

Caracterização de eventos de exceção e de seus respectivos impactos no sistema de transporte público por ônibus da cidade de São Paulo

Felipe Cordeiro Alves Dias

Orientador: Prof. Dr. Daniel de Angelis Cordeiro

Universidade de São Paulo

19 de Março de 2019

Introdução

Motivação

- Segregação urbana: dentre os mais de 12 milhões de habitantes da cidade de São Paulo, **10% estão localizados no Centro Expandido (CE) e 90% no Cinturão Periférico (CP).**
 - Problemas relacionados a mobilidade urbana.
- Legislação federal e municipal sobre mobilidade urbana.
 - Lei Federal 12.587/2012: **para desenvolvimento sustentável com a mitigação dos custos ambientais e socioeconômicos dos deslocamentos de pessoas.**
 - Decreto 56.834: institui o *PlanMob/SP 2015* **como instrumento de planejamento e gestão do Sistema Municipal de Mobilidade Urbana para os próximos 15 anos.**

Motivação

- *PlanMob/SP 2015*
 - Criação da Central Integrada de Mobilidade Urbana (CIMU): com o objetivo de integrar as áreas de mobilidade urbana subordinadas à Secretaria Municipal de Transportes (SMT).
 - A CIMU não processa conteúdo de Redes Sociais;
 - não aborda melhorias dos sistemas já existentes;
 - será integrada com o defasado SIM.
- Sistema Integrado de Monitoramento e Transporte (SIM).
- Sistemas de Transporte Inteligente (ITS — *Intelligent Transport System*).

Motivação

- A lei de mobilidade urbana (12.587/2012) e o *PlanMob/SP 2015* não mencionam explicitamente ITS e TIC.
- O transporte público pode se beneficiar ao integrar as Redes Sociais com o planejamento, gestão e as atividades operacionais dos transportes públicos, abordando seus respectivos fatores sócio-técnicos, como:
 - Analisar o impacto dos eventos de exceção na operação do sistema de transporte público por ônibus na cidade de São Paulo.

Definição do problema

- Problemas que envolvem a caracterização dos impactos dos eventos de exceção:
 - Corpus SPTrans.
 - Processamento de grandes volumes de dados;
 - identificação de padrões;
 - qualidade dos dados comprometida.
 - Corpus Twitter.
 - Identificação dos eventos de exceção nas publicações;
 - extração de endereços e geolocalização;
 - correlação de eventos de exceção com a base histórica.

Objetivos

Gerais

Caracterização de eventos de exceção e de seus respectivos impactos no sistema de transporte público por ônibus da cidade de São Paulo.

Específicos

- Identificar os eventos de exceção, quando existentes, dos tweets coletados (**classificação automatizada**).
- Extrair os endereços dos eventos de exceção identificados e geolocalizá-los.
- Criação de plataforma para exploração e visualização dos dados coletados e processados do Twitter e da SPTTrans.

Hipóteses

- É possível identificar e categorizar os eventos de exceção de acordo com os tipos de eventos encontrados pela Revisão Sistemática.
- Extração de características com o auxílio de Processamento de Linguagem Natural (NLP — *Natural Language Processing*) em conjunto com dicionários auxiliares para o contexto dos eventos de exceção mencionados.
- Extração dos endereços dos *tweets* utilizando a técnica de Expressão Regular e posterior geolocalização.

Fundamentação Teórica

Definição

São cidades sustentáveis e socialmente inclusivas, que utilizam Tecnologias da Informação e Comunicação (TICs) para gerir eficientemente seus respectivos recursos naturais.

- Método com viés tecnológico.
- Método com viés humano.

Cidades Inteligentes

- As TICs permeiam os sistemas urbanos e espaços físicos: dados voluntários, de sensores e Redes Sociais.
- Desafios relacionados a conectividade:
 - Infraestrutura de rede.
 - interoperabilidade;
 - padrões;
 - consumo de energia;
 - escalabilidade, dentre outros.
- Desafios relacionados aos dados:
 - Capacidade e local de armazenamento;
 - extração;
 - tratamento;
 - processamento;
 - análise;
 - integração;
 - agregação de dados, dentre outros.

Definição

Tem como fim utilizar TICs para resolver problemas relacionados ao transporte, tais como congestionamento, segurança, eficiência e conservação ambiental.

- Localização de Veículos Autônomos (*Automatic Vehicles Location*).

General Transit Feed Specification

Definição

É uma especificação de um formato comum para troca de informações estáticas sobre transporte público.

- *agency.txt*: agências de transporte público como fonte de dados.
- *stops.txt*: locais de embarque e desembarque.
- *routes.txt*: trajetos de um grupo de viagens.
- *trips.txt*: viagens de cada trajeto.
- *stop_times.txt*: horários de partida e chegada em paradas.
- *calendar.txt*: início, fim e dias disponíveis dos serviços.

Definição

Explora como computadores podem ser utilizados para entender e manipular texto ou fala em linguagem natural, o que envolve conhecimento interdisciplinar.

- Problemas de baixo nível (comuns a NLP).
- Problemas de alto nível (específicos e com base nos problemas de baixo nível).

Feature Engineering

Definição

Processo iterativo de construção, extração e seleção de variáveis (*features*), o qual depende do conhecimento de domínio e de suas respectivas métricas.

- Variáveis (*features*) binárias, categóricas ou contínuas.
- Pré-processamento: técnicas de padronização, normalização, remoção de ruídos, redução de dimensionalidade, discretização, expansão, etc.

Algorítimos de Aprendizado de Máquina Supervisionados

- Árvore de Decisão
- Floresta Aleatória
- K-ésimo Vizinho mais próximo
- Máquina de Vetores de Suporte
- Naive Bayes
- Redes Neurais
- Regressão Logística

Definição

TF-IDF é um algoritmo de ponderação de variáveis que combina as ponderações *frequência do termo* (TF — *Term Frequency*) e *inverso da frequência nos documentos* (IDF — *Inverse Document Frequency*) para calcular os pesos dos termos linguísticos (variáveis) em um determinado corpus.

Algoritmo Apriori

- **Suporte (support)**: indicador da frequência de determinados registros no conjunto de dados.
- **Confiança (confidence)**: frequência com que determinadas regras de associações entre registros são encontradas como verdadeiras.
- **Lift**: probabilidade de ocorrência de um consequente B no conjunto de dados ($lift > 1$ indica que a regra de associação em questão pode ser utilizada para predição de um consequente B em conjuntos de dados futuros).

A notação $A \rightarrow B$ se refere a antecedente e consequente, respectivamente.

Revisão Sistemática

Quais os tipos de problemas urbanos abordados utilizando processamentos de *tweets*? (QP1)

- *e-Participation.*
- Detecção de zoneamento urbano.
- Identificação de pontos de interesse.
- Mobilidade.
- Padrões demográficos.
- Poluição.
- Segurança Pública.
- Turismo.
- Tráfego.

Casos de uso relacionados ao transporte público (QP2)

- Impacto de eventos no transporte público relacionados:
 - Aos ataques terroristas em Paris no uso do transporte público.
 - Ao tráfego na demanda por bicicletas, em Nova Iorque e Washington D.C.
 - Aos pontos de interesse na demanda por transporte público.
 - Aos eventos anormais nas tomadas de decisão dos passageiros do Metrô de Tóquio.
 - A predição de fluxo de passageiros no Metrô de Nova Iorque.
- Planejamento e gestão do transporte público.

Técnicas estatísticas utilizadas no processamento de tweets (QP3)

- Análise de métricas relacionadas a desempenho.
- Semelhança de cosseno.
- F_1 score.
- Frequência do termo-inverso da frequência nos documentos (TF-IDF).
- Coeficiente de variação inversa.
- Método de reamostragem Jackknife.
- Indicadores locais de associação espacial (LISA).
- Local Moran's.
- Máxima verossimilhança.
- Média móvel integrada autoregressiva sazonal (SARIMA).
- Otimização e previsão com função de perda híbrida.

Revisão Sistemática

Paradigmas de processamento (QP4)

- *Processamento em lote (offline).*
- *Processamento em quase tempo real.*
- *Processamento em tempo real.*

Eventos de exceção relacionados ao transporte público (QP5)

- **Acidentes.**
 - Acidentes nas estações transporte.
 - Incêndio.
- **Espaço-temporais.**
 - Dia da semana.
 - Hora do dia.

Revisão Sistemática

Eventos de exceção relacionados ao transporte público (QP5)

- Eventos sociais.

- Feiras de rua.
- Festivais.
- Jogos esportivos.
- Passeatas e maratonas.

- Eventos urbanos.

- Relacionados ao tráfego.

- Desastres naturais.

- Tempestades.
- Terremoto.
- Tufões.

- Metereológicas.

- Dia claro, nublado, chuvoso, nevando, com neblina.
- Temperatura do ar.

Técnicas de Aprendizado de Máquina utilizadas no processamento de tweets (QP6)

- **Classificação bayesiana.**
- Algoritmo C5.0.
- Campo aleatório condicional com Regressão Logística.
- Alocação latente de Dirichle (LDA).
- Regressão Linear.
- Simulação Monte Carlo.
- Técnica inovadora que utiliza fatorização tensorial (*PairFac*).
- **Floresta Aleatória.**
- **Máquina de Vetores de Suporte.**
- Mapas auto-organizados.

Dados abertos relacionados ao transporte público e eventos de exceção

Corpus SPTTrans

Tabela: Arquivos e número de registros especificados na GTFS pela SPTTrans

Nome do arquivo	Número de registros
<i>agency.txt</i>	1
<i>calendar.txt</i>	6
<i>fare_attributes.txt</i>	6
<i>fare_rules.txt</i>	5.400
<i>frequencies.txt</i>	39.625
<i>routes.txt</i>	291.634
<i>shapes.txt</i>	800.767
<i>stop_times.txt</i>	95.134
<i>stops.txt</i>	19.933
<i>trips.txt</i>	2.273
Total	1.254.779

Tabela: Descrição do conjunto de dados AVL

Mês	Intervalo (dias)	Total de arquivos AVL	Espaço em disco (GB)
Janeiro ^a	1 - 31	744	102,44
Fevereiro	1 - 28	672	93,21
Março	1 - 31	744	102,64
Abril	1 - 30	720	97,04
Maio	1 - 31	744	101,46
Junho	1 - 30	720	97,13
Julho	1 - 31	744	104,95
Agosto	1 - 31	744	108,38
Setembro	1 - 30	720	109,89
Outubro	1 - 31	744	110,92
Novembro	1 - 30	717	108,16
Dezembro	1 - 31	738	110,89
Total	—	8.751	1.247,09

^a Arquivos Movto_20170111000_20170111100 com 35 campos na linha 60.025 e Movto_201701110900_201701111000 com 21 campos na linha 1.075.548, o esperado são 19 campos de acordo com os metadados fornecidos pela SPTrans.

