dados observados encontram-se disponíveis publicamente no site de http://compras.dados.gov.br.

As compras feitas pelo governo podem ser do tipo **com licitação** ou **sem licitação**, e são categorizadas com ajuda de um catálogo de **serviços** e **materiais** que agrupa compras de segmentos semelhantes. No enatnto, o volume de informações e a complexidade da base torna difícil contemplar as características gerais do comportamento de compra por parte das entidades públicas.

Afim de demonstrar o potencial de uma ferramenta automatizada para auxiliar na recuperação e análise em torno do texto dessas compras, foram selecionados apenas 2 serviços específicos do catálogo:

- Serviço 17663: Curso Aperfeiçoamento / Especialização Profissional
- Serviço 3239: Transporte Rodoviário - Pessoal por Automóveis

Em resumo, foram coletadas todas compras com e sem licitação desses dois serviços, e o conteúdo textual descritivo desses documentos foi extraído, processado e utilizado para gerar *insights*. As compras foram classificadas quanto à faixa de gasto a partir de uma análise da estatística descritiva. Em seguida, foram geradas nuvens de palavras destacando os termos descritivos de maior frequência para o **grupo de gastos menores** e para o **grupo de gastos maiores**. Um modelo de classificação *Naive Bayes* foi utilizado para verificar os termos que mais contribuiram para discriminar cada uma dessas classes. Também foi aplicada a técnica LDA (*Latent Dirichlet Allocation*) de detecção de tópicos em cada um desses grupos para verificar a co-ocorrência de termos nos conjunto de documentos. Por fim, uma estratégia de detecção de compras suspeitas foi proposta.

O sistema foi construído em Python 3.5 com ajudas de bibliotecas tais como *Pandas*, *Nltk*, *Scikit-Learn*, *Gensim*, *Matplotlib*, *Wordcloud* dentre outras. O código foi separado em classes conforme as responsabilidades do fluxo de trabalho: **Coletor**, **Extrator**, **Processador** e **Analisador**.

Cada uma das etapas será melhor descrita nas seções seguintes, bem como as observações e conclusões obtidas após as análises feitas.

Coleta dos documentos

O site http://compras.dados.gov.br não só permite navegar pelos dados através de sua interface em HTML, mas também oferece APIs que retornam documentos Json. Há uma API própria para as licitações, e outra API própria para as compras sem licitação. Ambas possuem características diferentes mas permitem

igualmente consultar uma numerosa lista paginada com todas as compras de um determinado serviço. Em ambas as APIs, uma compra pode envolver mais de um item, e assim é preciso também fazer novos acessos para encontrar os detalhes textuais daquele item.

Foi desenvolvida uma estratégia de coletada automatizada que busca todas as compras e licitações de um determinado serviço, com seus respectivos itens. A classe **Coletor** é responsável por essa parte do fluxo de trabalho, navegando pelas APIs, indo para as próximas páginas quando essas existem e guardando todas as respostas Json obtidas num **diretório de cache**. Dessa forma, ao ser re-executada, as compras já coletadas não são re-visitadas. Com isso, uma vez finalizada a coleta de todas as respostas das compras de um serviço é possível trabalhar *offline* sem a necessidade de acessar a API novamente.

Foi executada a coleta tanto para o serviço 17663 (Curso Aperfeiçoamento / Especialização Profissional) quanto para o serviço 3239 (Transporte Rodoviário - Pessoal por Automóveis). Ao fim das coletas, constaram mais de 40 mil arquivos JSON no direótório de cache. Novos serviços podem ser coletados se houver interesse.

Além da navegação nas APIs de compras, também foi feita uma coleta simples da página de divulgação oficial da taxa SELIC (https://www.bcb.gov.br/pec/copom/port/taxaselic.asp), afim de subsidiar a atualização monetária dos valores das compras durante a análise.

As coletas ocorreram entre 30/10/2018 e 15/11/2018.

Extração de dados

Mediante a coleta finalizada dos documentos de um serviço, o **Extrator** é responsável por fazer o *parse* dos documentos coletados, organizando as informações em um banco de dados relacional SQLite3 que torna fácil consultar as informações dos documentos.

Foram extraídos e armazenados como registros de uma tabela de documentos os seguintes dados:

- Id da Compra
- Id do Serviço
- Texto descritivo da compra
- Texto descritivos dos itens da compra
- Valor da compra (DOUBLE)
- Data da compra (DATE)
- Tipo (com licitação / sem licitação)

Após a extração dos dados das compras, o banco de dados apresentou 3396 documentos do serviço 17663 (especialização) e 1842 documentos do serviço 3239 (transporte rodoviário).

Cada serviço tem os seus dados armazenados num banco separado, facilitando o manuseio nas etapas posteriores, já que as análises planejadas são feitas separadamente por serviço. A presença da coluna **Tipo** torna o banco preparado para ser multi-serviço se necessáario. Ademais, a re-execução do **Extrator** apaga o banco e cria um novo, mas há como desligar essa abordagem.

O extrator também é responsável por fazer o *scraping* do HTML da página com o histórico da taxa SELIC mensal desde 1997, recuperando as células da tabela através de expressões *XPath* e armazenando os dados obtidos numa tabela onde consta:

- Data início (DATE)

- Data fim (DATE)
- Valor da taxa SELIC (DOUBLE)

Processamento dos dados

O **Processador** assume a existência do banco de dados relacional SQLite3 fruto da execução do **Extrator**, e a partir dele cria novas colunas e arquivos de apoio com dados que irão subsidiar a análise.

