Workshop de Informação, Dados e Tecnologia (WIDaT)

PROCESSAMENTO DE LINGUAGEM NATURAL E *MACHINE LEARNING* COMO APARATO PARA A CATEGORIZAÇÃO DE ARTIGOS CIENTÍFICOS

Ananda Fernanda de Jesus (UNESP)
Maria Lígia Triques (UEL)
José Eduardo Santarem Segundo (UNESP/USP)
Ana Cristina de Albuquerque (UEL)

2022

Contextualização

CATEGORIZAÇÃO NA CIÊNCIA DA INFORMAÇÃO

- A categorização perpassa o desenvolvimento e estabelecimento da Ciência da Informação
- Destaca-se a busca pela representação da informação e do conhecimento, visando a sua recuperação
- Em atividades como Catalogação, Classificação e Indexação

CATEGORIZAÇÃO NA ELABORAÇÃO DE PESQUISAS

- Necessidade de agrupamento dos resultados para possibilitar sua análise
- **Categorias a *priori*:** os resultados são agrupados em categorias criadas antes de sua análise
- **Categorias a posteriori:** as categorias são elaboradas com base em padrões identificados nos resultados obtidos.

PROCESSAMENTO DE LINGUAGEM NATURAL (PLN) E *MACHINE LEARNING* (ML)

- Potencialmente podem contribuir no processo categorização de artigos científicos
- Elaboração de novas categorias, com aplicação de algoritmos de aprendizado não-supervisionado
- Categorização em categorias pré-estabelecidas, com aplicação de algoritmos de aprendizado supervisionado

Proposta da pesquisa:

- Verificar o potencial de aplicação de técnicas de Processamento de Linguagem Natural (PLN) e de *Machine Learning* (ML) na automação do processo de categorização temática de artigos científicos.
- Partindo de um recorte experimental, buscou-se replicar as duas principais formas manuais de categorização: *priori* e a *posteriori*
- Categorização por meio de um **algoritmo supervisionado**, que executa suas funções utilizando categorias pré-estabelecidas manualmente pelos pesquisadores;
- Categorização por meio de um **algoritmo não supervisionado**, observando o potencial de sua aplicação na criação de novas categorias de análise.

Procedimentos Metodológicos: recorte experimental

- ***Corpus***: 46 artigos, obtidos por meio de levantamento bibliográfico, qualitativo e exploratório da literatura científica;
- ***Base consultada****:* Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação (BRAPCI);
- ***Estratégia de busca****: t*ermo “patrimônio cultural”;
- **Idioma**s: produção nacional em português;
- **Delimitação temporal**: 2012 a maio 2022;
- **Seleção**: identificação da presença do termo “patrimônio cultural” como descritor nas palavras-chaves das publicações e, posteriormente, identificado se havia a definição do termo em seu contexto de estudo;
- **Categorização manual**: identificação de duas categorias a *posteriori,* dentificar duas categorias: contexto de estudo relacionado ao meio digital (categoria A); contexto de estudo não relacionado ao meio digital (categoria B).

Procedimentos Metodológicos: recorte experimental

- Realizou-se a **tokenização** do conteúdo dos textos (divisão do conteúdo do texto em unidades menores, omitindo a pontuação);
- Aplicadas opções de transformação dos dados de modo a garantir a padronização, como remoção de URLs e demais *links e* uniformização em letra minúscula;
- Aplicação de filtros, que permitem remover ou manter uma seleção de palavras, como a definição por idioma, no caso, português, dada o *corpus* de análise;
- Processo de identificação e eliminação de *stopwords.*

Nuvem de palavras antes e depois do pré-processamento

Fonte: Elaborado pelos autores

Procedimentos Metodológicos: aplicação do algoritmos - categorias *a priori*

- Para a avaliação da aplicabilidade em categorização utilizando categorias construídas, que foram rotulados manualmente *a priori*, realizou-se a etapa de treino e teste de um conjunto de algoritmos supervisionados.
- ***Algoritmos testados:*** Rede Neural, KNN e *Random Forest*.
  - O procedimento de treino/teste teve o *corpus* (37 artigos) com uma aplicação de *K-fold cross validation para 20 repetições, com divisão de 70% para treino e 30% para teste*.
  - Utilizou-se a métrica de acurácia para avaliar o resultado dos algoritmos.

