UNIVERSIDADE FEDERAL DA FRONTEIRA SUL
CAMPUS CHAPECÓ
CURSO DE CIÊNCIA DA COMPUTAÇÃO

IVAIR PUERARI

ANÁLISE EXPLORATÓRIA SOBRE REGISTROS ELETRÔNICOS DE SAÚDE DO SETOR DE UNIDADE DE TERAPIA INTENSIVA UTILIZANDO MODELAGEM DE TÓPICOS

CHAPECÓ
2019

IVAIR PUERARI

ANÁLISE EXPLORATÓRIA SOBRE REGISTROS ELETRÔNICOS DE SAÚDE DO SETOR DE UNIDADE DE TERAPIA INTENSIVA UTILIZANDO MODELAGEM DE TÓPICOS

Trabalho de conclusão de curso de graduação apresentado como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação da Universidade Federal da Fronteira Sul.
Orientador: Prof. Dr. Denio Duarte

CHAPECÓ
2019

Bibliotecas da Universidade Federal da Fronteira Sul - UFFS

Puerari, Ivair
Análise exploratória sobre registros eletrônicos de saúde do setor de unidade de terapia intensiva utilizando Modelagem de Tópicos / Ivair Puerari. -- 2019.
58 f.:il.

Orientador: Dr. Denio Duarte.
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal da Fronteira Sul, Curso de Ciência da Computação, Chapecó, SC , 2019.

1. Modelagem de Tópicos. 2. Tópicos. 3. Registros Eletrônicos de Saúde. 4. Saúde. 5. LDA. I. Duarte, Denio, orient. II. Universidade Federal da Fronteira Sul. III. Título.

Elaborada pelo sistema de Geração Automática de Ficha de Identificação da Obra pela UFFS com os dados fornecidos pelo(a) autor(a).

IVAIR PUERARI

ANÁLISE EXPLORATÓRIA SOBRE REGISTROS ELETRÔNICOS DE SAÚDE NO SETOR DE UNIDADE DE TERAPIA INTENSIVA UTILIZANDO MODELAGEM DE TÓPICOS

Trabalho de conclusão de curso de graduação apresentado como requisito para obtenção do grau de Bacharel em Ciência da Computação da Universidade Federal da Fronteira Sul.

Orientador: Prof. Dr. Denio Duarte

Aprovado em: 05\12\2019

BANCA EXAMINADORA:

Dr. Denio Duarte - UFFS

Dr. Guilherme Dal Bianco - UFFS

Ma. Andressa Sebben - UFFS

RESUMO

O rápido crescimento dos registros eletrônicos de saúde traz o aumento de informações disponíveis sobre pacientes em hospitais. Essa massiva quantidade de informações em texto é adequada para a extração de informações desconhecidas sobre histórico médico, medicamentos, doenças, alergias, entre outras. A modelagem de tópicos é um problema de aprendizado de máquina que visa extrair, dada uma coleção de documentos, os principais tópicos que representam os assuntos abordados pela coleção. Em modelagem de tópicos, um documento pode ser definido como uma mistura de tópicos, sendo estes gerados a partir de diferentes distribuições probabilísticas de palavras, permitindo assim extrair assuntos em forma de tópicos de coleções de documentos. O objetivo deste trabalho foi realizar uma análise exploratória sobre duas coleções de registros eletrônicos de saúde separados por internações que obtiveram alta e internações que evoluíram a óbito no setor de Unidade de Terapia Intensiva utilizando modelagem de tópicos a fim de identificar os assuntos presentes nas coleções. Após a execução do modelo *Latent Dirichlet Allocation* (*LDA*) foram extraídos 11 tópicos para cada coleção de documentos óbito e alta. Como resultado, para a coleção de altas os assuntos com maior predominância são sistema respiratório, sistema renal, sistema neurológico, prematuridade e sistema cardíaco. Por outro lado, a coleção de óbitos apresenta os principais assuntos como sistema hepático, sistema cardiovascular, sistema neurológico e sistema respiratório. Foram analisadas as disjunções e intersecções dos assuntos definidos em cada coleção, e observado a infecção como importante fator contribuinte para evolução a óbito.

Palavras-chave: Modelagem de Tópicos; Tópicos; Registros Eletrônicos de Saúde; Saúde, LDA, Métricas.

