Aplicação de Modelos Probabilísticos de Tópicos para Artigos de Jornal em Português

Alex Siqueira Lacerda - 16/0047692, Marcos Paulo Diniz - 16/0035716

***Abstract*— Este trabalho visa analisar a aplicação de modelos probabilísticos de tópicos sobre um corpus de notícias de jornal em português, em termos da coerência e utilidade dos tópicos encontrados, além de possíveis aplicações para recuperação, classificação e clusterização de artigos de jornal.**

I. Introdução

Modelos Probabilísticos de Tópicos são um grupo de algoritmos de aprendizado de máquina não-supervisionados que buscam inferir, à partir das palavras contidas em uma grande quantidade de documentos textuais, a sua estrutura temática, ou seja, quais temas são abordados por aquele conjunto de documentos e como estes temas se interconectam.

Como explica [4], esses modelos têm grande utilidade na sumarização e exploração de arquivos textuais, especialmente para grandes quantidades de textos, que tornaria a análise manual muito onerosa. Além disso, relações semânticas entre documentos são mais facilmente identificadas quando aplicados estes modelos, abrindo espaço para aplicações em clusterização, taggeamento e sugestão de relacionados.

Neste trabalho, iremos analisar os tópicos gerados pelo modelo probabilístico *Latent Dirichlet Allocation* aplicado a um dataset de notícias de jornal publicadas nos anos de 1994 e 1995.

O primeiro problema a ser resolvido é a classificação mais granularizada dos artigos. Tradicionalmente, artigos de jornal são salvos por editoriais (Esporte, Política, Gastronomia, ...), porém, utilizando-se um modelo probabilístico de tópicos, esta classificação pode ter muito mais dimensões, tornando mais eficiente tal classificação.

Fazendo uso de documentos classificados com maior granularidade, iremos investigar se a recuperação de documentos similares a um novo artigo é mais precisa.

Outro problema é o taggeamento de artigos pequenos. Como retirar informação útil de um texto com 3-4 linhas? Por exemplo, de um texto sobre futebol, mas que não menciona a palavra "futebol", é possível automaticamente obter esta informação? Neste trabalho iremos analisar se, a partir dos tópicos que o modelo atribui a um artigo novo de tamanho pequeno, podemos obter informações semânticas importantes não contidas no texto. Podendo, assim, fazer um taggeamento mais eficiente.

II. Modelo

A. Latent Dirichlet Allocation (LDA)

Como dito anteriormente, modelos probabilísticos de tópicos se utilizam dos documentos e das palavras contidas neles para inferir quais são os tópicos por eles abordados e quais palavras caracterizam estes tópicos.

Latent Dirichlet Allocation (LDA) [2] (modelo mais simples e mais difundido) se utiliza do processo generativo para descrever a forma como documentos poderiam ser gerados a partir de distribuições conhecidas. Tais distribuições são:

Tópicos - Cada um dos tópicos é uma distribuição discreta de *Dirichlet* sobre todas as palavras do vocabulário. Esta distribuição tem um fator λ de concentração, que garante que um pequeno subconjunto de palavras terá maior probabilidade. A quantidade **K** de tópicos é definida pelo usuário.

Documentos Alocados - Distribuição de documentos sobre todos tópicos, que pretende refletir o assunto tratado por cada um deles. Novamente a distribuição de Dirichlet é utilizada para garantir que um pequeno conjunto de tópicos terá maior probabilidade para cada documento.

O **processo generativo** se inicia por selecionar aleatóriamente um tópico dentro da distribuição do documento que se está gerando. Em seguida, seleciona-se uma palavra aleatóriamente dentro do tópico encontrado no passo anterior. Por fim, a palavra é inserida no documento e o processo se repete até que o documento esteja completo.

Este exercício imaginativo é interessante para se entender o funcionamento e utilidade de tais distribuições. Entretanto, na prática, a única variável conhecida são os documentos e suas palavras. As distribuições de documentos e de tópicos são justamente as variáveis que queremos descobrir.

Como explica [1], para se calcular precisamente estas distribuições, todas as possibilidades de estruturas de tópicos deveriam ser analisadas, o que é intratável computacionalmente. Diante deste desafio, algumas técnicas de estatística podem ser utilizadas para aproximar estes valores. Uma destas técnicas é o Amostrador de Gibbs, que basicamente inicia as distribuições com valores aleatórios e os ajusta iterativamente a partir de amostras do dataset.