Procedimentos Metodológicos: aplicação do algoritmos - categorias *a priori*

- Apesar do pequeno *corpus*, o que normalmente não favorece um algoritmo de Rede Neural, ele teve melhor desempenho com uma acurácia de 88%, já o KNN atingiu 84% e o *Random Forest* com 82%.
- Para a validação, portanto, foi aplicado o algoritmo Rede Neural, levando em consideração 20% dos artigos (9 dos 46 artigos).
- Tais documentos já haviam sido previamente rotulados pelos pesquisadores de forma manual para permitir a checagem dos erros e acertos, mas essa rotulação não foi indicada ao algoritmo.

RESULTADOS

Algoritmos Supervisionados

Categorias *a priori*

- Na etapa de validação apenas 1 dos 9 artigos foi classificados incorretamente, obtendo-se uma acurácia aproximada de 89%;
- Compreendeu-se que acurácia do algoritmo é influenciada por diversos fatores como o detalhamento e rigor no pré-processamento e na limpeza dos dados, o tamanho e a representatividade da amostra escolhida.
- Quanto mais claras forem as características específicas de cada classe estabelecida a priori e quanto mais representativa for a mostra treino selecionada, maiores serão as chances de acerto do algoritmo.

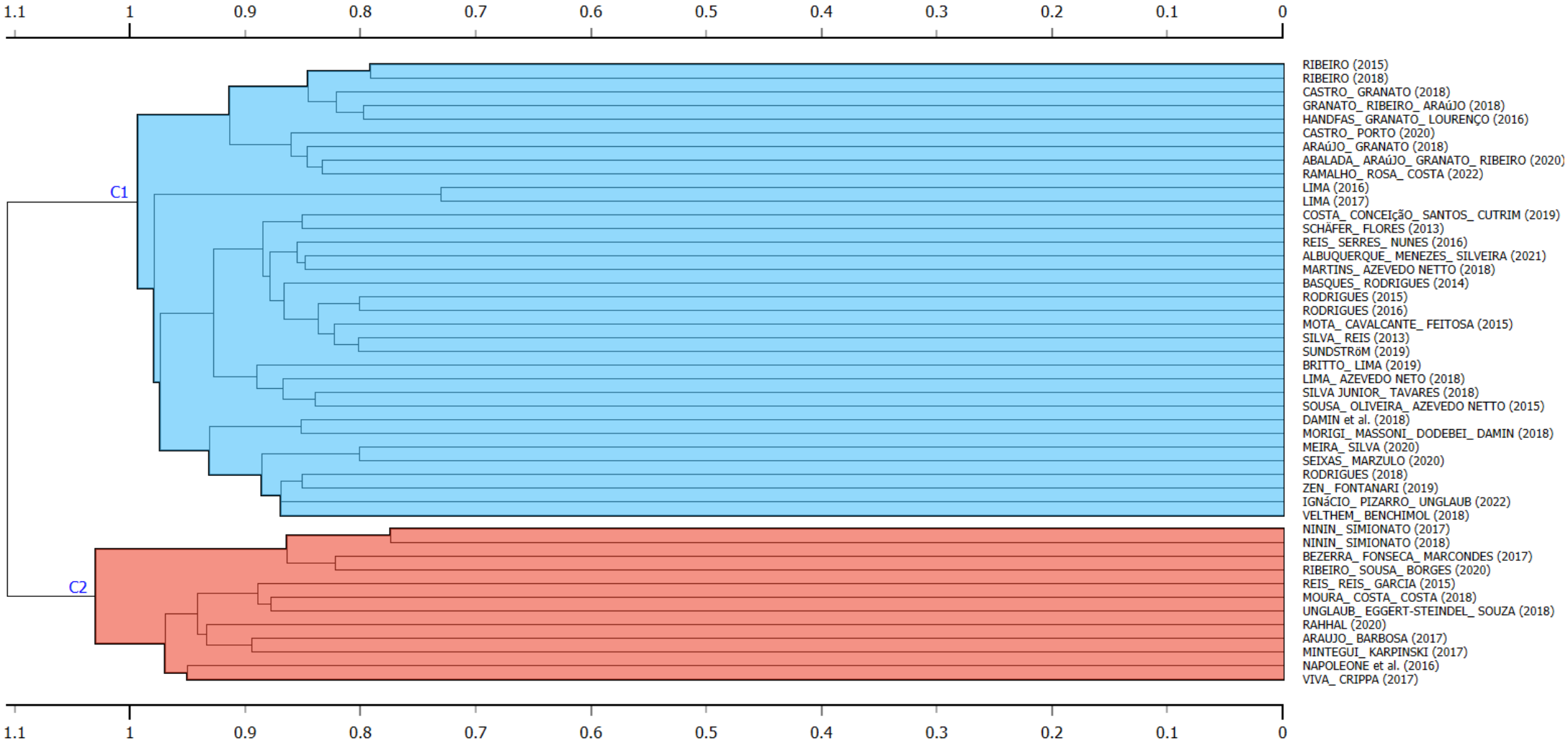
Procedimentos Metodológicos: aplicação do algoritmos - categorias *a posteriori*