Corpus Twitter

Tabela: Intervalo de tempo e número de tweets coletados

Perfil no Twitter	Total de tweets ^a	Timestamp 1 ^b	Timestamp 2 ^c
@BombeirosPMESP	6.632	2017-05-21	2017-12-01
@CETSP_	5.735	2017-02-20	2017-12-01
@CPTM_oficial	6.301	2017-04-24	2017-12-01
@governosp	6.011	2017-05-10	2017-12-01
@metrosp_oficial	8.621	2017-06-07	2017-12-01
@Policia_Civil	3.417	2015-04-15	2017-11-30
@PMESP	4.365	2016-06-02	2017-11-30
@saopaulo_agora	3.960	2016-11-18	2017-11-30
@smtsp_	1.316	2017-04-26	2017-12-01
@SPCEDEC	1.301	2015-06-09	2017-12-01
@sptrans_	9.956	2017-06-13	2017-12-01
@TurismoSaoPaulo	3.369	2012-06-12	2017-11-29
Total	60.984	—	—

^a Número de tweets coletados.

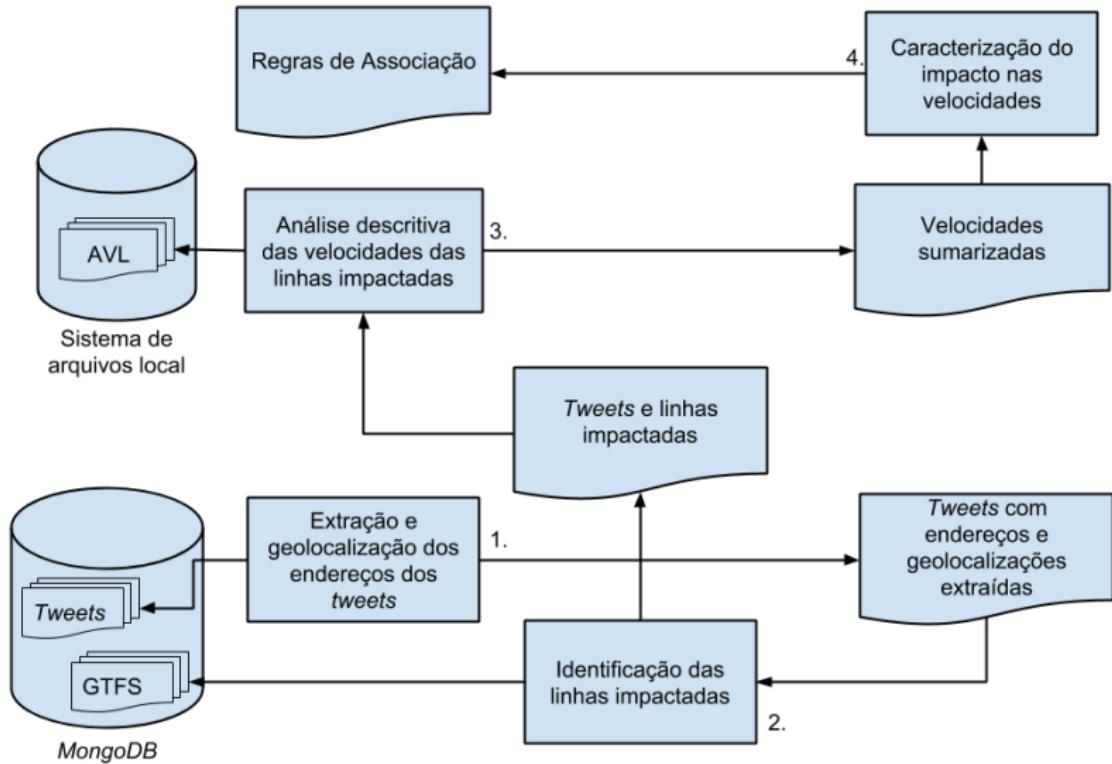
^b Timestamp mais antigo.

^c Timestamp mais recente.

Pré-processamento

- *Case folding*: processamento de normalização de todas as letras do texto (de A-Z) para minúsculas.
- Remoção de *URLs* e menções a outros *tweets*.
- Remoção de acentos, *emoticons* e pontuações substituídas por espaços vazios.
- *Stemming*.

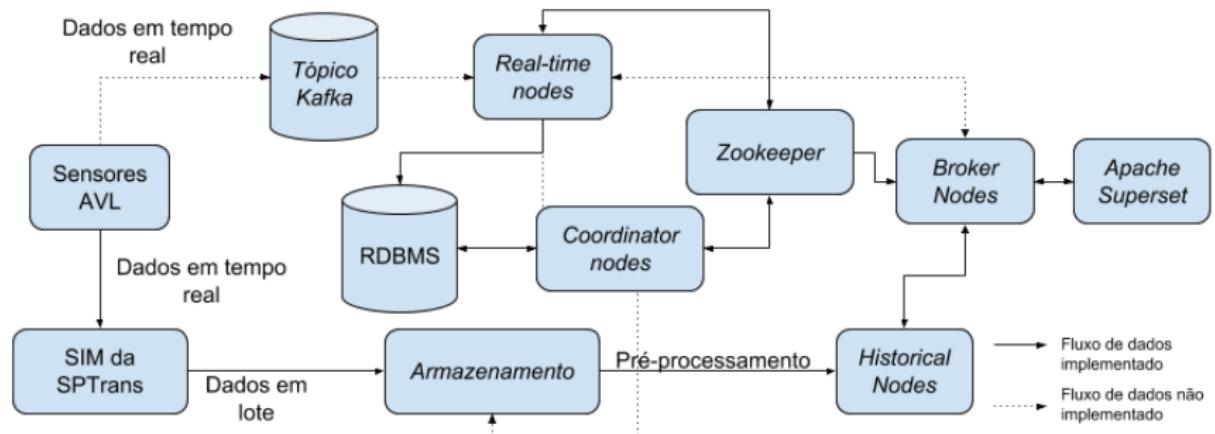
Correlação entre os tweets, dados AVL e GTFS da SPTrans



Exploração e visualização de grandes volumes de dados

Arquitetura proposta

Figura: Arquitetura utilizada no estudo de caso



Arquitetura proposta

Druid

- O Druid é um banco de dados para análises exploratórias em tempo real (latências abaixo de sub-segundos) em grandes conjuntos de dados.
- Arquitetura distribuída composta por um cluster com diferentes tipos de nós (real-time, historical, broker e coordinator nodes).
- Nós independentemente uns dos outros e possuem interação mínima entre eles.

Arquitetura proposta

Real-time nodes

Real-time nodes

- Ingerir, indexar e consultar fluxos de eventos.
- Índices localmente persistidos são mesclados em **blocos imutáveis** de dados com todos os eventos ingeridos em um período de tempo.
- **Não há perda de dados e a imutabilidade dos blocos permite a consistência de leitura e um modelo de paralelização simples:** *historical nodes* podem simultaneamente examinar e agregar blocos imutáveis de forma não bloqueante.

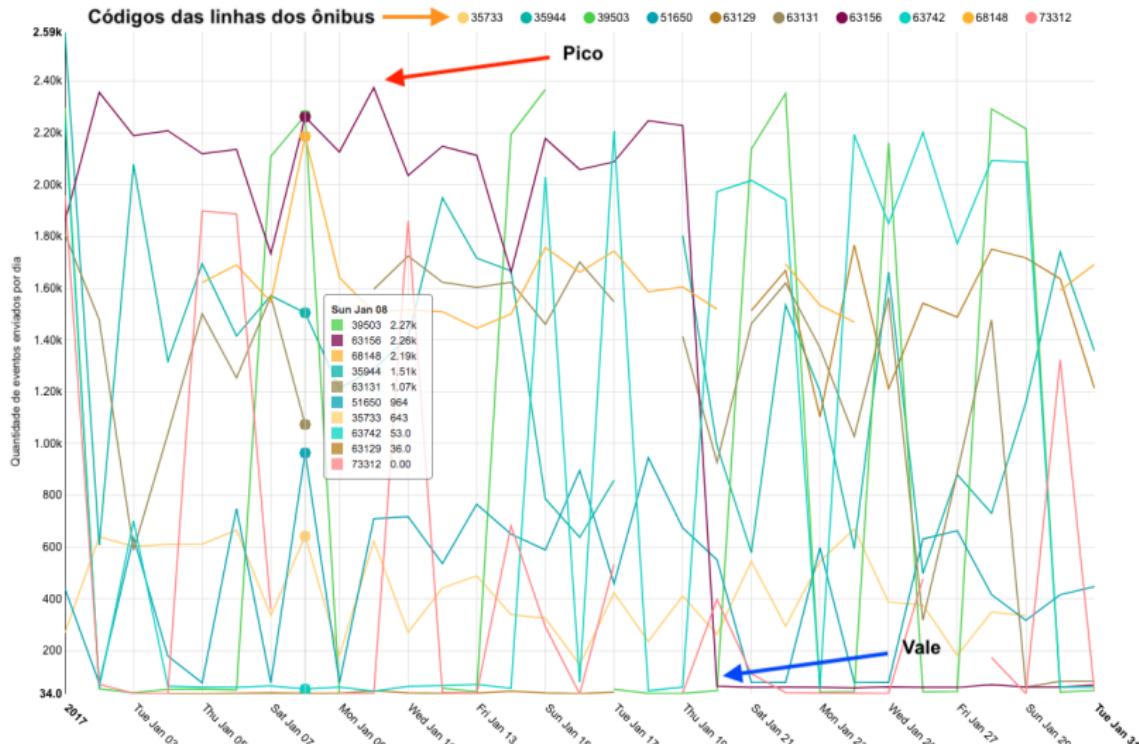
Arquitetura proposta

Historical, broker e coordinator nodes

- **Historical nodes:** são responsáveis por carregar, descartar e servir segmentos imutáveis.
- **Broker nodes:** são responsáveis por receber consultas e mesclar resultados parciais dos historicals e real-time nodes antes de retornar um resultado final consolidado para o cliente.
- **Coordinator nodes:** são responsáveis pelo gerenciamento e distribuição dos dados nos historical nodes, exigindo destes o carregamento, descarte e replicação dos dados.