O LDA é uma ferramenta poderosa, e uma de suas vantagens é poder ser usada como acessório na resolução de problemas mais complicados. Desde que foi proposto, LDA é alterado, extendido e adptado para diversas aplicações.

B. Avaliação de Modelos Probabilísticos de Tópicos

No trabalho *Evaluation Methods for Topic Models* [9], foram propostos métodos intrínsecos de avaliação de Modelos Probabilísticos de Tópicos, baseando-se em held-out likelihood e perplexidade. Porém, Chang em [5] propôs novas formas de avaliação de modelos a partir da análise semântica das palavras mais importantes dos tópicos. Os resultados demonstraram que a análise semântica dos tópicos eram **negativamente correlacionados** com os métodos intrínsecos utilizados anteriormente. Demonstrando que a avaliação

semântica dos tópicos é mais efetiva que a avaliação intrínseca dos modelos que os geram.

A forma de avaliação automática mais comum é a proposta por David Newman em [8], que calcula o Pointwise Mutual Information (PMI) de todas as combinações de pares de palavras do tópico. Este calculo é feito calculando-se a co-ocorrência de cada par de palavras em uma base de conhecimento externa (geralmente a Wikipedia).

$$PMI(w_i, w_j) = \log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}$$

Neste trabalho, iremos utilizar a sua versão normalizada, Normalized Pointwise Mutual Information (NPMI), proposta em [3] e que já demonstrou ótimos resultados em alguns estudos de comparação [6].

$$NPMI(w_i, w_j) = \frac{PMI(w_i, w_j)}{-\log P(w_i, w_j)}$$

III. Material e Métodos

Todos os códigos utilizados neste trabalho estão disponibilizados em `https://github.com/siqueiralex/lda_newsarticles_ptbr`.

Utilizamos a implementação de Latent Dirichlet Allocation presente na biblioteca python gensim. Será utilizada a interface disponível em `https://github.com/siqueiralex/LdaMalletHandler`.

Para avalição dos tópicos, utilizamos a API 'pt-topic-check', cujo código pode ser encontrado em `https://github.com/siqueiralex/pt-topic-check`. Nela, avaliações de tópicos são feitas utilizando as métricas PMI e NPMI na Wikipedia em português.

O dataset consite em notícias de 1994 e 1995 dos jornais "O Público" e "Folha de São Paulo". Disponibilizados em `https://www.linguateca.pt/chave/`. Este corpus não possui informações detalhadas sobre os artigos, neste trabalho utilizamos apenas a data e o texto puro de cada notícia.

A partir do texto puro, realizamos o pré-processamento que consiste em:

- **Cleaning**: Eliminação de de caracteres não alfanuméricos, vírgulas, acentos e pontuação. Parsing de todas as letras para minúsculo.
- **Tokenização**: Transformação do texto em uma lista de palavras.
- Retirada de **Stopwords**, que são palavras de pouco valor semântico ('de', 'a', 'ou', ...)

Para o pré-processamento, utilizamos as bibliotecas disponíveis para python do scikit-learn e nltk.

IV. Solução e Análise

A. Seleção da quantidade de tópicos

Inicialmente, testamos diferentes quantidades de tópicos para o modelo LDA, a fim de encontrar o melhor número. Também fizemos o mesmo processo utilizando outro modelo mais geral de redução de dimensionalidade, Non-negative Matrix Factorization (NMF) [7]. Rodamos ambos modelos para 50, 100, 150 e 200 tópicos. A comparação dos resultados foi feita utilizando a métrica NPMI, considerando a probabilidade de co-ocorrência de cada par de palavras em uma janela de 10 palavras dentro da Wikipedia em português. Para avaliar cada modelo, consideramos a média do NPMI de todos os tópicos gerados.