A primeira responsabilidade do **Processador** é pré-processar o conteúdo textual de cada documento para tornar possível a criação de um bag-of-words mais otimizado que dará suporte às análises sobre os termos. O pré-processamento do texto incluiu:

- **Texto em minúsculo** - Optou-se por tratar todas as palavras em minúsculo.
- **Remoção de acentos** - Foi verificado navegando nos documentos extraídos, há ocorrência de palavras iguais com e sem acentuação ao longo dos textos, o que prejudica a correta contagem da frequência dos termos.
- **Remoção da pontuação** - Como não houve necessidade de preservar as sentenças, todas as palavras ficaram separadas por um único espaço, facilitando tratamentos posteriores. Isso foi feito com ajuda de um **RegexpTokenizer** que trasnformou o texto numa lista de palavras, ignorando espaços adjacentes e pontuação.
- **Remoção de tokens numéricos** - Tokens apenas numéricos não foram incluídos no texto processado final pois não apresentaram benefício para a análise.
- **Remoção de palavras do domínio** - Algumas expressões específicas do assunto Compras Goveernamentais estavam presentes nos documentos mas não contribuiram para uma boa compreensão do conteúdo das compras através da frequência de teremos. Sendo assim alguns termos foram removidos:

  ```
  REMOVER = [ 'pregao eletronico', 'pregao', 'aquisicao', 'valor',
    'limite' 'licitacao', 'licitacao', 'justificativa', 'edital',
    'contratacao', 'fornecimento', 'prestacao', 'precos', 'preco',
    'formacao','empresa', 'servico', 'servicos', 'inscricao',
    'pagamento', 'taxa','para', 'objeto' ]
  ```

- **União de palavras quebradas:** Ao investigar a base de documentos visualmente, foi verificada uma grande ocorrência de palavras quebradas que deveriam estar unidas (ex: ca pacaitacação -> capacitação, traba lho -> trabalho). Para tenta resolver esse problema, foi proposta uma heurística sobre a sequência de tokens do texto. Se o *token i* concatenado ao *token i+1* formar uma palavra uma palavra integrante do vocabulário composto por todos documentos em questão, cuja frequência dessa palavra unida seja maior que 25% da frequência dos tokens separados, então os 2 tokens adjacentes são transformados num único token concatenado.

  Para que essa estratégia funcionasse foi preciso realizar uma primeira passada em todos os documentos para criar esse vocabulário e calcular as frequências dos tokens. Os resultados foram satisfatórios e trouxeram maior qualidade para a etapa de análise.

O trecho de LOG abaixo demonstra algumas uniões de palavras que aconteceram durante o processamento:

```
...
2018-11-15 10:44:11,072 [DEBUG] - Unindo crit+erio
2018-11-15 10:44:11,073 [DEBUG] - Unindo maqu+ina
2018-11-15 10:44:11,099 [DEBUG] - Unindo mentori+ng
2018-11-15 10:44:11,102 [DEBUG] - Unindo minis+trar
2018-11-15 10:44:11,109 [DEBUG] - Unindo mer+cado
2018-11-15 10:44:11,109 [DEBUG] - Unindo doce+ntes
2018-11-15 10:44:11,109 [DEBUG] - Unindo integr+ada
2018-11-15 10:44:11,109 [DEBUG] - Unindo oite+nta
2018-11-15 10:44:11,126 [DEBUG] - Unindo univ+ersitaria
2018-11-15 10:44:11,130 [DEBUG] - Unindo tra+nsferencia
2018-11-15 10:44:11,132 [DEBUG] - Unindo amb+iente
2018-11-15 10:44:11,137 [DEBUG] - Unindo enc+adernacao
2018-11-15 10:44:11,142 [DEBUG] - Unindo w+indows
2018-11-15 10:44:11,144 [DEBUG] - Unindo execut+iva
2018-11-15 10:44:11,149 [DEBUG] - Unindo f+ormacao
2018-11-15 10:44:11,152 [DEBUG] - Unindo vi+deo
2018-11-15 10:44:11,163 [DEBUG] - Unindo te+cnicos
...
```

- **Stemming** - Por fim, a sequência de termos pré-processados é reduzida ao seu radical usando o *stemmer* para Português **nltk.stem.RSLPStemmer**. Como o objetivo era entender o panorama das compras governamentais de um determinado serviço, era importante também contemplar termos legíveis nas análises. Para tanto, ao realizar o *stemming*, foi armazenado num dicionário as frequências de cada variação do radical, para que num pós-processamento a top-palavra fosse utilizada como representante daquele conjunto de termos.

  Abaixo segue uma entrada do dicionário de frequências:

```
"estim": {
    "estimativas": 24,
    "estimada": 4,
    "estimados": 1,
    "estimativa": 1,
    "estimado": 3,
    "estimadas": 1
},
```

  No caso acima, a palavra **estimativas** é a melhor representante do radical **estim**, e será usada por exemplo para representar todas as demais palavras desse radical numa nuvem de palavra.