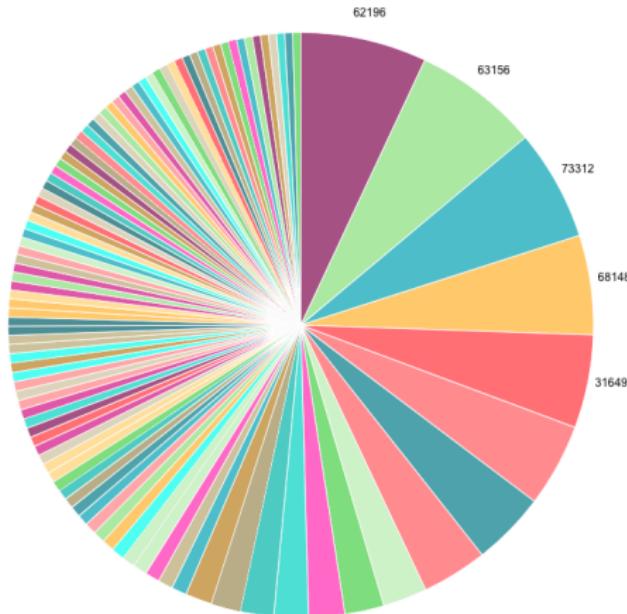
Validação da arquitetura proposta

Quantidade de dados enviados por dia por ônibus (selecionados aleatoriamente) em janeiro de 2017



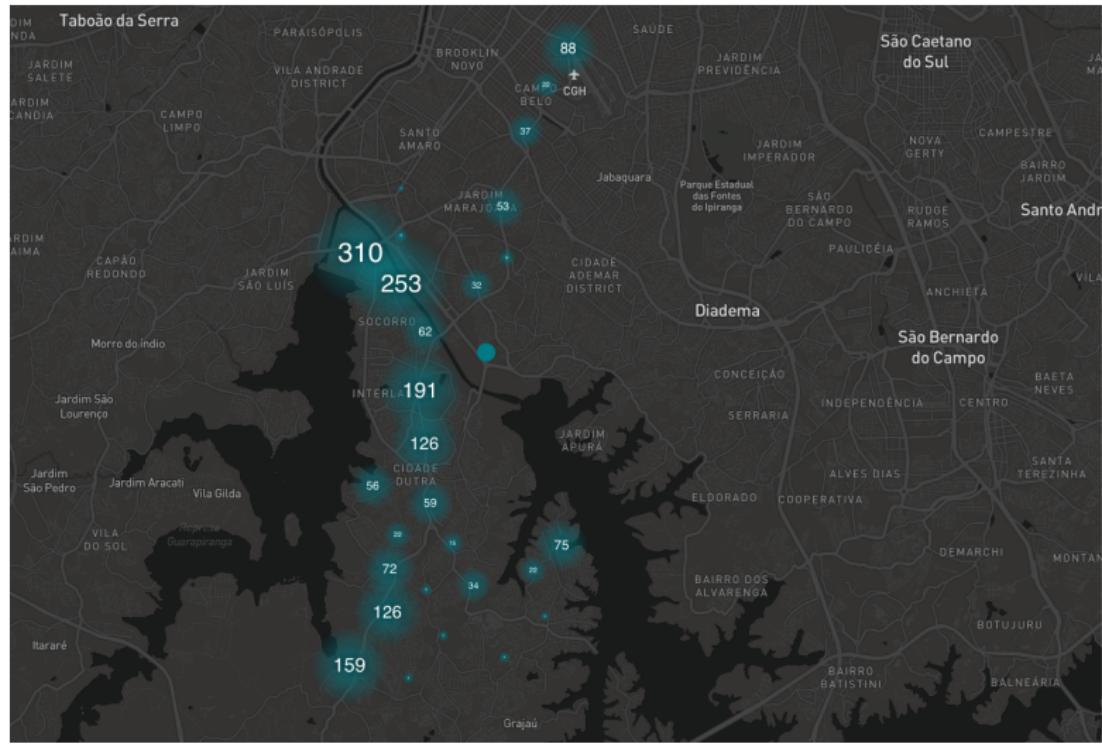
Validação da arquitetura proposta

Distribuição da quantidade de dados enviados por ônibus (selecionados aleatoriamente) em janeiro de 2017



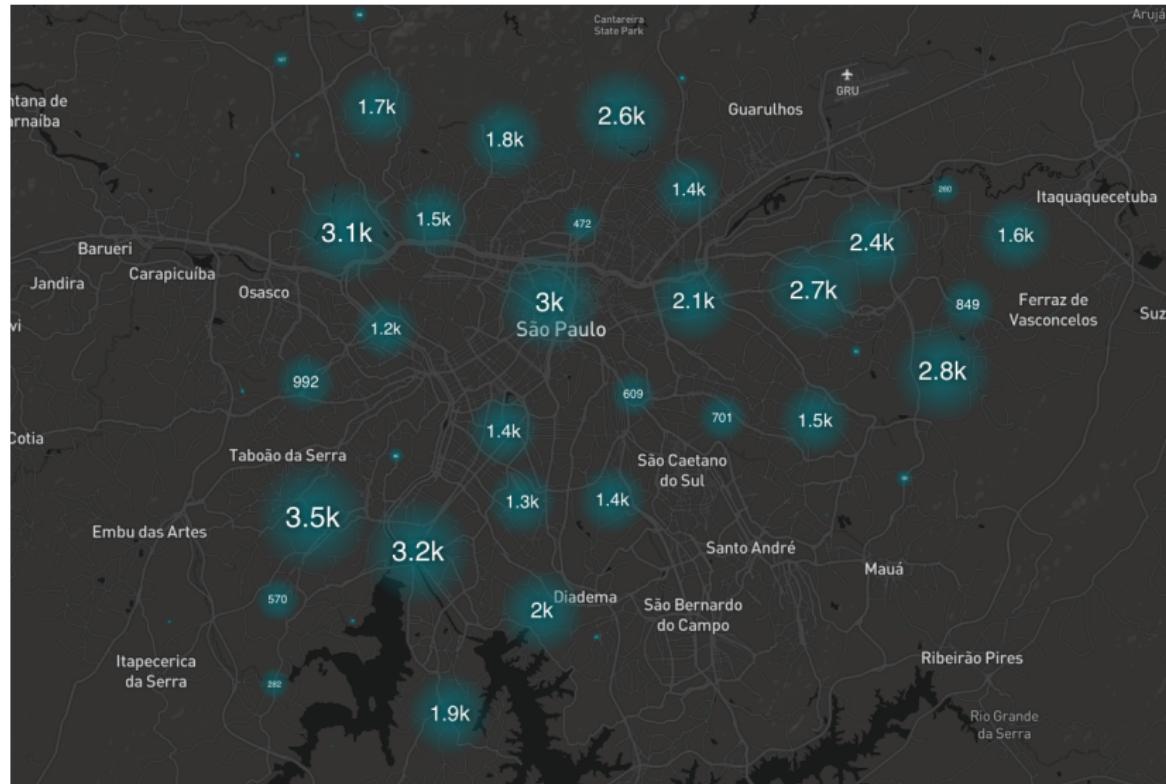
Validação da arquitetura proposta

Localizações enviadas em Janeiro de 2017 de uma linha de ônibus selecionada aleatoriamente



Validação da arquitetura proposta

Localizações dos ônibus referente a movimentação de Janeiro de 2017

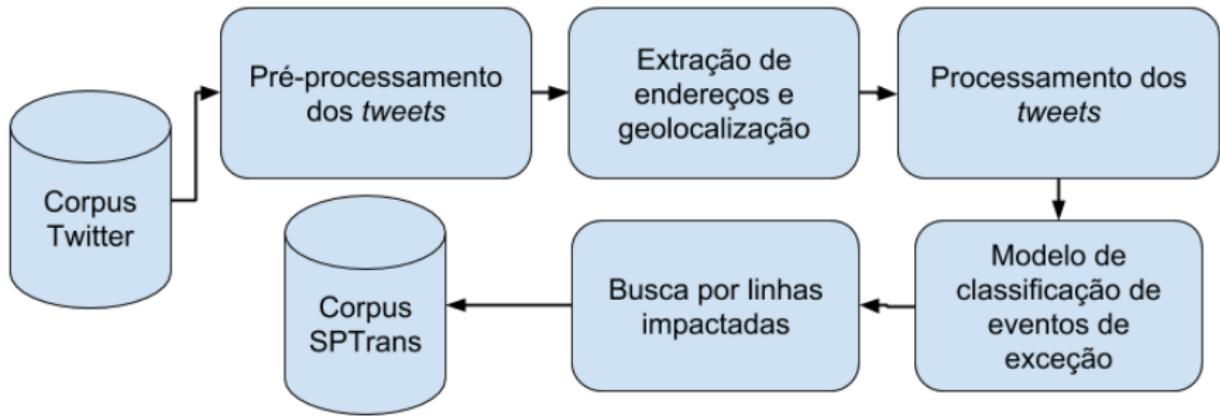


Consideração sobre a arquitetura utilizada para exploração e visualização dos dados AVL da SPTrans

- Estudo de caso relacionado à visualização de grandes conjuntos de dados, utilizando dados dos ônibus da cidade de São Paulo.
- Mostramos que é possível encontrar padrões complexos e incomuns e possíveis eventos de exceção em grandes conjuntos de dados por meio da visualização.
- O Druid e o Apache Superset demonstraram suporte a agregação, exploração e visualização de grandes conjuntos de dados.

Mineração e geolocalização automatizada de eventos de exceção a partir de dados do *Twitter*

Mineração e geolocalização automatizada de eventos de exceção a partir de dados do Twitter



Expressão regular para extração de endereços:

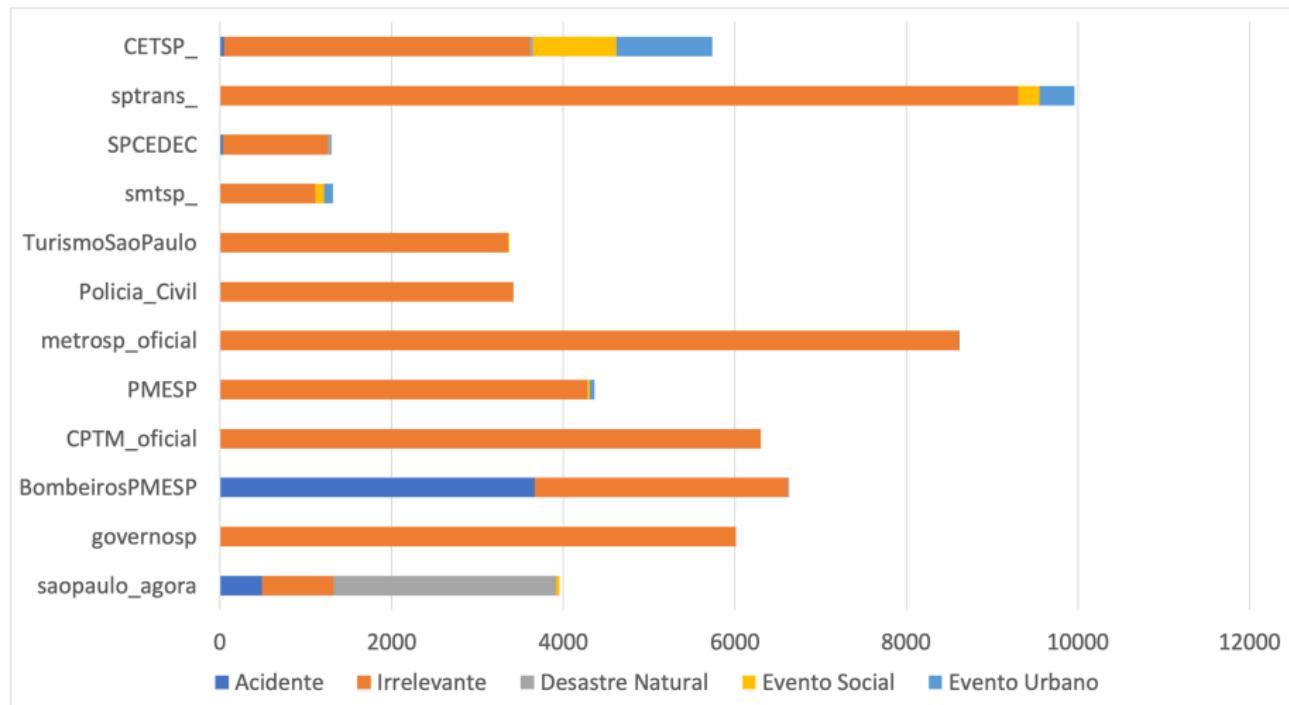
$$ER = \{L_1|S_1|L_2|S_2|\dots|L_n|S_n\}\{[a-z\dot{A}-\ddot{Y}_-]+\} \quad (1)$$

Geolocalização dos endereços usando a API do Google Geocoding.