ABSTRACT

The rapid growth of electronic health record systems brings the increase of available information about patients in hospitals. This massive amount of text information is suitable for the extraction of unknown information about medical history, medication, deseases, allergies, among others. Topic modeling is a machine learning problem, which aims to extract, given a collection of documents, the main topics that represent the subjects covered by a text collection. In the topic model, the documents can be composed of a mixture of topics with a certain probability. This work aims to make an exploratory analysis of two collections of electronic records health from an intensive care unit. The collection is split into two subcollections: discharged patients and patients who progressed to death. We apply the Latent Dirichlet Allocation (LDA) algorithm in both collections, setting the number of topics to 11. As a result, discharged patients collection shows the following predominant topics: respiratory system, renal system, neurological system, prematurity, and cardiac system. On the other hand, the death collection presents as main topics subjects about the hepatic system, cardiovascular, neurological, and respiratory system. We also analyze the correlation of the topics inter collections, and we observed that the infection as a significant contributing factor to progress to death.

Keywords: Topic Modeling; Topics; Electronic Health Record, Health, LDA, Metrics.

LISTA DE FIGURAS

Figura 2.1 – Exemplo de Modelagem de Tópicos (BLEI, 2012). 16
Figura 2.2 – Exemplo de Modelagem de Tópicos. Gráficos de tópicos (BLEI, 2012). 17
Figura 3.1 – Visão geral da base de dados de cuidados intensivos MIMIC III (JOHNSON et al., 2016). 27
Figura 3.2 – Diagrama de Classes relacionamento entidade (MILES, 2017). 28
Figura 5.1 – Consultas SQL para busca na base de dados..... 35
Figura 5.2 – Função que gera um documento e coleção de documentos. 36
Figura 5.3 – Fragmento de um documento pertencente à coleção de documentos. 37
Figura 5.4 – StopWords utilizadas no pré-processamento..... 38
Figura 5.5 – Fragmento de um documento pertencente à coleção de documentos pós-processado. 39
Figura 5.6 – Representação matriz documento-termo. 40
Figura 5.7 – Gráfico de coerência de cada tópico inicial. 41
Figura 5.8 – Gráfico de coerência de cada tópico. 41
Figura 6.1 – Gráfico de coerência de cada α. 43
Figura 6.2 – Gráfico de documentos por tópicos da coleção de óbitos. 44
Figura 6.3 – Gráfico de documentos por tópicos da coleção de altas. 45
Figura 6.4 – *Top-5* tópicos da coleção de óbitos. 46
Figura 6.5 – *Top-5* tópicos da coleção de altas. 46
Figura 6.6 – Tabela de respostas da coleção de alta. 47
Figura 6.7 – Tabela de respostas da coleção de óbito. 48
Figura 6.8 – Intersecções e disjunções das coleções de documentos..... 50

LISTA DE TABELAS

Tabela A.1 – Tópicos 1-4 extraídos da coleção de óbitos. 57
Tabela A.2 – Tópicos 5-8 extraídos da coleção de óbitos. 57
Tabela A.3 – Tópicos 9-11 extraídos da coleção de óbitos. 57
Tabela A.4 – Tópicos 1-4 extraídos da coleção de altas. 58
Tabela A.5 – Tópicos 5-8 extraídos da coleção de altas. 58
Tabela A.6 – Tópicos 9-11 extraídos da coleção de altas. 58

LISTA DE APÊNDICES

APÊNDICE A – Tópicos extraídos 56

LISTA DE ABREVIATURAS E SIGLAS

CID	*Classificação internacional de Doenças*
CID-9	*Classificação internacional de Doenças Versão 9*
LDA	*Latent Dirichlet allocation*
MIMIC III	*Medical Information Martfor Intensive Care III*
MIT	*Instituto Tecnológico de Massachussets*
NPMI	*Normalized Pointwise Mutual Information*
PMI	*Pointwise Mutual Information*
REP	*Rochester Epidemiology Projects*
SVTM	*Survival Topic Model*
UTI	*Unidade de Terapias Intensivas*