As decisões de pré-processamento do texto acima descritam foram feitas iterativamente com as análises, principalmente observando a qualidade da nuvem de palavra. O pré-processamento foi primordial para gerar uma nuvem com menos 'sujeira' e com frequências mais significativas para determinados assuntos. Abaixo segue um exemplo do texto antes e depois do processamento:

Texto puro	Texto processsado
Frete de veiculo no percurso redencao/kikretum/redencao.. Objeto: Pregão Eletrônico - Contratação de emp resa especializada em serviço de instalação de linha de gases especiais.Justificativa: Conduzindo professores para a aldeia kikretum.	*frete veiculos percurso redencao kikretum redencao empresa especializada instalacao linha gases especial conduzir professores aldeia kikretum*

O processamento final feito sobre o texto foi o ajuste de um objeto do tipo **sklean.feature_extraction.text.TfidfVectorizer** que recebeu como entrada o texto pré-processado de cada documento, e ajustou-se para fazer o cálculo do TF-IDF (Term Frequency - Inverse Document Frequncy), que é uma medida que traduz a frequência de um termo naquela coleção de documentos, levando em conta também que termos frequentes em mutos documentos são menos discriminantes.

O vetorizador foi configurado para trabalhar com 2000 palavras, cada uma sendo uma dimensão do *bag-of-words* final que pode representar um documento no espaço vetorial. O vetorizador também foi configurado para ignorar stopwords da língua Portuguesa através do **nltk.corpus.stopwords**.

Esse vetorizador foi serializado para ser usado durante a análise sempre que fosse necessário avaliar as frequências dos termos da coleção, ou vetorizar um conjunto de documentos.

Outra parte do processamento, não relacionada ao texto, foi o atualização monetária dos valores das compras pela taxa SELIC, permitindo assim uma análise mais justa das faixa de gasto maior e menor das compras durante a análise.

Análise

Definição das faixas de gasto

Uma análise estatística descritiva foi feito sobre os valores das compras de ambos os serviços. Em ambos os casos, verificou-se que, conforme é possível ver nas tabelas e nos gráficos, os valores de compras mais altos só ocorrem próximos do percentil 98, 99. Ou seja, a maior parte das compras tem valores moderados se comparadas com os valores máximos.

Tabela 1: estatística descritiva do serviço 17663 (Cursos)

descritiva	valor
count	2765
mean	47737
std	389498
min	8
0%	8
15%	2002
30%	3384
45%	5499

descritiva	valor
60%	8522
75%	15915
90%	46917
93%	63320
96%	122307
99%	830910
max	15324904

Figura 1 - Histograma dos valores do Serviço 17663 (Cursos)