Resultados da classificação, manual, pré-processamento, processamento dos *tweets* e extração de endereços

- Pré-processamento e processamento dos tweets: Corpus com **414.637 palavras**, vocabulário com **13.915 palavras**, comprimento máximo das sentenças do conjunto de dados com **19 caracteres**.
- **60.984 tweets** classificados manualmente.
- **10.027 tweets** classificados manualmente em eventos de exceção e desse subconjunto foram encontrados **8.112 endereços**.

Resultado da distribuição das classes dos eventos de exceção do Corpus Twitter

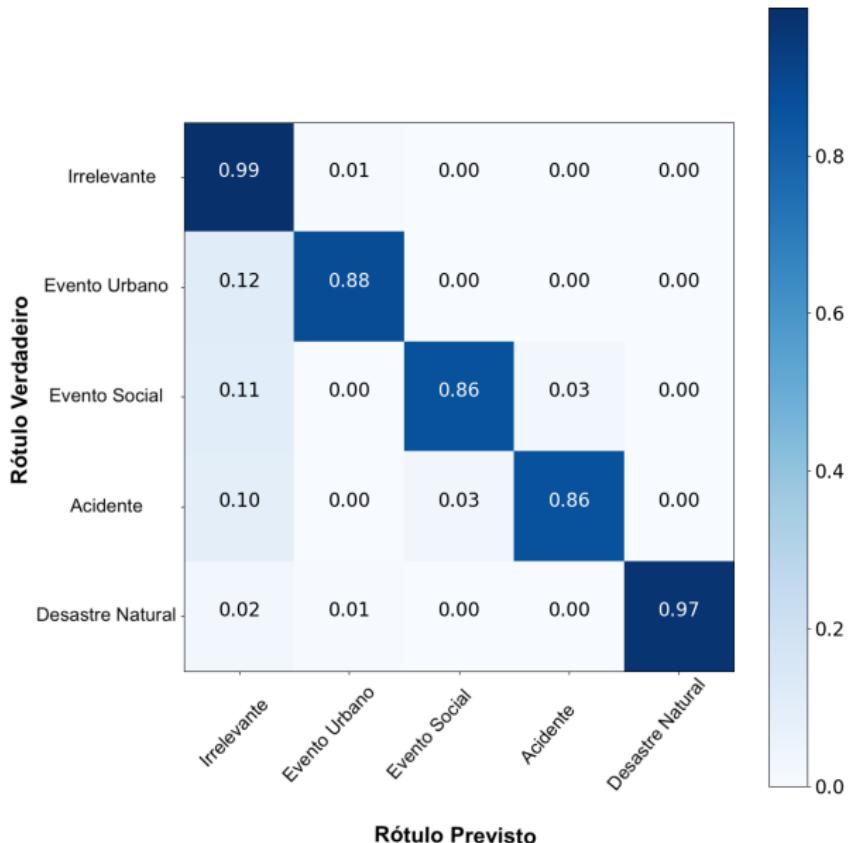


Resultados dos modelos para classificação automatizada dos eventos de exceção

Tabela: Métricas das avaliações dos algoritmos utilizados para classificação dos tweets em eventos de exceção

Algoritmo	ACC	PPV	TPR	f1-score
Naive Bayes Complementar	0,941	0,949	0,941	0,944
Árvore de Decisão	0,965	0,965	0,965	0,965
K-ésimo Vizinho mais Próximo	0,970	0,971	0,970	0,970
Regressão Logística	0,969	0,968	0,969	0,968
Perceptron multicamadas	0,973	0,972	0,973	0,972
Naive Bayes Multinomial	0,953	0,952	0,953	0,949
Floresta Aleatória	0,970	0,970	0,970	0,970
Máquina de Vetores de Suporte	0,833	0,694	0,833	0,757

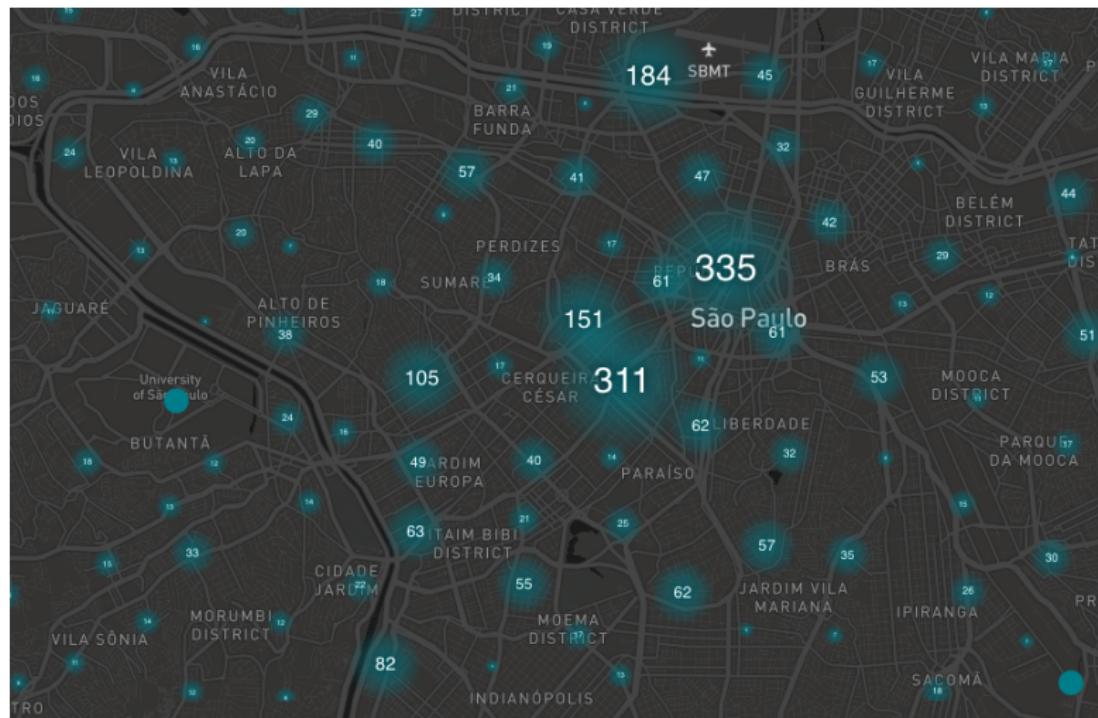
Resultados da MC do modelo Perceptron multicamadas



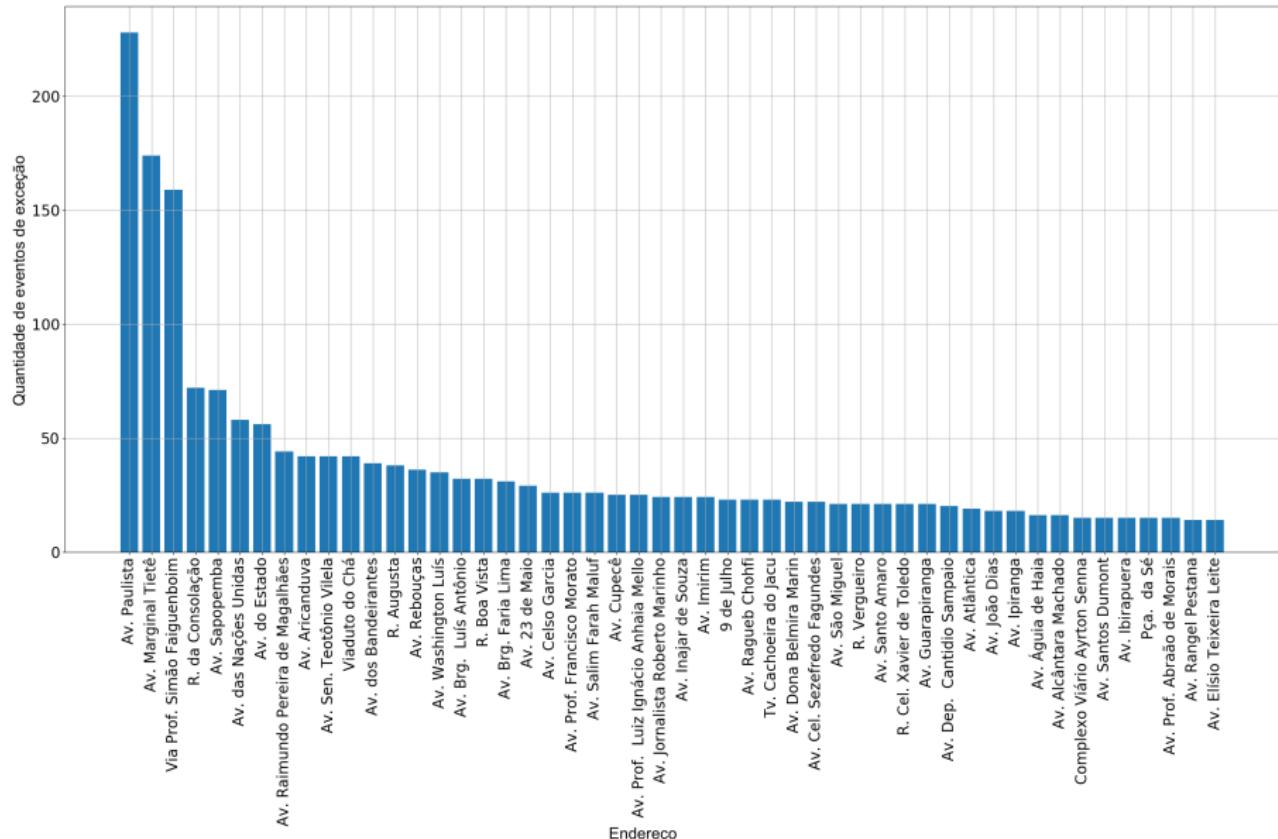
Resultados da distribuição de endereços extraídos por classe

- ① Tweets apenas com o ponto de interesse, ou seja, não consta explicitamente o endereço.
- ② Tweets sem informação de endereço.
- ③ Tweets com nome de logradouro incomum (por exemplo *passagem, complexo viário, ligação sentido*).
- ④ Tweets com endereços com palavras concatenadas (por exemplo *avenidapaulista*).