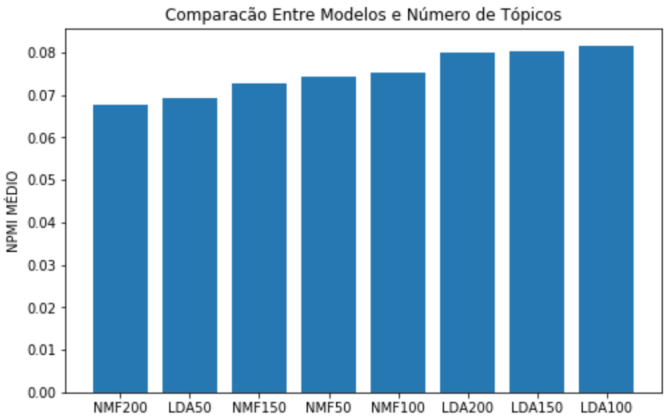

Fig. 1. Comparação entre modelos

Os resultados sugerem que o LDA com 100 tópicos divide o corpus em assunto mais semânticamente coerentes e, portanto, utilizamos este modelo para prosseguir com os nosso trabalho.

B. Comparação entre documentos

Para comparação, consideramos cada documento como um vetor de N-dimensões (N = número de tópicos). De posse da distribuição sobre tópicos dos documentos, utilizamos a distância cosseno pra avaliar a proximidade entre eles.

$$Similaridade(D_i, D_j) = \frac{D_i \cdot D_j}{\|D_i\| \, \|D_j\|}$$

Para artigos novos, primeiramente o classificamos utilizando o modelo pré-treinado, em seguida, a partir da distribuição sobre tópicos do documento, fazemos uma busca por documentos mais próximos dentre os já classificados.

Para recuperação de documentos mais representativos de um tópico, o mesmo calculo é feito, porém considerando a distância para um vetor unitário, com apenas uma componente valor 1 no tópico de que se deseja obter os documentos mais representativos. Ficando a similaridade do documento D_i para o tópico n assim:

$$Representatividade(D_i, n) = \frac{D_{i,n}}{\|D_i\|}$$

Sendo $D_{i,n}$ a probabilidade do documento D_i pertencer ao tópico n.

V. Resultados

Rodamos o modelo LDA para 100 tópicos. Configuramos o modelo para estimar as distribuições através do Amostrador de Gibbs em 1000 iterações. Com os 100 tópicos gerados, separamos as 10 palavras mais significativas de cada um

e, segundo a métrica PMI, selecionamos os 5 mais bem avaliados:

1) carro carros fabrica veiculos modelo veiculo motor ford fiat volkswagen (**Indústria Automotiva**)
2) deputado camara pmdb deputados senado comissao pfl senador parlamentares congresso (**Política**)
3) empresa mercado lojas produtos vendas produto consumidor marketing venda loja (**Comércio**)
4) candidato pt lula partido campanha eleitoral psdb eleicao candidatos governador ((**Eleições/PT**))
5) saude medico hospital medicos doenca casos aids virus tratamento hospitais (**Saúde/AIDS**)

Nota-se que todos são constituídos por palavras fortemente relacionadas. As palavras destacadas entre parêntese são rótulos sugerido por nós a partir das palavras constituintes dos tópicos.

Pra análise mais detalhada neste trabalho, escolhemos dois tópicos:

1) real urv precos inflacao salarios mp plano valor medida conversao (**Plano Real**)
2) senna piloto prova corrida ano equipe carro brasileiro mundial ayrton (**F1/Senna**)

Vale ressaltar que os rótulos 'Plano Real' e 'F1/Senna' não são dados pelo modelo, mas atribuídos pelos autores deste artigo para melhor se referir aos tópicos.

A. Artigos Mais Representativos

Para o tópico **Plano Real**, as duas notícias mais representativas foram:

"1994-01-28 - O governo editará medida provisória para definir os parâmetros de conversão em URVs (Unidade Real de Valor) de preços, salários, tarifas públicas, aluguel, salário mínimo e mensalidade escolar. O secretário de Política Econômica da Fazenda, Winston Fritsch, defende a conversão de preços e salários pela média. No caso das tarifas públicas, elas serão convertidas em URV pelo pico (...)"

e

"1994-07-03 - Os aluguéis residenciais que permaneceram em cruzeiros reais serão convertidos para o real através do cálculo da média em URV. A fórmula é semelhante à dos salários, mas com um detalhe: os aluguéis ficam meses e meses com o mesmo valor nominal. Os salários vinham tendo reajustes mensais embora expurgados (...)"

Tais documentos indicam que este é um tópico que agrupa notícias que abordam questões quanto à moeda de transição do Plano Real, o URV.