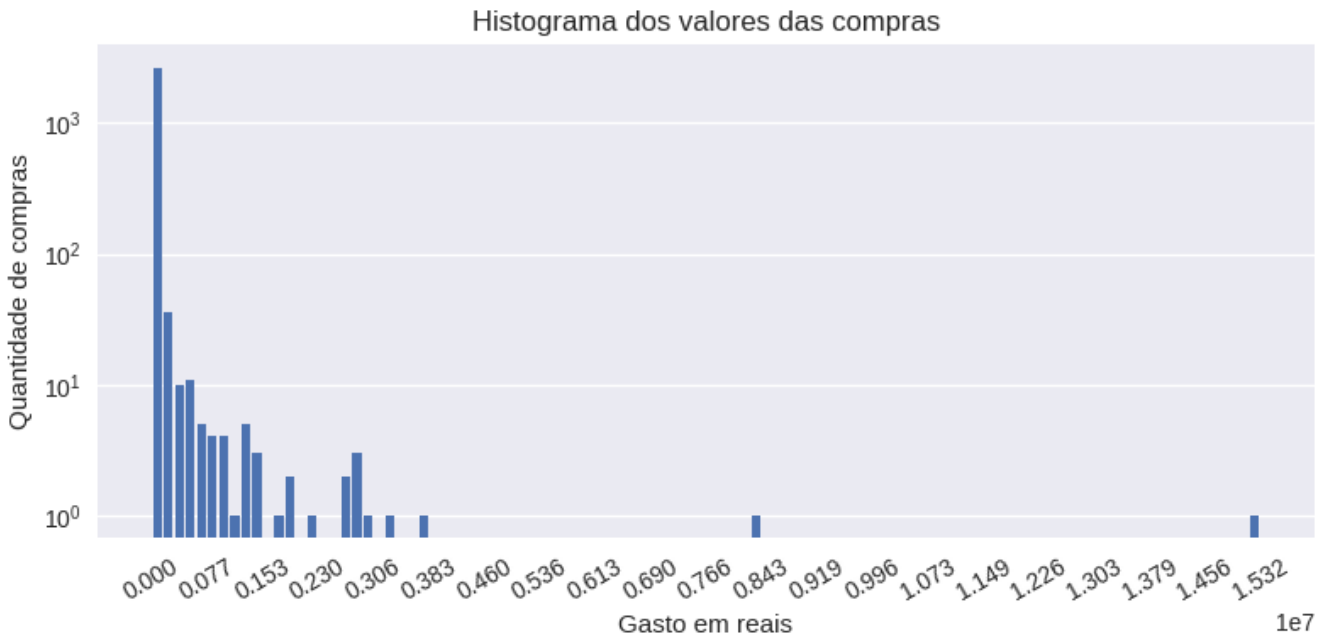
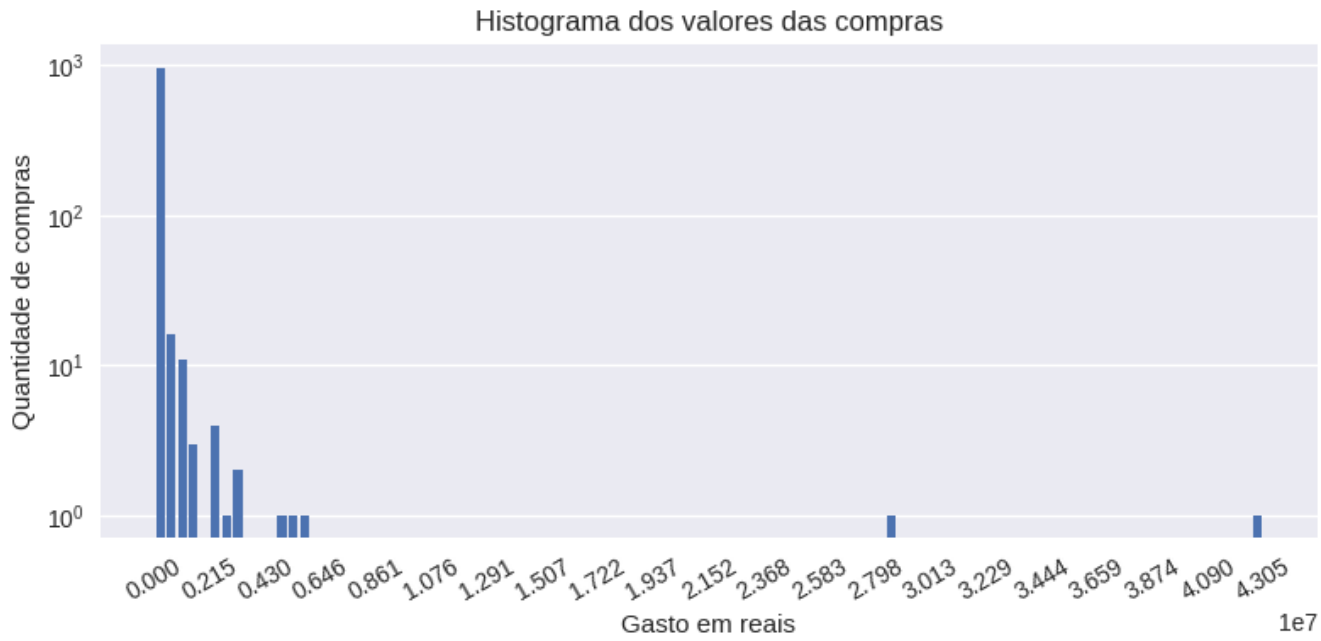


Tabela 1: estatística descritiva do serviço 3239 (Transporte)

descritiva	valor
count	1011
mean	148983
std	1673170
min	7
0%	7
15%	1609
30%	2247
45%	4183

descritiva	valor
60%	7687
75%	18283
90%	68555
93%	126200
96%	442888
99%	2188247
max	43048801

Figura 1 - Histograma dos valores do Serviço 17663 (Cursos)



Para definir o ponto de corte da faixa de menor gasto para para a faixa de maior gasto, foram feitas algumas tentativas. Por fim, optou-se por uma regra que demonstrou chegar num valor adequado para ambos os serviços. O **valor de corte** foi definido como sendo a **média** somada com **1 desvio**.

Com essa abordagem, o ponto de corte do seviço 17663 foi o percentil 98.227848, e o do serviço 3239 foi o percentil 98.813056.