Resultado da análise visual da distribuição dos eventos de exceção na região central de São Paulo



Resultado dos endereços mais impactados



Resultados das linhas de ônibus mais impactadas por eventos de exceção

Tabela: Linhas de ônibus mais impactadas por eventos de exceção^a

Código da linha	# eventos de exceção	Letreiro
33389	1301	TERM. PINHEIROS / METRÔ TUCURUVI
33284	1176	ITAIM BIBI / METRÔ SANTANA
33121	1023	TERM. PRINC. ISABEL / TERM. STO. AMARO
32805	1006	TERM. PRINC. ISABEL / CHÁC. SANTANA
33112	933	TERM. PQ. D. PEDRO II / JD. SÃO SAVÉRIO

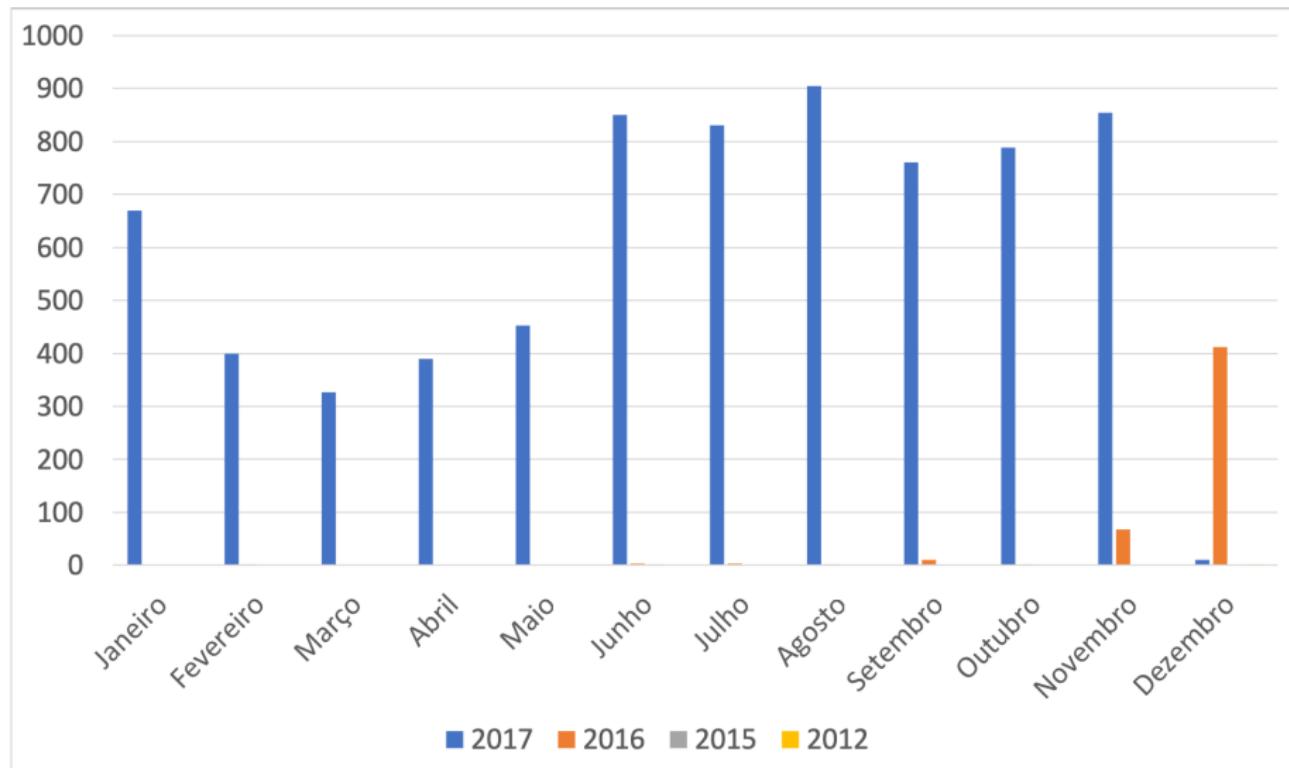
^a Tabela completa no Apêndice D.

Considerações finais sobre a metodologia desenvolvida

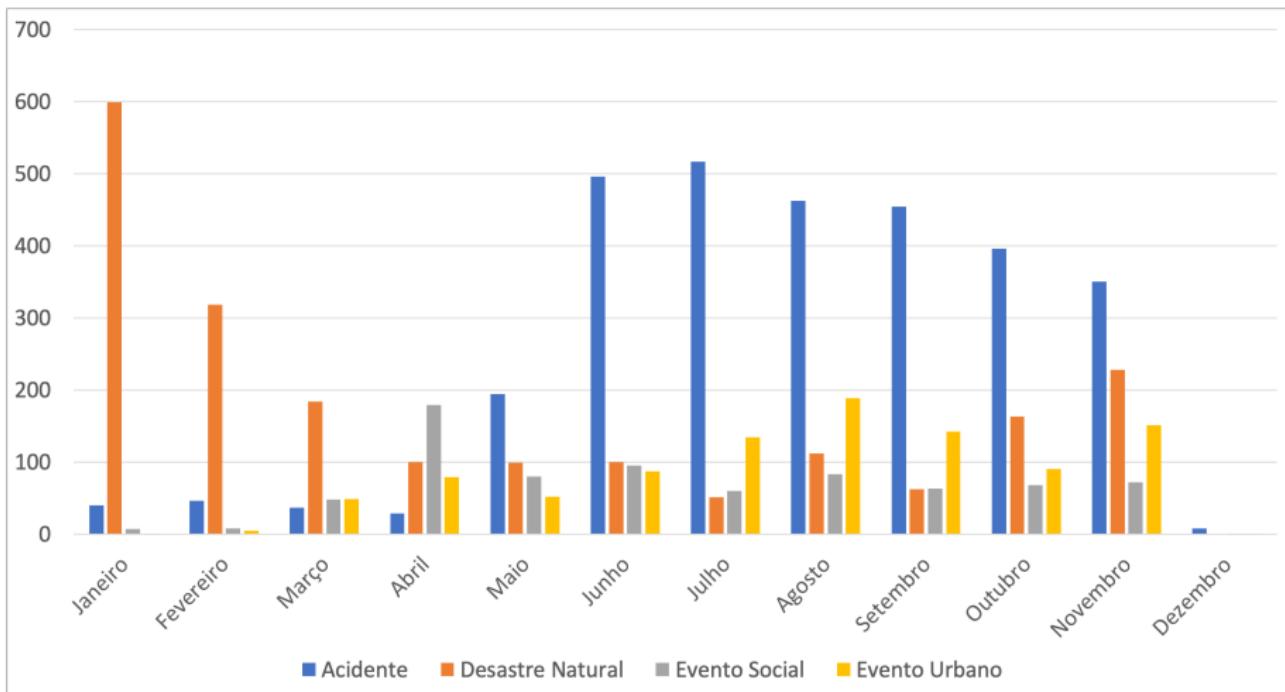
- Uma nova metodologia para classificação de eventos de exceção e analisa seus respectivos impactos no sistema de transporte coletivo por ônibus da cidade de São Paulo.
- O algoritmo *Multi-layer Perceptron* obteve **0,973 de acurácia** para classificação de *tweets* em eventos de exceção.
- É possível extrair endereços de *tweets* semi-estruturados usando apenas expressões regulares.
- A **classificação automatizada** desses eventos é o primeiro passo para entender melhor como os eventos de exceção afetam a rede de transporte público.
- Metodologia aplicável em **diferentes idiomas e cidades** (a GTFS é um formato ubíquo para o transporte público e ferramentas como a **NLTK suporta vários idiomas**).

Caracterização do impacto dos eventos de exceção

Distribuição do número de eventos de exceção geolocalizados ao longo dos meses dos anos 12/15/16 e 17

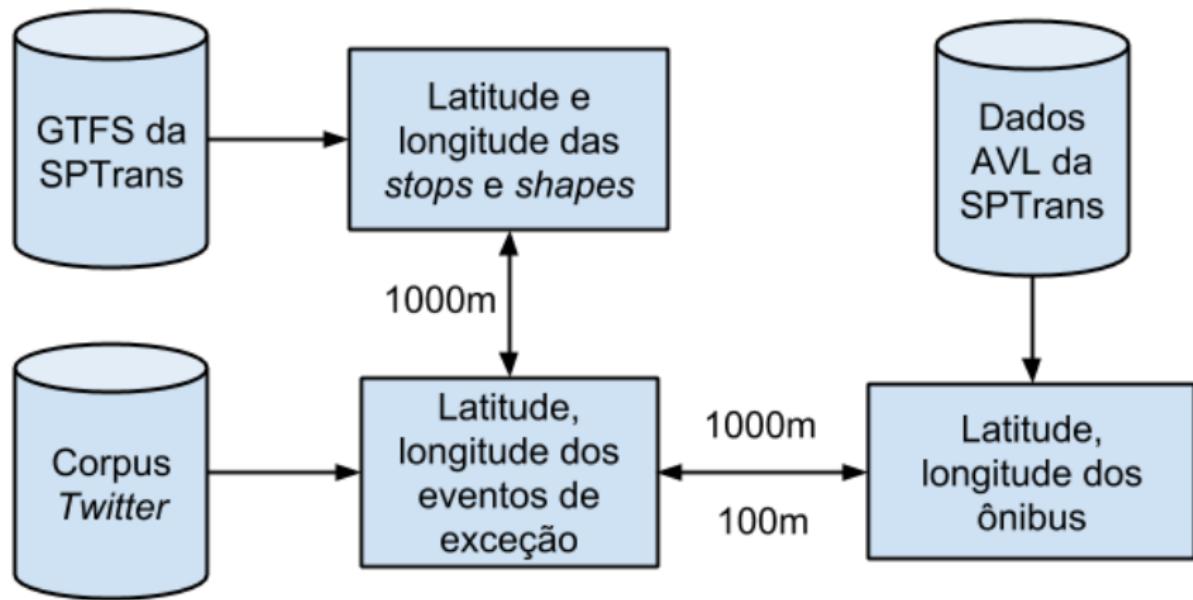


Distribuição das classes de eventos de exceção geolocalizados ao longo dos meses do ano de 2017



Processo para correlação entre os dados AVL, GTFS e tweets para análise do impacto dos eventos de exceção

Figura: Processo para correlação entre os dados AVL, GTFS e tweets para análise do impacto dos eventos de exceção



Equação utilizada para identificar velocidade mediana esperada

$$f(n) = \begin{cases} 0 & \text{se vel. mediana do dia do evento} > \frac{\text{vel. mediana dos dias da semana}}{\text{total de vel. medianas}} \\ 1 & \text{se vel. mediana do dia do evento} \leq \frac{\text{vel. mediana dos dias da semana}}{\text{total de vel. medianas}} \end{cases} \quad (2)$$