SUMÁRIO

1 INTRODUÇÃO 12
1.1 Estrutura da Monografia 14
2 MODELAGEM DE TÓPICOS 15
2.1 Tópicos 15
2.2 Documentos 17
2.3 Algoritmos 18
2.3.1 *Latent Dirichlet Allocation (LDA)* 18
2.4 Métricas 19
2.4.1 *Pointwise Mutual Information* (PMI) 20
2.4.2 *Normalized Pointwise Mutual Information* (NPMI) 20
2.4.3 *UCI Coherence* 20
2.4.4 *NPMI Coherence* 20
2.4.5 C_v *Coherence* 20
3 REGISTROS ELETRÔNICOS DE SAÚDE 22
3.1 Prontuário do paciente 22
3.1.1 Prontuário eletrônico do paciente 23
3.2 Registro eletrônico de saúde x Prontuário eletrônico do paciente 24
3.3 Atendimento em Unidades de Terapia Intensiva (UTI) 24
3.4 Base de Dados 25
4 TRABALHOS RELACIONADOS 30
4.1 ***Discovering Associations Among Diagnosis Groups Using Topic Modeling*** 30
4.2 ***Redundancy-Aware Topic Modeling for Patient Record Notes*** 31
4.3 ***Survival Topic Models for Predicting Outcomes for Trauma Patients*** 31
4.4 ***Using topic modeling to infer the emotional state of people living with Parkinson's disease.*** 32
5 PROJETO DE EXPERIMENTO 34
5.1 Configuração do ambiente 34
5.2 Geração das coleções de documentos (Óbito e Alta) 34
5.3 Pré-processamento das coleções 37
5.4 Extração dos tópicos 39
6 EXPERIMENTO 42
6.1 Execução 42
6.2 Extração de tópicos 43
6.3 Resultados 48
6.3.1 Tópicos (Assuntos) da coleção de alta 48
6.3.2 Tópicos (Assuntos) da coleção de óbitos 49
6.3.3 Intersecções e disjunções das coleções de documentos 50
7 CONCLUSÃO 52
7.1 Trabalhos Futuros 52
REFERÊNCIAS 53
APÊNDICES 55

1 INTRODUÇÃO

O volume de dados na área da saúde é de grande escala, com informações de pacientes sendo coletadas rotineiramente (ROSE, 2018). Os dados gerados por meio de avaliações contínuas formam os registros eletrônicos de saúde. Estes, por sua vez, armazenam as informações processáveis sobre o paciente (PANITZ, 2014).

O prontuário eletrônico do paciente, por exemplo, é um elemento constituinte dos registros eletrônicos de saúde, definido como o documento único. O prontuário é formado por um conjunto de informações provenientes de outros documentos, tais como, evolução de enfermagem, anotações de enfermagem e resumo de alta-óbito-transferência, dos quais, grande parte desses dados se apresentam em forma de texto livre (PANITZ, 2014).

Dentre as áreas de saúde, a unidade de terapia intensiva (UTI) é um ambiente hospitalar destinado a assistir pacientes graves e instáveis, considerado de alta complexidade (BACKES; ERDMANN; BüSCHER, 2015). De acordo com BACKES; ERDMANN; BüSCHER (2015) os procedimentos realizados são agressivos e invasivos para reverter o quadro em que o paciente se encontra, mas muitas vezes o óbito do paciente é inevitável. A assistência aos pacientes na UTI está relacionada ao cuidado direto, intensivo e ao monitoramento permanente por meio controle rigoroso dos parâmetros vitais, gerando um grande volume de dados (BACKES; ERDMANN; BüSCHER, 2015).

Dados gerados na área de saúde são de grande interesse em muitas aplicações no desenvolvimento de métodos para extração automática de informações úteis, porém escondidas ao olho humano (ROSE, 2018). Para este fim, a modelagem de tópicos apresenta um conjunto de algoritmos estatísticos que analisam palavras, visando extrair de uma coleção de documentos (textos), os principais tópicos que representam os assuntos abordados pela coleção (BLEI, 2012).

Modelagem de tópicos considera que tópicos são formados por distribuições probabilísticas de palavras, sendo os documentos baseados na ideia que são formados por uma mistura de tópicos. Assim, documentos com diferentes assuntos podem ser gerados a partir de diferentes distribuições sobre tópicos (BLEI, 2012).

Desse modo, a modelagem de tópicos apresenta uma tarefa que automatiza a extração de informações em grandes coleções de documentos. Assim, por meio de técnicas estatísticas é possível inferir o conjunto de tópicos que geraram uma coleção de documentos, e consequente-

mente descrever os assuntos abordados (BLEI, 2012).

A Modelagem de tópicos tem sido aplicada em registros eletrônicos de saúde com o objetivo de encontrar conceitos clínicos relevantes e relações entre pacientes (WANG et al., 2011).

ZHANG; JIANG; PETZOLD (2017) citam que ferramentas que podem fornecer avaliações rápidas e precisas sobre, por exemplo, a condição de um paciente, serão úteis nas tomadas de decisões críticas feitas por profissionais de saúde.