Frequência de termos

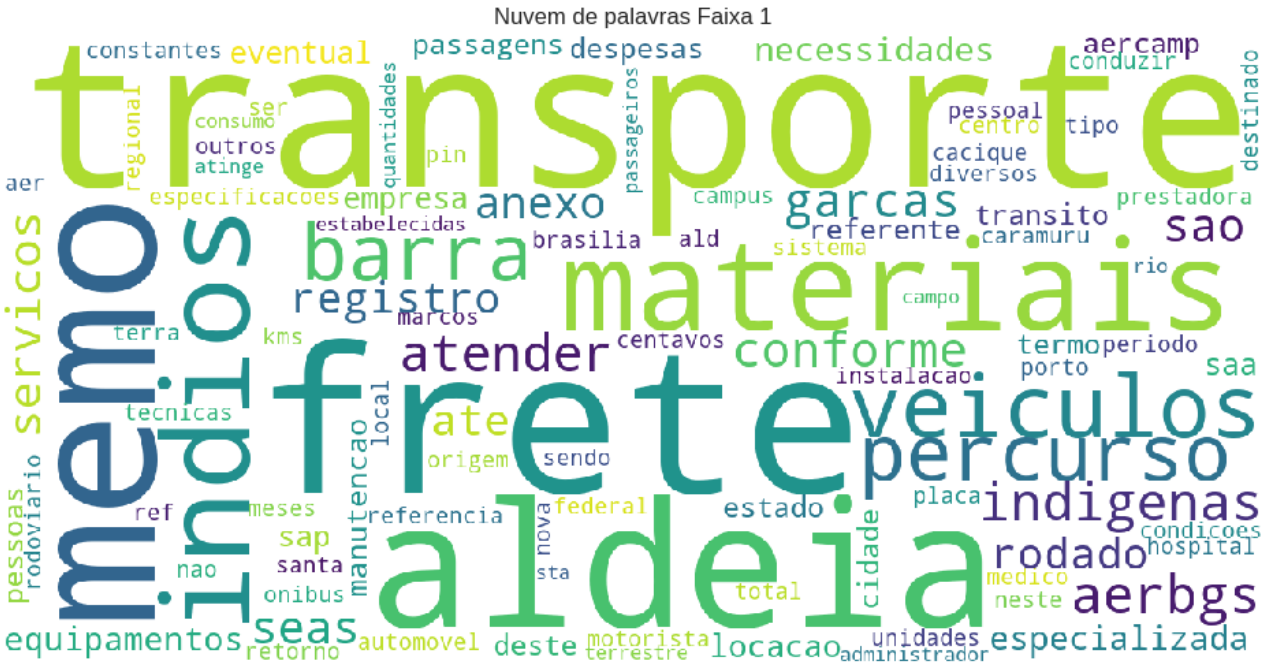
O **Analisador** utiliza os valores TF-IDF advindos do processamento para gerar nuvens de palavras em que os termos com valores de frequências mais altas ficam em destaque. As nuvens foram construídas através do objeto **wordcloud.WordCloud** e foram criadas nuvens separadas para cada faixa de gasto, sendo que a **Faixa 1** é a faixa de menor gasto, e a **Faixa 2** é a faixa de maior gasto.

Também foram gerados na saída, arquivos contendo a lista de palavras mais frequentes, já que a nuvem, apesasr de visualmente atraente, não é muito boa para uma análise cautelosa comparativa entre os conjuntos de termos. (vide arquivos **termos_Faixa1.md** e **termos_Faixa2.md** nas pasta **out**)

Nuvens de palavras do serviço 17663 (Cursos de Especialização)

É possível observar nas nuvens do serviço 17663 que grande parte das compras se relacionam de fato com investimento em aperfeiçoamento de servidores, sendo eles através de cursos, treinamentos, especializaçõões, materiais educativos, congressos. Destaca-se na nuvem da Faixa 2 o termo **saúde** que sugere que podem existir gastos mais elevados nessa área de formação.

Nuvens de palavras do serviço 3239 (Transporte Rodoviário de Pessoas)



As nuvens de palavras do serviço 3239 sugerem compras em torno de serviços de transporte, frete, contratação de motorista, aluguel de veículos. Mas chama a atenção na Faixa 1, de menor gasto, as palavras índios, indigena e aldeia, revelando um tipo de gasto governamental de pouca visibilidade.

Anális de Termos Discriminantes

Afim de comprovar de fato os termos que discriminam uma faixa de gasto da outra, surgiu a idéia de ajustar um classificador *Naive Bayes* para que as informações do cálculo das probabilidades de cada classe pudesse ser usado como um ranking das palavras mais importantes. Dessa forma, foi utilizado o **sklearn.naive_bayes.GaussianNB**, primeiramente fazendo um treino com validação cruzada de 5 *folds* para medir sua acurácia, e em seguida ajustando-se com toda a base para gerar a informação probabilística.

Foi observado a informação da variância (*sigma*) e da média (*theta*)da feature por classe para rankear as top-palavras de cada uma das faixas de gasto.

Top-20 palavras mais discriminantes para a Faixa 1 do serviço 17663:

ranking	palavra	sigma	theta
1	servidores	0.020315	0.128345
2	participacao	0.015974	0.048399
3	referente	0.015537	0.045957
4	periodo	0.012891	0.065668
5	brasilia	0.012198	0.040523
6	maria	0.010902	0.029835
7	realizado	0.010624	0.046874
8	lei	0.010547	0.033297
9	material	0.009802	0.052602
10	aperfeicoamento	0.009550	0.027284
11	registro	0.008832	0.033202
12	manutencao	0.008363	0.029003
13	horas	0.008012	0.020143
14	empresa	0.007952	0.028426
15	participar	0.007557	0.026616
16	atender	0.007556	0.041355
17	consumo	0.007145	0.021562
18	paulo	0.007127	0.020399
19	equipamentos	0.006857	0.026778
20	estabelecido	0.006670	0.026210
21	deste	0.006634	0.025207
22	nacional	0.006415	0.020672
23	direta	0.006321	0.015083
24	curso	0.006310	0.080502
25	mat	0.006262	0.012909
26	sao	0.005792	0.018641
27	conforme	0.005789	0.030746

ranking	palavra	sigma	theta
28	ser	0.005766	0.029811
29	administracao	0.005748	0.018331
30	publica	0.005708	0.024211