Para o tópico **F1/Senna**, as duas notícias mais representativas foram:

"1995-09-24 - A Williams não deu chances aos rivais e confirmou a primeira fila do grid no GP de Portugal, que acontece hoje, a partir das 10h (de Brasília), no circuito do Estoril, com transmissão pela TV. Ontem, na segunda sessão de treinos oficiais, David Coulthard foi o mais rápido (...)"

e

"1994-05-15 - Em homenagem a Senna e Ratzenberger, a 1ª fila não será ocupada. 1. 2. 3. Michael Schumacher (ALE)/Benetton - 1min18s560 4. Mika Hakkinen (FIN)/McLaren - 1min19s488 (...)"

Como a notícia mais representativa não fala sobre o piloto brasileiro, pode-se concluir que o tópico inclui não só notícias sobre Ayton Senna mas todas as notícias que tratam de corridas de Fórmula 1.

B. Análise Temporal

Para cada tópico, extraímos as 500 notícias mais representativas e as agrupamos pelos meses dos anos de 1994 e 1995, para analisar como elas se mostram temporalmente.

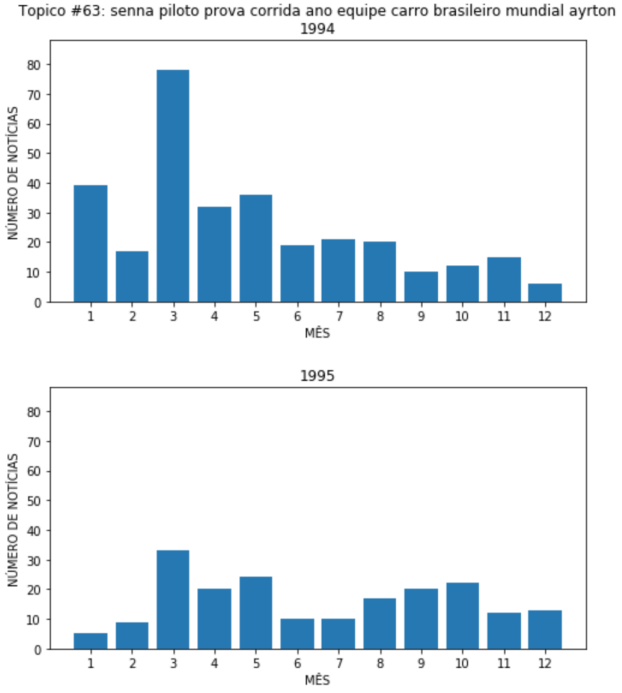

Fig. 2. Artigos do tópico **F1/Senna** no tempo

Para ambos os tópicos o espalhamento temporal se mostrou dentro do esperado. O tópico **F1/Senna** apresenta um pico no mês de fevereiro, que é esperado pois este é o mês em que o GP do Brasil ocorria. Um arrefecimento na quantidade de notícias pode ser notada, devido ao falecimento do Piloto Brasileiro em maio de 1994. Já o tópico **Plano Real** também possui um pico no primeiro semestre de 1994, especialmente nos meses de Fevereiro e Março que foram os meses em que o plano econômico estava sendo implementado.

Este tipo de análise demonstra o poder de exploração que o modelo propicia, dando ao usuário uma ferramenta para analisar a distribuição de centenas de documentos relacionados, sem a necessidade de lê-los.

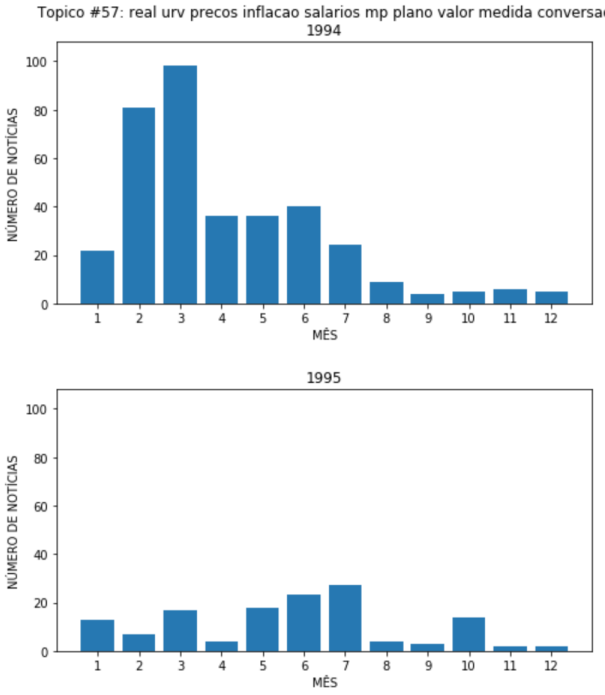

Fig. 3. Artigos do tópico **Plano Real** no tempo

C. Aplicação para Novo Documento

Para simular a escrita de um novo artigo de jornal, selecionamos o primeiro parágrafo de uma notícia presente em um grande periódico on-line. Então, pré-processamos e aplicamos esta nova notícia ao nosso modelo treinado para classificação e recuperação de documentos relacionados.