Em vista dos fatores apresentados, a modelagem de tópicos se mostra pertinente como método para extração de informações na área da saúde. A UTI demonstra ser um ambiente crítico de alto risco, com dados sendo gerados sistematicamente. Logo, identificar assuntos abordados em coleções de documentos da UTI a partir de modelagem de tópicos podem contribuir na informação sobre pacientes e ambiente.

Considerando a relevância e os possíveis benefícios proporcionados, este estudo visa desenvolver uma análise exploratória sobre duas coleções de registros eletrônicos de saúde separados por internações que obtiveram alta e internações que evoluíram a óbito, utilizando modelagem de tópicos, a fim de definir os assuntos abordados em cada coleção e identificar intersecções e disjunções nos tópicos pertencentes às duas coleções (alta e óbito). Os registros eletrônicos de saúde são provenientes da base de dados *Medical Information Mart for Intensive Care III* (MIMIC III), disponível gratuitamente, que inclui dados relacionados à saúde de 53.423 internações hospitalares distintas de pacientes que permaneceram em UTIs do Centro Médico Beth Israel Deaconess em Boston, Massachussets entre 2001 e 2012.

Para construção desta proposta foi realizada a definição das coleções de documentos em alta e óbito. Logo, para geração de tópicos, houve a definição de hiper-parâmetros para configuração do modelo e aplicação de métricas para validação da qualidade e coerência dos tópicos criados. Com os tópicos gerados, foi identificado qual assunto é representado dentro de cada tópico e quais são os tópicos mais presentes em cada coleção. Assim foram analisadas intersecções e disjunções entre os tópicos gerados pelas coleções e apresentados os resultados obtidos.

Dentre os resultados obtidos, através da análise dos tópicos, foram identificados assuntos como sistema respiratório, diabete e sistema cardíaco contidos na coleção de alta, enquanto assuntos como sistema neurológico, infecção e sistema hepático estavam presentes na coleção de óbitos.

1.1 Estrutura da Monografia

O trabalho está estruturado na seguinte forma: o Capítulo 2 discorre sobre modelagem de tópicos, apresentando conceitos sobre documentos, tópicos, algoritmos e métricas. No Capítulo 3 são apresentados os conceitos de registros de eletrônicos de saúde e prontuário eletrônico do paciente, como também é dissertado sobre o atendimento em UTI e a base de dados utilizada neste trabalho. No Capítulo 4 são apresentados os trabalhos relacionados ao tema. O capítulo 5 apresenta o projeto de experimento que define as etapas e procedimentos que foram utilizados para atingir o objetivo. Na sequência o Capítulo 6 expõe o experimento realizado e o capítulo 7 apresenta as conclusões.

2 MODELAGEM DE TÓPICOS

O conhecimento coletivo acontece por meio de notícias, blogs, sites, artigos, livros, entre outros. Estes formatos apresentam dificuldades na extração de informações, de modo que, enquanto o número de textos disponíveis cresce, a capacidade humana de extrair informações relevantes sobre todos os dados se mantém (BLEI, 2012). Para este fim, pesquisadores de Aprendizado de Máquina desenvolveram a modelagem probabilística de tópicos.

Modelagem de tópicos dispõe de um conjunto de algoritmos, que visa extrair, dada uma coleção de documentos, os principais tópicos que representam os assuntos abordados pela coleção (BLEI, 2012). Modelagem de tópicos se encontra na classe de algoritmos não supervisionados, onde dados de entrada não possuem rótulos, ou seja, não possuem um resultado para cada exemplo (STEYVERS; GRIFFITHS, 2007).