Top-20 palavras mais discriminantes para a Faixa 2 do serviço 17663:

ranking	palavra	sigma	theta
1	treinamento	0.010706	0.039954
2	ambiental	0.009945	0.024885
3	graduacao	0.009906	0.038901
4	saude	0.009798	0.052138
5	pos	0.008984	0.032768
6	transito	0.007490	0.020014
7	componentes	0.007230	0.017397
8	educacao	0.006404	0.025989
9	horas	0.006333	0.021662
10	atender	0.005431	0.037598
11	producao	0.005213	0.014358
12	ufpe	0.005105	0.010312
13	servidores	0.005038	0.034969
14	naval	0.004922	0.010127
15	informatica	0.004836	0.018229
16	instituicao	0.004813	0.037921
17	capacitacao	0.004749	0.038096
18	humanos	0.004725	0.016365
19	recursos	0.004528	0.018485
20	especializacao	0.004423	0.041247
21	contratos	0.004229	0.023515
22	resfriamento	0.004215	0.009370
23	profissional	0.004155	0.029968
24	turma	0.004105	0.012278

ranking	palavra	sigma	theta
25	seguranca	0.003964	0.009087
26	ensino	0.003936	0.023565
27	tecnica	0.003848	0.024772
28	interna	0.003832	0.013105
29	atualizacao	0.003766	0.018478
30	projeto	0.003713	0.022890

É possível observar que na faixa 2 inclui assuntos como: ambiental, saúde, trânsito, informática, recursos humanos, segurança. Esses assuntos não constam na lista da Faixa 1. Isso pode sugerir que o compras com treinamento especializado em torno desses temas possam ser maiores ou mais numerosos no governo.

Top-20 palavras mais discriminantes para a Faixa 1 do serviço 3239:

ranking	palavra	sigma	theta
1	materiais	0.050605	0.106390
2	indigenas	0.045294	0.087304
3	servicos	0.029950	0.061576
4	rodado	0.025620	0.042201
5	equipamentos	0.025571	0.055718
6	ate	0.024179	0.047951
7	veiculos	0.023904	0.053977
8	conforme	0.020759	0.066162
9	anexo	0.019888	0.062316
10	especializada	0.015370	0.043200
11	necessidades	0.014916	0.040655
12	atender	0.013932	0.050508
13	pessoal	0.012821	0.024324
14	locacao	0.012735	0.030054
15	manutencao	0.012541	0.032847
16	onibus	0.012337	0.024075
17	sendo	0.011998	0.028858
18	local	0.011329	0.024251

ranking	palavra	sigma	theta
19	prestadora	0.010821	0.023853
20	medico	0.010401	0.018035
21	centro	0.009670	0.022158
22	hospital	0.009652	0.018050
23	sistema	0.009576	0.021816
24	tecnicas	0.009062	0.021343
25	rodoviario	0.008991	0.017003
26	pessoas	0.008973	0.023904
27	referencia	0.008864	0.027221
28	federal	0.008656	0.021758
29	diversos	0.008299	0.018132
30	termo	0.008096	0.027533

Top-20 palavras mais discriminantes para a Faixa 2 do serviço 3239:

ranking	palavra	sigma	theta
1	diaria	0.033232	0.080719
2	horas	0.028042	0.074800
3	nuraf	0.022485	0.045212
4	meses	0.018017	0.068731
5	rodado	0.017411	0.039785
6	media	0.017411	0.039785
7	categoria	0.017411	0.039785
8	ano	0.017411	0.039785
9	pequenas	0.015892	0.081587
10	perimetro	0.015620	0.037683
11	irrigados	0.015620	0.037683
12	pecas	0.014804	0.060123
13	manutencao	0.014797	0.060058
14	locacao	0.014189	0.070948
15	salgado	0.013555	0.035103

ranking	palavra	sigma	theta
16	internacional	0.013555	0.035103
17	filho	0.013555	0.035103
18	alegre	0.013555	0.035103
19	aeroporto	0.013555	0.035103
20	motorista	0.012552	0.093567
21	tipo	0.012027	0.062625
22	ser	0.011199	0.047283
23	ate	0.009923	0.056612
24	firma	0.009004	0.028611
25	veiculos	0.008929	0.115578
26	ans	0.008734	0.034988
27	pernoite	0.008520	0.031458
28	passageiros	0.007946	0.041299
29	cargas	0.007826	0.066090
30	taxi	0.007693	0.026445

Dentre os termos mais discrimantes da Faixa 1 do serviço 3239, de fato está a palavra **indígenas**, comprovando que é um assunto em destaque dessa categoria. Nessa faixa surgem também as palavras **médico**, **hospital**, sugerindo deslocamentos de pessoas para fins de cuidados com a saúde. Já na Faixa 2, esses termos não ocorrem, e dão lugar à termos como **aeroporto**, **taxi**, **pernoite**, **internacional**, sugerindo uma temática ligada ao traslado de pessoas que viajam muito de avião.

Detecção de Tópicos

Para confirmar algumas das suspeitas de temas em torno das compras, foi executado o algoritmo do LDA que observa a co-ocorrência de palavras nos documentos, gerando grupos de palavras que juntas representam tópicos que podem resumir os assuntos mais tratados em uma coleção de documentos. Para tanto foi utilizada a biblioteca **gensim**.

O modelo de LDA foi ajustado para encontrar 3 tópicos com 10 passadas pela coleção de documentos de cada faixa de cada serviço. É sabido que o LDA é uma abordagem supevisionada, e logo, não é possível saber de antemão a quantidade de tópicos e de palavras nos tópicos que melhor representará os assuntos em torno da coleção. Dessa forma, também foi utilizada a biblioteca **pyLDAvis** que gera visualizações navegáveis em HTML dos tópicos obtidos. (vide arquivos **lda.html** na pasta **out**)

Para fins de insights, foram incluídos abaixo uma breve lista das palavras dos tópicos obtidos.