Resultados da caracterização dos impactos em relação às velocidades medianas dos ônibus

Tabela: Porcentagem de ônibus dos grupos de linhas afetadas por eventos de exceção, a 1.000 m e 100 m de distância a partir dos pontos de parada, respectivamente, que tiveram a velocidade mediana reduzida nos meses do ano de 2017

Mês	Acidente		Desastre Natural		Evento Social		Evento Urbano	
	1.000 m	100 m	1.000 m	100 m	1.000 m	100 m	1.000 m	100 m
Janeiro	83,33	100	64,23	98,00	100	—	100	—
Fevereiro	70,58	100	66,25	100	100	100	80	—
Março	50,00	—	66,66	100	85,00	100	68,18	100
Abril	87,50	100	61,11	100	82,75	100	76,92	100
Maio	65,13	100	58,82	100	93,33	100	50,00	100
Junho	54,46	100	61,53	100	76,47	100	72,41	100
Julho	61,48	98,41	66,66	100	69,23	100	58,13	100
Agosto	57,86	87,17	55,35	100	85,54	100	68,10	90,90
Setembro	64,21	100	42,10	100	92,30	100	62,06	100
Outubro	70,49	—	56,81	—	80,00	—	61,11	—
Novembro	66,66	100	57,99	100	92,85	100	74,35	100
Dezembro	—	—	—	—	—	—	—	—
Total	66,51	98,39	59,77	99,80	87,04	100	70,11	98,86

Resultados da caracterização dos impactos em relação às velocidades medianas dos ônibus

Tabela: Porcentagem de impacto na velocidade média dos grupos de linhas afetadas por eventos de exceção a 1.000 m e 100 m de distância dos pontos de rota, respectivamente, nos meses do ano de 2017

Mês	Acidente		Desastre Natural		Evento Social		Evento Urbano	
	1.000 m	100 m	1.000 m	100 m	1.000 m	100 m	1.000 m	100 m
Janeiro	66,66	100	47,68	78,49	100	100	100	—
Fevereiro	35,29	100	49,09	81,25	100	100	40,00	100
Março	66,66	100	42,85	62,5	90,00	72,22	50,00	53,84
Abril	62,50	60,00	47,05	100	76,11	77,27	89,47	90,90
Maio	49,09	77,77	64,70	100	73,33	80,00	40,00	50,00
Junho	47,78	79,76	46,15	70,00	61,76	61,29	72,41	77,77
Julho	44,85	75,55	66,66	83,33	48,14	75,00	41,86	61,53
Agosto	49,49	75,36	44,44	71,42	72,72	72,72	70,00	56,75
Setembro	49,47	79,16	36,84	54,54	76,92	58,33	55,17	73,91
Outubro	56,06	78,26	58,69	90,00	90,00	75,00	55,00	60,00
Novembro	54,32	66,66	44,00	74,07	85,71	85,71	67,50	72,97
Dezembro	—	—	—	—	—	—	—	—
Total	52,92	81,13	49,83	78,69	79,51	77,95	68,14	69,76

Trabalhos relacionados a identificação de padrões de velocidade média dos dados AVL

- Identificar **padrões relacionados a transferência (entre metrô e ônibus)**, por meio dos dados dos cartões inteligentes usados no transporte público da China.
- Identificar e apresentar visualmente **padrões de movimentação humana**, no transporte público de Singapura.
- Identificar **padrões de rotas de táxi**, na cidade de Pequim, China.
- Identificar **anomalias no comportamento do transporte público rodoviário** da cidade do Rio de Janeiro.

Identificação de padrões de velocidade média dos dados AVL

A proposta desse trabalho se diferencia das demais por encontrar os padrões de velocidade média existentes nos dados do transporte público por ônibus da cidade de São Paulo, considerando ainda a correlação com eventos de exceção extraídos de Redes Sociais.

Resultados da identificação de padrões de velocidade média dos dados AVL

Tabela: Análise *Apriori* aplicada as velocidades médias (intervalos de 5 minutos) ao conjunto de dados AVL da SPTrans

Mês	Regra de associação	Support	Confidence	Lift
Fevereiro	$7 \rightarrow 8$	0,101	0,496	3,586
Abril	$7 \rightarrow 8$	0,108	0,456	3,188
Maio	$7 \rightarrow 8$	0,108	0,570	4,375
Outubro	$8 \rightarrow 7$	0,100	0,595	3,433
Novembro	$8 \rightarrow 7$	0,104	0,446	3,369
Janeiro	$11 \rightarrow 12$	0,137	0,476	1,729
Junho	$11 \rightarrow 12$	0,129	0,632	1,656
Julho	$11 \rightarrow 12$	0,204	0,694	1,934
Agosto	$11 \rightarrow 12$	0,169	0,670	1,662
Outubro	$11 \rightarrow 12$	0,119	0,601	1,669

Resultados da identificação de padrões de velocidade média dos dados AVL

Tabela: (Continuação) Análise *Apriori* aplicada as velocidades médias (intervalos de 5 minutos) ao conjunto de dados AVL da SPTrans

Mês	Regra de associação	Support	Confidence	Lift
Fevereiro	12 → 11	0,126	0,582	1,770
Março	12 → 11	0,134	0,621	1,627
Abri	12 → 11	0,123	0,601	2,013
Maio	12 → 11	0,137	0,645	1,703
Setembro	12 → 11	0,163	0,608	1,863
Novembro	12 → 11	0,154	0,531	1,875
Dezembro	12 → 11	0,143	0,432	2,073
Fevereiro	12 → 13	0,123	0,375	1,956
Março	12 → 13	0,158	0,415	1,766
Junho	12 → 13	0,141	0,370	1,907

Resultados da identificação de padrões de velocidade média dos dados AVL

Tabela: (Fim da continuação) Análise *Apriori* aplicada as velocidades médias (intervalos de 5 minutos) ao conjunto de dados AVL da SPTrans

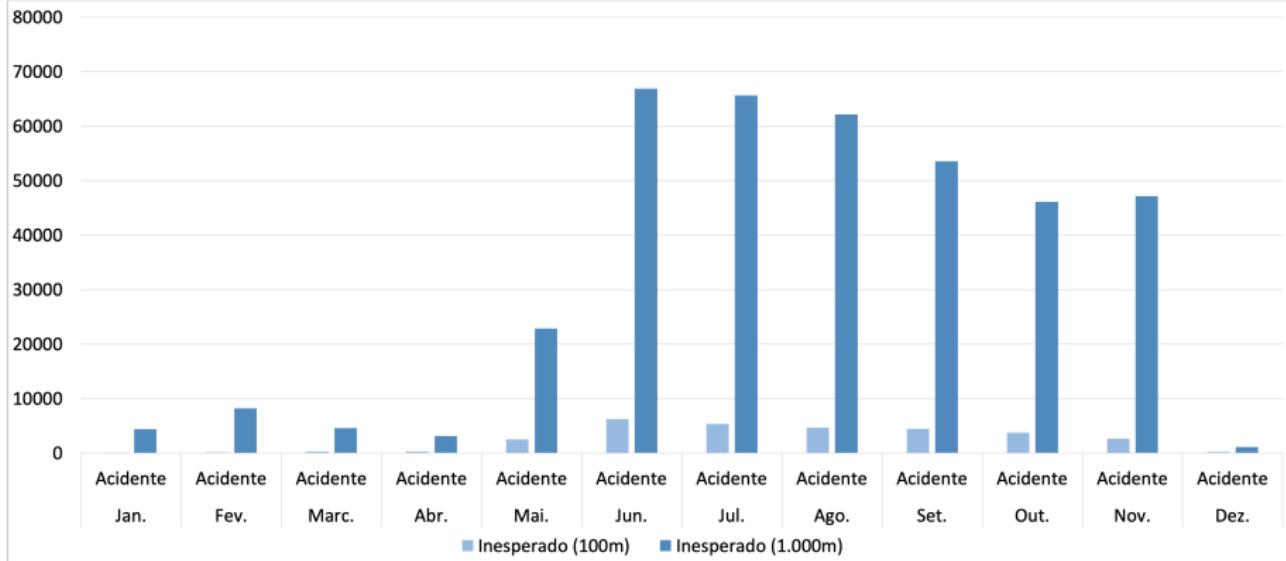
Mês	Regra de associação	Support	Confidence	Lift
Abril	13 → 12	0,109	0,367	2,280
Maio	13 → 12	0,161	0,425	1,942
Agosto	13 → 12	0,147	0,366	1,830
Outubro	13 → 12	0,150	0,417	1,737

Resultados das velocidades
médias inesperadas correlacionadas
aos eventos de exceção