- Para a avaliação da aplicabilidade em categorias construídas *a posteriori* foi utilizado o processo não supervisionado, isto é, utilizando padrões estatísticos por meio de um algoritmo de clusterização
- Algoritmo de clusterização hierárquica (*hierarchical clustering algorithm*), tendo como métrica de distanciamento escolhida o índice Jaccard;
- Foi retirada a *feature alvo* que indicava categorização (*target*) dos textos selecionados;
- Foi escolhido os parâmetros de frequência dos termos e dos documentos de forma que fosse calculada a importância de uma palavra em um documento em relação a uma coleção de documentos, e não somente as palavras de maior ocorrência total.

RESULTADOS

Algoritmos não-supervisionados

Categorias a posteriori



Fonte: Elaborado pelos autores

RESULTADOS

Algoritmos não-supervisionados

Categorias a posteriori

- Ao marcar as duas categorias de maior nível na clusterização (C1 e C2), é possível gerar uma nuvem de palavras para cada uma
- A primeira (C1) se aproxima da categoria B identificada manualmente, isto é, referente a um contexto não vinculado ao digital.
- A segunda (C2), se aproxima da categoria A identificada manualmente, o que corresponde ao contexto digital.
- Palavras como ‘museus’, ‘objetos’ e ‘sociais’, são as mais relevantes de C1, enquanto que ‘interoperabilidade’, ‘metadados’ e ‘web’ são os principais destaques da C2.

RESULTADOS

Algoritmos não-supervisionados

Categorias a posteriori

C 1

C 2

Considerações Finais

- Em relação a predição de novos documentos com base em categorias obtidas *a priori*, conclui-se que não é possível excluir a participação do pesquisador no processo de categorização
- O número de acertos faz com que a aplicação do processo seja relevante e possibilite imaginar um novo cenário: algoritmo atuando como uma pré-classificação e o pesquisador como validador, (redução significativa de trabalho manual).
- Outro papel importante do pesquisador no processo será o de seleção da amostra utilizada para treino do algoritmo e estabelecimento correto das características que diferenciam os conjuntos de dados e que permitem a criação das categorias.
- Conclui-se ainda que o potencial de contribuição desse procedimento seria ampliado em análise de grandes volumes de documentos.

Considerações Finais

- Já em relação a criação de categorias a posteriori, aplicando técnicas de PLN e ML, conclui-se que os resultados são mais promissores, tendo em vista que com base na aplicação desse procedimento é possível identificar novos padrões que poderiam passar despercebidos pelos próprios pesquisadores.
- O potencial do procedimento seria ampliado em um contexto de grandes volumes de documentos cuja análise aprofundada pelo pesquisador seria um processo longo e exaustivo, e em muitas situações inviável.

Pesquisas futuras

- Tendo em vista a ampliação do potencial dessas técnicas na análise de grandes conjuntos de documentos, como estudos futuros, pretende-se ampliar a amostra utilizada
- Pretende-se ainda realizar teste com outros algoritmos e esgotar (saturar) o uso de seus parâmetros com busca de melhores resultados

Agradecimentos

Agradecemos à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP - Processo n° 2021/03349-0) e Coordenação de Aperfeiçoamento de Pessoal de Nível Superior pelo financiamento recebido para o desenvolvimento dessa pesquisa.

af.jesus@unesp.br

mligia.triques@uel.br

santarem@usp.br

albuanati@uel.br

OBRIGADO!

2022