O parágrafo selecionado da notícia "Brasil registra queda de 16% no número de detecções de Aids" do G1 foi:

"O Brasil registrou uma redução de 16% no número de detecções de Aids em 2018, segundo o Boletim Epidemiologico divulgado nesta terça-feira."

1) Classificação: O tópico que o modelo indica como tendo maior probabilidade para a notícia é:

saude medico hospital medicos doenca casos aids virus tratamento hospitais

Note que, além de classificar corretamente a nova notícia, o tópico sugerido adiciona valor semântico ao parágrafo. Por exemplo, a palavra 'saúde' não está presente na notícia, mas é trazida como a mais importante do tópico, e é obviamente relacionada ao assunto. Isto torna mais eficiente atividades como o taggeamento e anotação semântica do novo artigo.

2) Notícias Mais Similares:: As notícias indicadas pelo modelo como mais similares ao novo artigo foram:

"1995-10-14 - Acontece no próximo dia 19, às 15h, no Colégio Rio Branco (zona oeste de SP), o "Fórum de Debates sobre Prevenção de Drogas e Aids em Escolas". O debate será coordenado pelo professor Claude Olievenstein, diretor do Centre Medical Marmottan de Paris, especializado na recuperação de farmacodependentes."

e

"1995-01-05 - O Comitê da Cidadania dos Funcionários do Banco do Brasil inaugura hoje um posto de saúde na favela do Pirambu, considerada a maior de Fortaleza. Foram investidos R$ 12 mil no posto, que contará com serviço de atendimento médico e odontológico e uma farmácia. O posto ocupará uma área de 100 m2."

Nota-se que existe similaridade entre os documentos. Levando em conta o tamanho da notícia classificada (apenas um pequeno parágrafo), pode-se inferir que o modelo classifica e recupera documentos eficientemente.

VI. Conclusão

Após a análise dos resultados obtidos, percebe-se o sucesso do algoritmo de aprendizado não supervisionado na clusterização e classificação das notícias.

Primeiramente, o algoritmo do LDA obteve resultados como o esperado, conseguido fazer a distribuição dos tópicos com uma análise horizontal entre as notícias, isto é, analizar todas as notícias e verificar a possível relação entre notícias semelhantes e, com isso, gerar os tópicos dos assuntos abordados no corpus. Com essa funcionalidade, conseguiu-se para um determinado tópico retornar as notícias mais relevantes, obedecendo a escala de probabilidade, e conforme visto no trabalho realizado os resultados obtidos foram bem satisfatórios.

Outro ponto relevante, foi conseguir aplicar o modelo para uma nova notícia. Nesse caso, a partir de um artigo novo, o modelo foi capaz de identificar qual tópico teria a maior probabilidade de ter o conjunto de notícias simalares a notícia nova adicionada. Sendo essa funcionalidade importante para a recuperação de informações e classificação de novos artigos.

References

[1] David M. Blei. Introduction to Probabilistic Topic Models. *Communications of the ACM*, 55, 2011.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[3] Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. 2009.

[4] Jordan Boyd-Graber, Yuening Hu, and David Mimno. *Applications of Topic Models*, volume 11 of *Foundations and Trends in Information Retrieval*. NOW Publishers, 2017.

[5] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-graber, and David M. Blei. Reading Tea Leaves: How Humans Interpret Topic Models. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 288–296. Curran Associates, Inc., 2009.

[6] Jey Han Lau and David Newman. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. April 2014.

[7] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, NIPS'00, pages 535–541, Cambridge, MA, USA, 2000. MIT Press.

[8] David Newman, Sarvnaz Karimi, and Lawrence Cavedon. External Evaluation of Topic Models. 2009.

[9] Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1105–1112, New York, NY, USA, 2009. ACM.