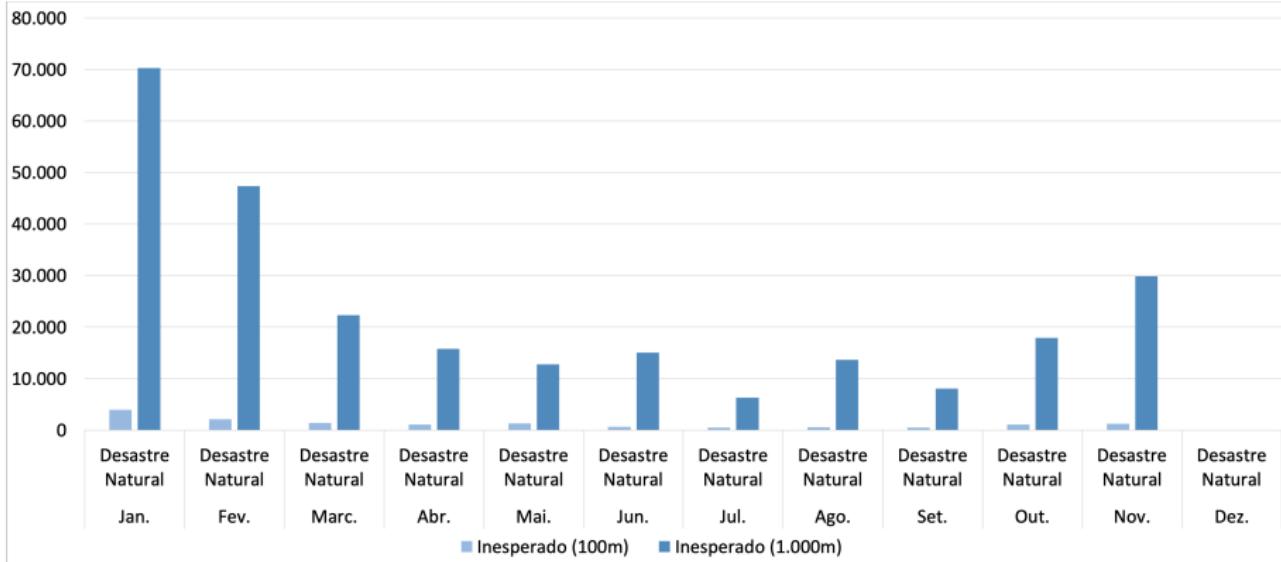
Ref. aos **pontos de parada**

Ao longo dos meses do ano de 2017

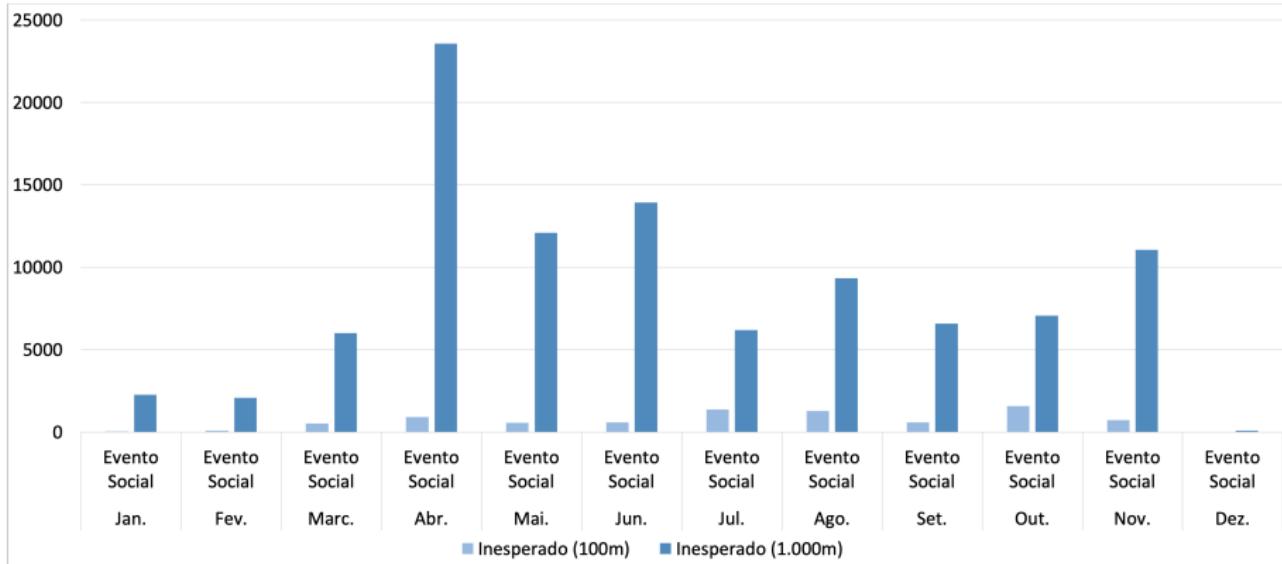
Velocidades inesperadas relacionadas a acidentes



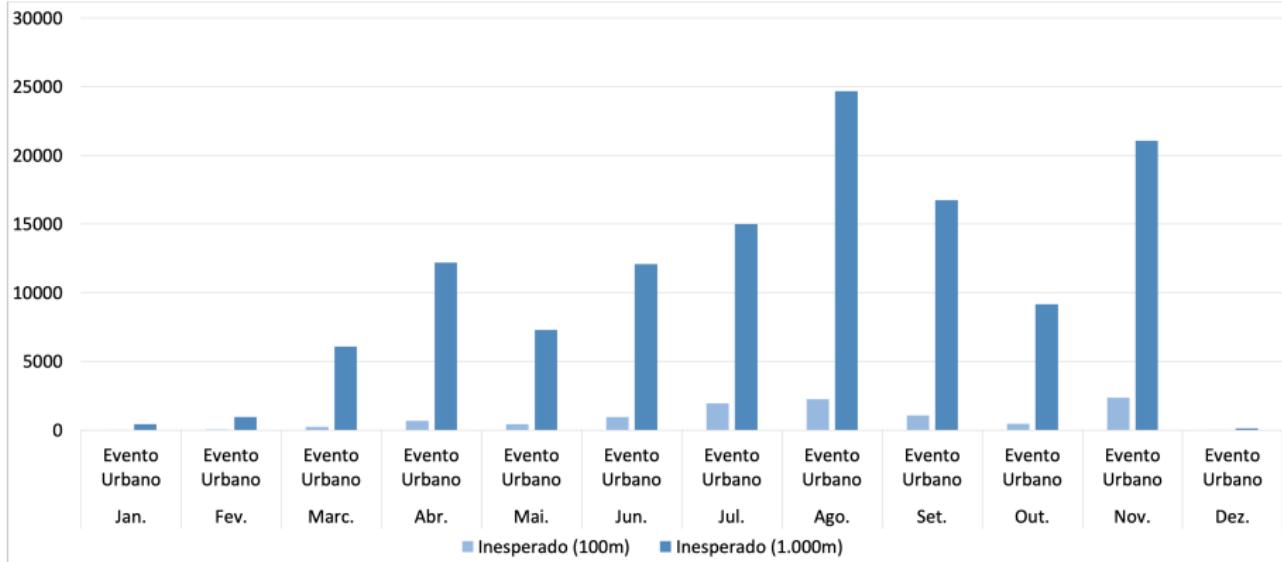
Velocidades inesperadas relacionadas aos desastres naturais



Velocidades inesperadas relacionadas aos eventos sociais



Velocidades inesperadas relacionadas aos eventos urbanos

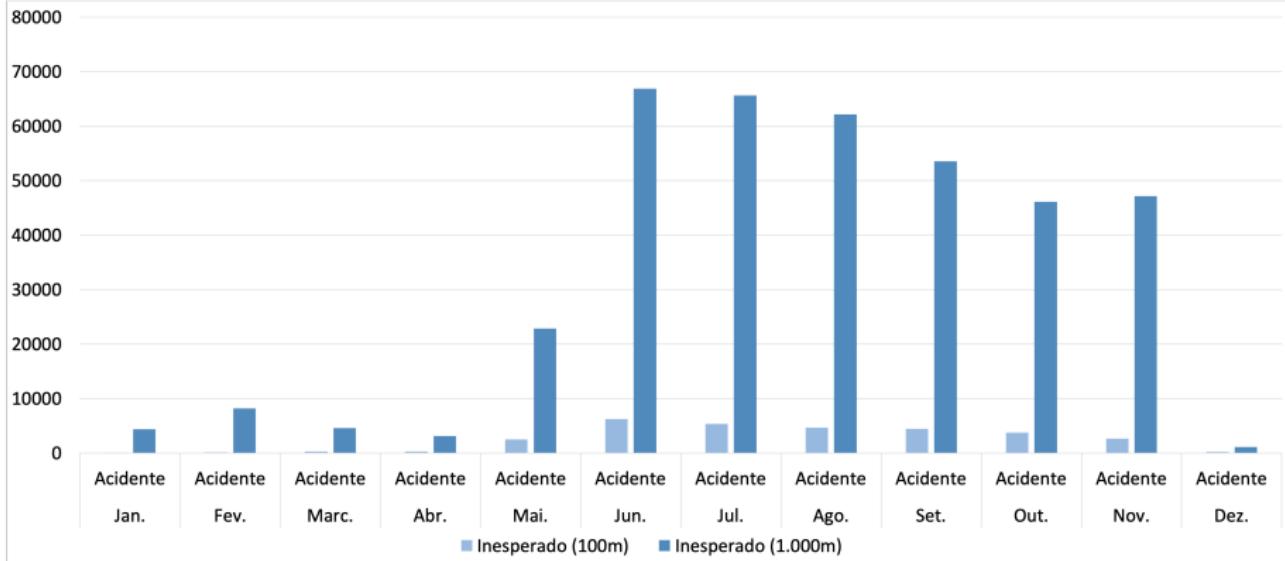


Resultados das velocidades
médias inesperadas correlacionadas
aos eventos de exceção

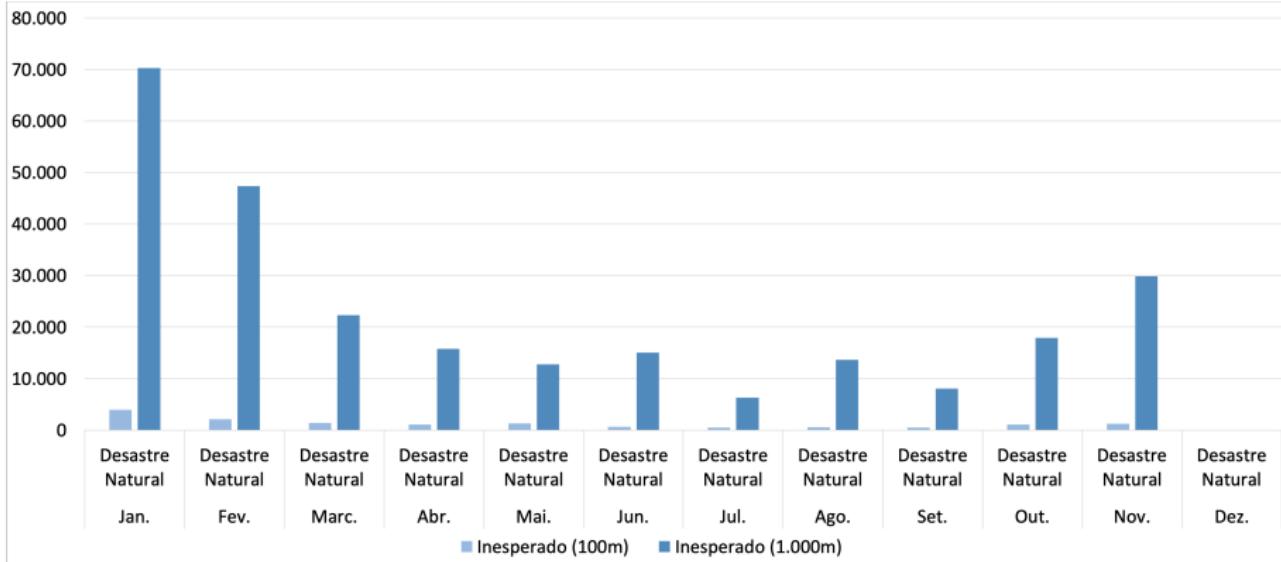
Ref. aos pontos de rota

Ao longo dos meses do ano de 2017

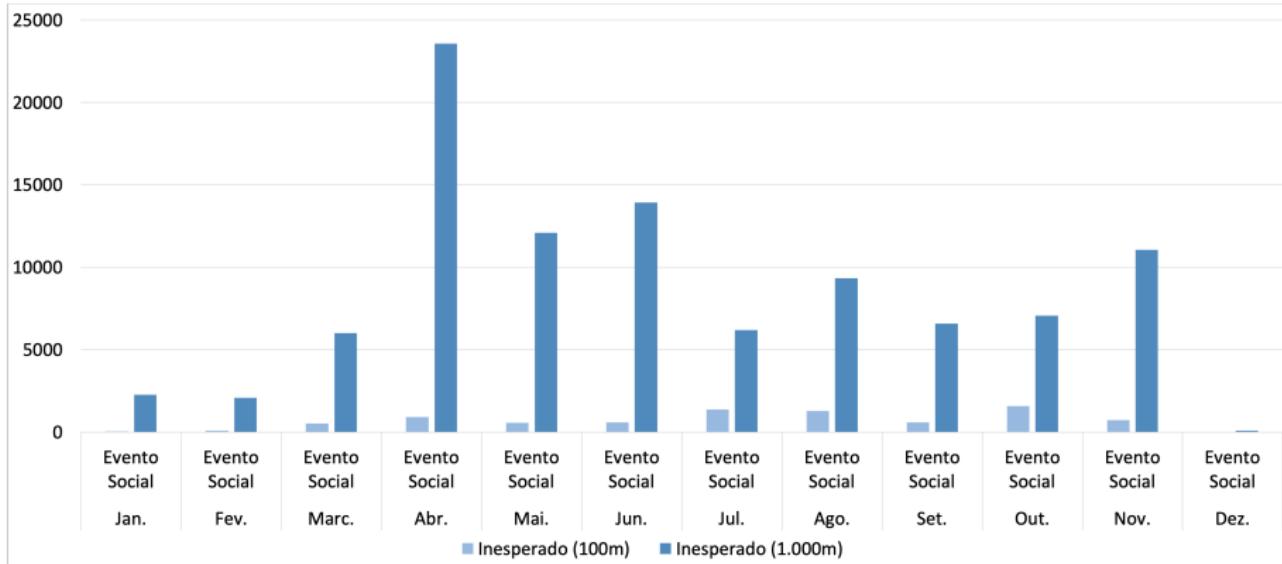
Velocidades inesperadas relacionadas a acidentes