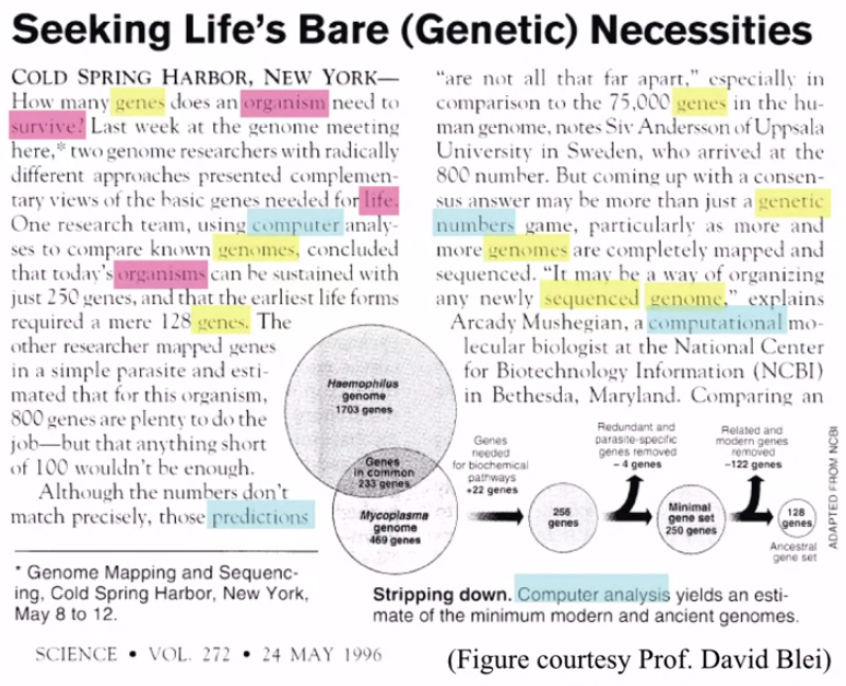
Conforme STEYVERS; GRIFFITHS (2007), um modelo de tópico é um modelo generativo para documentos. Baseado em métodos probabilísticos que especifica um procedimento de como palavras em documentos podem ser geradas com base em variáveis latentes. Variáveis latentes são variáveis que não são diretamente observadas, mas possíveis de serem inferidas a partir de outras variáveis observáveis. O intuito é descobrir um conjunto de variáveis latentes que possam explicar os dados observados. E para este fim, utiliza-se técnicas probabilísticas que fazem o processo reverso (STEYVERS; GRIFFITHS, 2007).

Em outras palavras, modelagem de tópicos assume que qualquer texto é formado selecionando palavras pertencentes a um conjunto de palavras, denominado na área de modelagem de tópicos de saco de palavras. Cada saco de palavras corresponde a um tema (variável latente). Partindo deste princípio, torna-se possível decompor matematicamente um texto nos sacos de palavras prováveis, analisando de onde as palavras (dados observados) vieram primeiro. Este processo é realizado diversas e diversas vezes, até gerar à distribuição mais provável de palavras, em sacos de palavras, que serão os temas (tópicos).

2.1 Tópicos

Tópicos são formados por distribuições probabilísticas de palavras. Um conjunto de palavras que pela relação de ordem, frequência e semântica representam determinados assuntos (temas). Assim, por essas relações, é possível definir um tema como um tópico, ou seja, uma

distribuição probabilística de palavras com frequência e semântica que fazem sentido dentro do contexto do tópico.



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

(Figure courtesy Prof. David Blei)

Figura 2.1: Exemplo de Modelagem de Tópicos (BLEI, 2012).

Segundo BLEI (2012), a partir de um texto, como do artigo *Seeking Life's Bare (Genetic) Necessities* vide Figura 2.1 que se refere à análise de dados para determinar o número de genes, é possível analisar palavras e inferir os principais tópicos contidos no texto.

Para definir um tópico, é necessário demonstrar a frequência e semântica das palavras dentro do tema. Com este intuito quatro tipos de cores diferentes foram utilizados, sendo que, cada cor representa um tema. A cor amarela para palavras referente à genética, em azul palavras sobre computação, rosa sobre biologia evolucionária e em verde sobre a anatomia.

Cada palavra com sentido dentro do contexto do tema foi demarcada com uma determinada cor. Por exemplo, o tema de cor amarela corresponde ao tópico Genética, contém *top-3 words*, *gene*, *dna* e *genetic* sendo as três palavras com maior probabilidade dentro do tema/tópico.

Deste modo, a Figura 2.2 apresenta possíveis tópicos pertinentes ao artigo. Cada tópico é formado por um conjunto de palavras com diferentes probabilidades de frequência e semântica que fazem sentido dentro do contexto do tópico.

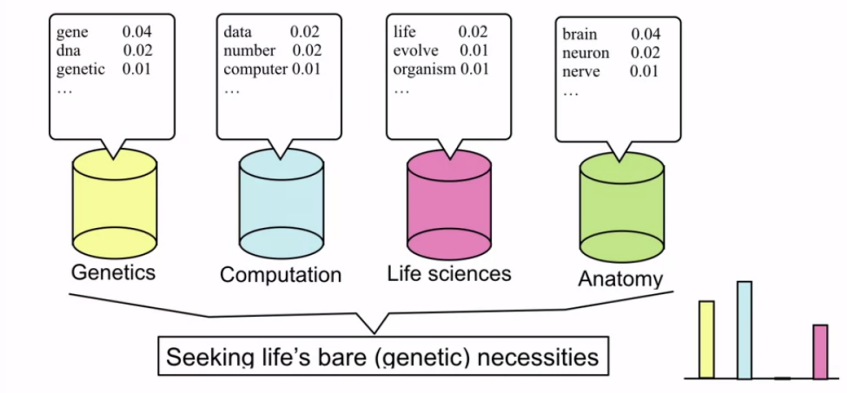


Figura 2.2: Exemplo de Modelagem de Tópicos. Gráficos de tópicos (BLEI, 2012).