Tópicos identificados pelo LDA para a Faixa 1 do serviço 17663:

- **Tópico 1**: 0.032*"servidores" + 0.029*"curso" + 0.013*"material" + 0.012*"periodo" + 0.008*"realizado" + 0.008*"atender" + 0.007*"participacao" + 0.006*"brasilia" + 0.006*"capacitacao" + 0.006*"abaixo"

- **Tópico 2**: 0.014*"servidores" + 0.010*"periodo" + 0.009*"curso" + 0.008*"congresso" + 0.007*"brasilia" + 0.007*"memo" + 0.007*"silva" + 0.006*"atender" + 0.006*"realizado" + 0.006*"maria"

- **Tópico 3**: 0.029*"curso" + 0.010*"especializacao" + 0.009*"conforme" + 0.008*"anexo" + 0.008*"ser" + 0.007*"atender" + 0.007*"especializada" + 0.006*"servidores" + 0.005*"aperfeicoamento" + 0.005*"profissional"

Tópicos identificados pelo LDA para a Faixa 2 do serviço 17663:

- **Tópico 1**: 0.023*"curso" + 0.015*"saude" + 0.015*"treinamento" + 0.010*"atender" + 0.010*"especializacao" + 0.009*"horas" + 0.008*"material" + 0.007*"ser" + 0.007*"capacitacao" + 0.006*"profissional"

- **Tópico 2**: 0.033*"curso" + 0.021*"transito" + 0.017*"tecnica" + 0.014*"elaboracao" + 0.014*"area" + 0.012*"destinado" + 0.011*"sistema" + 0.010*"aplicacao" + 0.010*"capacitacao" + 0.009*"gestores"

- **Tópico 3**: 0.025*"curso" + 0.015*"educacao" + 0.014*"graduacao" + 0.013*"ambiental" + 0.013*"material" + 0.011*"capacitacao" + 0.010*"instituicao" + 0.010*"processo" + 0.009*"pos" + 0.009*"servidores"

No geral os tópicos giram em torno da capactiação de servidores, mas na Faixa 2 é possível de fato observar temática **saúde** no 1º tópico, **gestores** no 2º e **ambiental** no 3º.

Tópicos identificados pelo LDA para a Faixa 1 do serviço 3239:

- **Tópico 1**: 0.030*"frete" + 0.028*"aldeia" + 0.023*"memo" + 0.021*"transporte" + 0.016*"barra" + 0.014*"percurso" + 0.013*"indios" + 0.012*"aerbgs" + 0.012*"ate" + 0.012*"sao"

- **Tópico 2**: 0.017*"transporte" + 0.013*"frete" + 0.009*"atender" + 0.009*"memo" + 0.008*"onibus" + 0.008*"conforme" + 0.007*"anexo" + 0.006*"materiais" + 0.006*"caramuru" + 0.005*"veiculos"

- **Tópico 3**: 0.036*"transporte" + 0.013*"veiculos" + 0.012*"frete" + 0.012*"indigenas" + 0.009*"atender" + 0.009*"materiais" + 0.006*"pessoas" + 0.006*"conforme" + 0.006*"aldeia" + 0.006*"especializada"

Tópicos identificados pelo LDA para a Faixa 2 do serviço 3239:

- **Tópico 1**: 0.060*"veiculos" + 0.046*"motorista" + 0.046*"locacao" + 0.046*"pecas" + 0.046*"manutencao" + 0.039*"nuraf" + 0.019*"atender" + 0.018*"ans" + 0.018*"categoria" + 0.018*"media"

- **Tópico 2**: 0.040*"transporte" + 0.036*"pequenas" + 0.029*"veiculos" + 0.026*"cargas" + 0.022*"pessoas" + 0.022*"porto" + 0.022*"ate" + 0.022*"motorista" + 0.019*"meses" + 0.019*"especializada"

- **Tópico 3**: 0.047*"diaria" + 0.035*"tipo" + 0.024*"veiculos" + 0.023*"transporte" + 0.023*"meses" + 0.023*"inca" + 0.023*"horas" + 0.021*"feira" + 0.021*"unidades" + 0.019*"anexo"

No caso do serivço 3239, é possível verificar na Faixa 1 de fato a ocorrência de tópicos que incluem as palavras **transporte** juntamente com **aldeia**, **índios** e **indígenas**, reforçando a importância do tema na coleção. Já na Faixa 2, esses assuntos não ocorrem.

Deteção de compras suspeitas

Com base nas análises anteriores, e assumindo ingenuamente que a análise dos textos das faixas de gasto de um serviço podem discriminar compras de alto valor de compras de baixo valor, sugere-se que um modelo de automático possa aprender a classificar uma compra quanto a sua faixa de gasto com base no seu texto, para em seguida classificar as compras coletadas e levantar suspeitas de comrpas que estão na Faixa 2 mas deveriam estar na Faixa 1.

Para tanto foi utilizado o classificador Random Forest com ajuda da classe **sklearn.ensemble.RandomForestClassifier**. Dado que temos 2000 features obtidas do vetorizador TF-IDF criado anteriormente, o classificador Random Forest foi testado e ajustado até apresentar uma alta acurácia, mas ainda sim gerando alguns falsos positivos para a Faixa 1, que serão tratados como suspeitos. A configuração configuração sugerido foi trabalhar com 5 árvores estimadoras e profundidade máxima 10 em cada árvore.