Velocidades inesperadas relacionadas aos desastres naturais



Velocidades inesperadas relacionadas aos eventos sociais



Velocidades inesperadas relacionadas aos eventos urbanos

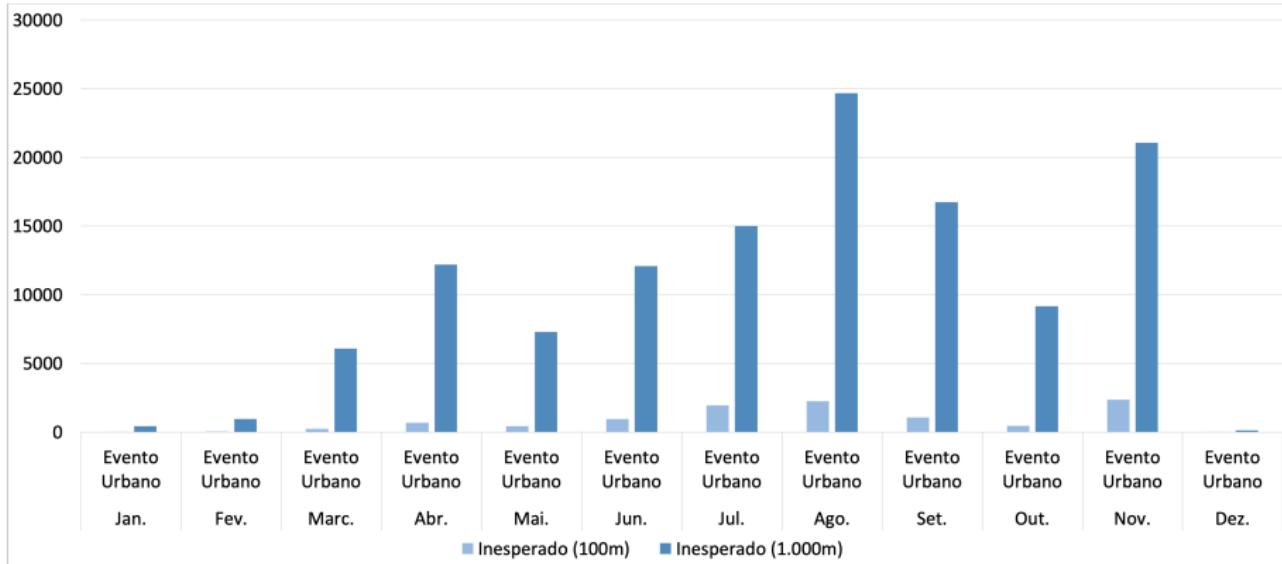


Tabela: Análise *Apriori* aplicada as velocidades médias (intervalos de 5 minutos) ao conjunto de dados AVL da SPTrans correlacionados aos eventos de exceção (a distância de 100 m^f e 1.000 m^g, respectivamente, dos pontos de parada de ônibus) dos meses do ano de 2017

Classe do evento	Total de eventos ^a	Total de Regras de Associação ^b	Esperadas ^c	Não esperadas ^d	Parcialmente inesperadas ^e
Acidente	1.677	315.063	278.493	30.804	5.766
Desastre Natural	912	115.301	99.206	14.282	1.813
Evento Social	506	61.927	52.403	8.245	1.279
Evento Urbano	596	93.513	81.261	10.480	1.772
Total	3.691	585.804	511.363	63.811	10.603
Classe do evento	Total de eventos ^a	Total de Regras de Associação ^b	Esperadas ^c	Não esperadas ^d	Parcialmente inesperadas ^e
Acidente	3.029	3.980.542	3.415.780	385.728	179.034
Desastre Natural	2.016	2.624.415	2.253.123	259.285	112.007
Evento Social	764	1.262.805	1.118.546	100.224	44.035
Evento Urbano	980	1.481.040	1.296.476	125.803	58.761
Total	6.789	9.348.802	8.083.925	871.040	393.837

^a Total de eventos de exceção.

^b Total de correlações de velocidade média.

^c Regras esperadas ($Lift > 1$, $Support > 0,05$)

^d Regras de associação inesperadas ($Lift = 1$).

^e Regras de associação parcialmente inesperadas ($0 < Lift < 1$).

^f 3.545 eventos de exceção não atingiram linhas de ônibus no raio de 100 m.

^g 447 eventos de exceção não atingiram linhas de ônibus no raio de 1.000 m.

Tabela: Análise *Apriori* aplicada as velocidades médias (intervalos de 5 minutos) ao conjunto de dados AVL da SPTrans correlacionados aos eventos de exceção (a distância de 100 m^g e 1.000 m^h, respectivamente, dos pontos de rota dos ônibus) dos meses do ano de 2017

Classe do Evento	Total de Eventos ^b	Qtd. Regras de Associação ^c	Esperadas ^d	Não Esperadas ^e	Parcialmente inesperadas ^f
Acidente	2.367	3.390.690	3.164.726	171.860	54.104
Desastre Natural	1.307	1.342.048	1.247.219	75.981	18.848
Evento Social	704	1.522.423	1.433.700	67.835	20.888
Evento Urbano	825	1.602.343	1.499.305	79.155	23.883
Total	5.203	7.857.504	7.344.950	394.831	117.723

Classe do evento	Total de eventos ^a	Total de Regras de Associação ^b	Esperadas ^c	Não esperadas ^d	Parcialmente inesperadas ^e
Acidente	3.035	2.772.368	2.259.806	365.234	147.328
Desastre Natural	2017	1.876.843	1.545.172	239.897	91.774
Evento Social	764	683.037	588.385	63.549	31.103
Evento Urbano	980	963.892	805.901	111.898	46.093
Total	6.796	6.296.140	5.199.264	780.578	316.298

^a Total de eventos de exceção.

^b Total de correlações de velocidade média.

^c Regras esperadas ($Lift > 1$, $Support > 0,05$)

^d Regras de associação inesperadas ($Lift = 1$).

^e Regras de associação parcialmente inesperadas ($0 < Lift < 1$).

^f 2.033 eventos de exceção não atingiram linhas de ônibus no raio de 100 m.

^g 440 eventos de exceção não atingiram linhas de ônibus no raio de 1.000 m.

Considerações sobre a caracterização dos impactos dos eventos de exceção

- De acordo com os estudos realizados é possível caracterizar padrões inesperados e reduções de velocidades relacionadas aos eventos de exceção.
- Tais padrões foram validados de acordo com os períodos de sazonalidade e dos eventos de exceção identificados nos tweets.
- Notícias veiculadas na mídia correlacionadas aos padrões identificados:
 - Greve geral, manifestações.
 - Ataques aos ônibus de São Paulo.
 - Ataques aos ônibus do Ceará.

Conclusão e contribuições

- Estudo realizado para caracterização de eventos de exceção e de seus respectivos impactos no sistema de transporte público por ônibus da cidade de São Paulo, com dados reais obtidos de fontes públicas e heterogêneas: *tweets*, dados históricos dos módulos AVL e da GTFS.
- Uma nova metodologia para extração e geolocalização dos endereços contidos nas publicações dos órgãos responsáveis por reportar eventos de exceção da cidade de São Paulo é proposta e validada.
- Uma arquitetura distribuída para exploração e visualização de dados AVL.

Conclusão e contribuições

- Uma nova metodologia desenvolvida para extração e geolocalização automática de endereços a partir de mensagens postadas no Twitter, adequada para as contas governamentais responsáveis pelas notificações de eventos de exceção da Cidade de São Paulo.
- Modelos de classificação automatizada de eventos de exceção, treinados com 60.984 tweets classificados manualmente.
- A abordagem que utiliza as coordenadas espaciais dos pontos de parada de ônibus como referência mostrou-se mais adequada do que a que usa os pontos de rota. Devido aos resultados semelhantes obtidos, menor custo computacional e margem de erro.

Trabalhos atuais

Trabalhos publicados

DIAS, F. C. A.; CORDEIRO, D. *Visualizing large datasets: A case study with data of the buses of São Paulo city.* In: *1st Workshop on the Distributed Smart City (WDSC'2018)*, 2018, Salvador, BA. *Proceedings of the 37th IEEE International Symposium on Reliable Distributed Systems*, 2018. p. 10-13.

Trabalhos submetidos

DIAS, F.C.A; CORDEIRO, D. *Characterization of exception events and their respective impacts on the public transport system by bus of São Paulo.* Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC), 2019.

Publicações

- Uso dos *tweets* na caracterização dos impactos dos eventos de exceção.
 - Identificação de padrões inesperados nas velocidades correlacionados aos eventos de exceção, por meio do Algoritmo *Apriori*.
 - Revisão Sistemática.
-
- Implementar o fluxo de processamento de dados em *streaming*, em um cenário de exploração e visualização de dados quase em tempo real.
 - Estabelecer uma cooperação entre a Acadêmia e a SPTrans para aplicação cotidiana dos estudos realizados por esse trabalho e outros relacionados a análise de grandes volumes de dados de transportes públicos.
 - Aplicar os estudos realizados por este trabalho a publicações de usuários que representam a sociedade civil.

Caracterização de eventos de exceção e de seus respectivos impactos no sistema de transporte público por ônibus da cidade de São Paulo

Felipe Cordeiro Alves Dias

Orientador: Prof. Dr. Daniel de Angelis Cordeiro

Universidade de São Paulo

19 de Março de 2019