Processamento de Linguagem Natural (PLN) e Aplicações em Políticas Públicas

Prof. Ricardo Ceneviva

PPG-PP/UFABC

Texto como base de dados:
descoberta, mensuração e inferência

Objetivos da aula

- Compreender as três tarefas: descoberta, mensuração e inferência.
- Reconhecer escolhas-chave: seleção do corpus e representação textual.
- Conhecer modelos básicos: BoW, TF-IDF, tópicos, embeddings e classificadores.
- Relacionar medidas textuais a perguntas de políticas públicas.
- Antecipar riscos de validade e cuidados éticos/LGPD.

Por que texto como base de dados?

- Fontes abundantes e estratégicas:
  - Discursos legislativos, ementas, regulamentos, pareceres, auditorias.
  - Diários oficiais, mídias/redes, registros administrativos (campos livres).
- Valor analítico:
  - Medir agendas, enquadramentos, crédito por políticas e prioridades.
  - Apoiar avaliação e desenho de políticas com evidência textual mensurável.
- **Mensagem: métodos ajudam, mas não substituem teoria e julgamento humano.**

Texto como base de dados: visão geral

- Três tarefas articuladas:
- Descoberta
  - Explorar regularidades/temas sem rótulos prévios.
- Mensuração
  - Construir variáveis confiáveis a partir do texto.
- Inferência
  - Responder a perguntas preditivas ou causais.
- **O método adequado depende da tarefa e da pergunta de pesquisa.**

Seis princípios orientadores

1. Ancorar em teoria e conhecimento substantivo.
2. Métodos ampliam, não substituem, a leitura humana.
3. Iterar e acumular: modelos são aproximações úteis, não fins em si.
4. Modelos extraem generalizações; evite ‘verdades’ rígidas.
5. Escolhas dependem da tarefa (descoberta, mensuração, inferência).
6. Validação é central e específica à tarefa.

Seleção do corpus: o que entra no seu estudo?

- Defina população e quantidade de interesse:
  - Quem fala? Onde? Quando? Sobre o quê?
- Quatro vieses na coleta:
  - Recursos (custos), incentivos (o que é publicado), meio (formato), recuperação (busca/raspagem).
- Documente decisões de inclusão/exclusão e implicações para validade externa.

Representação: panorama de opções

- Bag-of-Words (BoW) e matriz documento-termo (interpretável, simples).
- TF-IDF para salientar termos distintivos por documento.
- Embeddings (palavra/sentença) para semântica e similaridade.
- Sequências e rótulos linguísticos (POS, NER, dependências) para IE/extração.

Bag-of-Words na prática

- Pré-processamento consciente:
  - Normalização (lowercase), remoção de pontuação/números, stopwords.
  - Stemming/lematização: ganhos vs. perda de nuances.
- Construção da DTM/DFM:
  - Filtros por frequência mínima/máxima; n-gramas para expressões frequentes.
- Cuidado: pequenos ajustes mudam resultados; registre suas escolhas.

Modelo multinomial por trás do BoW

- Intuição: documentos geram contagens de termos sob um processo multinomial.
- Regularização (Dirichlet/Laplace) estabiliza estimativas com baixa frequência.
- Base para vários métodos (tópicos, Naive Bayes, Wordscores).

Espaço vetorial e TF-IDF

- Representar documentos como vetores → medir similaridade/distância.
- TF-IDF reduz peso de termos onipresentes e destaca termos informativos.
- Útil para busca, agrupamento, classificação linear e visualizações.

Embeddings de palavras e sentenças

- Ideia: palavras em contextos parecidos têm vetores próximos.
- Modelos distribuídos e contextualizados (ex.: BERT), pré-treinados.
- Agregação ao nível do documento (médias, CLS); valide e audite vieses.

Representações por sequência e IE

- Anotar sequência com POS/NER/dependências aumenta granularidade.
- Casos de uso: atores, órgãos, programas e valores orçamentários em textos.
- Integração com regras ou ML para extração de informação (IE).

Descoberta: princípios gerais

- Nem sempre há ‘ground truth’; foque no conceito e no uso substantivo.
- Separe dados para avaliar descobertas e evitar overfitting interpretativo.
- Prefira abordagens interativas e documentação da rotulagem humana.

Clustering de documentos

- k-means e métodos hierárquicos organizam documentos por similaridade.
- K é decisão substantiva; compare soluções e justifique.
- Interpretação humana é indispensável; explicite regras de rotulagem.

Modelos de tópicos: LDA

- Intuição: documentos são misturas de temas; temas, distribuições de termos.
- Decisões: número de tópicos (K), priors, filtros; rotulagem transparente.
- Avaliação: coerência, estabilidade e utilidade para a pergunta.

Structural Topic Model (STM)

- Inclui covariáveis documentais: prevalência (quem/quando fala sobre quê).
- Modela conteúdo por grupo (ex.: partido) e variação temporal.
- Interprete efeitos estimados; não confunda com causalidade.

Mensuração: princípios práticos

- Defina o construto e a hipótese de uso antes do modelo.
- Garanta fonte pública e processo reprodutível (dados + código).
- Planeje validação e reporte limitações explicitamente.

Dicionários e Wordscores

- Dicionários: simples e transparentes; sensíveis a domínio e ambiguidade.
- Wordscores: posiciona documentos a partir de textos ancorados.
- Valide com amostras rotuladas e medidas externas (construct).

Classificação supervisionada

- Pipeline: rotular (amostra), treinar (NB/linear/árvores), avaliar e aplicar.
- Conjunto de treino deve cobrir variação relevante do corpus.
- Relate métricas (precisão, recall, F1) e análise de erros.

Validação de medidas

- Com ‘gold’: holdout, cross-validation, acurácia e erros sistemáticos.
- Sem ‘gold’: rótulos substitutos, checagens por não-especialistas, convergência externa.
- Teste robustez a pré-processamento e hiperparâmetros.

De descoberta para medida (repurposing)

- Workflow: 1) explorar; 2) definir medida; 3) congelar; 4) validar em dados novos.
- Riscos: instabilidade, opacidade, superinterpretação.
- Mitigue com documentação, sementes fixas e avaliação independente.

Inferência: princípios

- Distinguir previsão e inferência causal; validações diferentes.
- Texto como feature, tratamento, desfecho ou confundidor.
- Causalidade requer desenhos (RCT, DiD, IV, RD) e suposições claras.

Previsão com texto e nowcasting

- Casos: previsão de decisões, detecção de risco, monitoramento quase em tempo real.
- Perigos: concept drift, mudanças institucionais e incentivos de publicação.
- Boas práticas: re-treino, janelas móveis e auditoria fora da amostra.

Causalidade com texto

- Defina antes a função g(texto) → medida; evite ‘pescaria’ pós-hoc.
- Separe treino/teste; considere experimentos sequenciais quando viável.
- Explicite ameaças à validade (omissos, mediadores, seleção).

Aplicações em políticas públicas (exemplos)

- Agenda legislativa e crédito por políticas em discursos e ementas.
- Monitoramento de implementação: relatórios e diários oficiais.
- Análise de transparência: linguagem em auditorias e TCs.

Boas práticas, ética e LGPD

- Minimização e anonimização: colete apenas o necessário; proteja metadados.
- Transparência: descreva fonte, scraping, pré-processamento e versões de modelos.
- Governança: versionamento, reprodutibilidade (scripts, seeds), plano de riscos.