A acurácia foi verificada com validação cruzada de 5 *folds*, e em seguida toda a coleção foi classificada automaticamente. Os suspeitos foram listados em um arquivo separado, ordenados do maior valor de compra para o menor, além de oferecer o link para o detalhe da compra/licitação no site do governo. (vide o arquivo **suspeitas.txt** no diretório **out**)

Resultado da identificação de compras suspeitas para o serviço 17663 (Cursos de Especialização):

Classes:

- Faixa 1 (gasto até 437236.17) - 2716 registros.
- Faixa 2 (acima de 437236.17) - 49 registros.

Acurácias obtidas na validação cruzada com 5 folds: 98,19%, 98,19%, 98,01%, 98,19%, 98,36%.

Foram encontrada 41 suspeitas. Algumas delas:

- A compra #78 de valor 3723691.94 é da Faixa 2 mas parece ser da Faixa 1. (http://compras.dados.gov.br/compraSemLicitacao/doc/compra_slicitacao/20001206000012001)
- A compra #2237 de valor 3359457.66 é da Faixa 2 mas parece ser da Faixa 1. (http://compras.dados.gov.br/compraSemLicitacao/doc/compra_slicitacao/15308006000101999)
- A compra #2621 de valor 3054234.70 é da Faixa 2 mas parece ser da Faixa 1. (http://compras.dados.gov.br/compraSemLicitacao/doc/compra_slicitacao/25000506001022002)

A primeira suspeita acima, da compra #78, foi avaliada no site do governo e apresentou dados de fato confusos. Sua descrição indica *"prestação de serviços de limpeza e conservação prediais"*. Já sua justificativa alega "entidade Jurídica de direito privado, sem fins lucrativos, especializada para desenvolvimento dos cursos de qualificação". Ademais, a descrição do único item que compõe a compra é: "realização de 50 cursos, atendendo 11 Estados da Federação, para qualificação de 2.000 (dois mil)

Agentes Municipais de Trânsito, sendo 50 (cinquenta) turmas de 40 (quarenta) alunos cada." É importante destacar que pode também se tratar de um erro na base de dados.

Resultado da identificação de compras suspeitas para o serviço 3239 (Transporte Rodoviário de Pessoas):

Classes:

- Faixa 1 (gasto até 1822154.05) - 999 registros.
- Faixa 2 (acima de 1822154.05) - 12 registros.

Acurácias obtidas na validação cruzada com 5 folds: 98,52%, 98,52%, 99,00%, 99,00%, 99,00%

Foram encontrada 2 suspeitas, conforme segue:

- A licitacao #1063 de valor 43048801.96 é da Faixa 2 mas parece ser da Faixa 1. (http://compras.dados.gov.br/licitacoes/doc/licitacao/25005203000892000)
- A licitacao #1725 de valor 3260659.88 é da Faixa 2 mas parece ser da Faixa 1. (http://compras.dados.gov.br/licitacoes/doc/licitacao/25001502000012002)

A primeira suspeita acima, da licitação #1063, foi avaliada no site do governo. É uma compra do Instituto Nacional do Câncer (INCA). Sua descrição é *"Serviços de locação de diversos tipos de veículos, ambulâncias e caminhões, para atenderem às necessidades das diversas Unidades do INCA, conforme especificações constantes do edital."* Os item de compra dela são milhares de diárias de ambulâncias para transporte de pacientes. Um fato que pode ser observado nesse caso, é que outras variáveis devem ser levadas em conta pelo classificador de suspeitas além da informação textual. Uma delas é a quantidade e a unidade do item de compra, pois ela pode revelar dimensões importantes sobre a compra que está sendo feita, ajudando o classificador por exemplo a diferenciar uma compra cara e numerosa e uma compra barata de poucas unidades de um mesmo tipo de material/serviço.

Considerações Finais

Foi apresentado um sistema automatizado de coleta, extração, processamento e análise do texto das compras governamentais, sob a das faixas de menor e maior gastos de um determinado serviço. A execução do fluxo para dois serviços específicos foi realizada e os resultados foram apresentados para ilustrar o potencial analítico do ferramental proposto. Foi possível exercitar diversos assuntos pertinentes aos tema recuperação da informação e mineração de textos.

O sistema demonstra alto grau de automatização, bastanto definir qual serviço será avaliado no arquivo **constantes.py**, e executando as etapas do fluxo através do arquivo **main.py**. As respostas da coleta são armazenadas no diretório **cache**, os dados intermediários do fluxo são armazenados no diretório **data**, separados por serviço. Os documentos e figuras frutos da análise são armazenados no diretório **out** separados por serviço.

O trabalho pode ser expandido tanto com a coleta de dados de novos serviços, quanto com o aprimoramento do sistema. Algumas melhorias futuras incluem: fazer a coleta de materiais (hoje só coleta serviços), incluir novas dimensões na análise além do texto (ex: levar em conta a data, o local, dados do órgão que originou a compra, dados do fornecedor que ofereceu o serviço/material, etc.), além de ajustes finos nas estratégias definidas.

Fim