2.2 Documentos

A modelagem de tópicos se baseia na ideia de que documentos são misturas de tópicos, ou seja, documentos exibem múltiplos tópicos (STEYVERS; GRIFFITHS, 2007). Com esse conceito, documentos podem ser gerados a partir de diferentes distribuições sobre tópicos. Um documento pode ser desde um artigo ou comentário em uma rede social, textos longos ou curtos. Logo, uma coleção de documentos é formada por inúmeros documentos.

Em modelagem de tópicos, grande parte das abordagens de algoritmos assumem o documento como uma *bag-of-words*, um saco de palavras em que a ordem das palavras contidas no documento não importa. Para isto, inicialmente é realizado um pré-processamento sobre o documento, para que, posteriormente seja utilizado por um modelo de tópico.

O pré-processamento pode consistir as seguintes etapas:

- Remoção de *stop-words*: técnica que remove palavras consideradas irrelevantes dentro conjunto de Documentos, exemplo como, pronomes, artigos, preposições e plurais;
- *Tokenization*: transformação de um texto em apenas palavras. Exemplo, o texto "Escola Regional de Banco de Dados" para [Escola, Regional, Banco, Dados];
- *Stemming*: técnica para cortar palavras com sufixos como "conseguindo" para "conseguir";
- *Lemmatization*: utiliza-se do vocabulário para padronizar palavras, exemplo "Estou, Estás, Está" para "Estar".

2.3 Algoritmos

Os modelos probabilísticos de tópicos buscam descobrir estruturas temáticas ocultas em grandes coleções de documentos. Dentre os modelos, destaca-se o *Latent Dirichlet Allocation (LDA)* (BLEI, 2012).

2.3.1 *Latent Dirichlet Allocation (LDA)*

A forma mais simples de modelar tópicos é através da Alocação Latente de *Dirichlet* e serve como base para várias outras abordagens de extração de tópicos. O *LDA* se caracteriza por atribuir as palavras de cada documento como variáveis observáveis e as variáveis ocultas como a estrutura de tópico.

A alocação de tópicos é realizada utilizando-se da distribuição de *Dirichlet*, que considera que todos os documentos dentro da coleção compartilham o mesmo conjunto de tópicos. Porém, para cada documento, cada tópico possui uma diferente probabilidade de pertencer ao documento. Segundo BLEI (2012), o problema central computacional em modelagem de tópicos é inferir a estrutura de tópicos ocultos, a partir dos dados observados. Para isto, *LDA* realiza adaptações a fim de se aproximar da distribuição condicional das variáveis ocultas dados os documentos.

O *LDA* compreende que os documentos são *bag of words* e assume que o número de tópicos é conhecido e não sofrerá modificações. Em outras palavras, um modelo generativo, em que cada documento é gerado palavra por palavra, pela combinação de tópicos.

O Algoritmo 1 descreve um pseudocódigo do modelo generativo para *LDA*. O primeiro laço se refere à criação da distribuição de tópicos em todos os documentos, o segundo laço distribui os tópicos para cada documento e o terceiro laço repete a distribuição de tópicos internamente para as palavras de um documento, o responsável por realizar a mistura dos tópicos.

Algoritmo 1: Modelo generativo para LDA

```
1  início
       /* Nível de tópicos:                                          */
2      para cada tópico k ∈ [1, k] faça
3          aplica a priori da distribuição de Dirichlet relacionado a tópicos em todos os
           documentos.
4      fim
       /* Nível de Documentos:                                       */
5      para cada documento d ∈ [1, D] faça
6          aplica a priori da distribuição de Dirichlet relacionadas aos tópicos para
               cada documento /* Nível de palavras:                 */
7          para cada palavra n ∈ d faça
8              aplica a distribuição dos tópicos internamente para as palavras de um
               documento e realiza mistura de tópicos.
9          fim
10     fim
11 fim
```

2.4 Métricas

Métricas são instrumentos que podem ser utilizados para medir e avaliar resultados. Em modelagem de tópicos existem dificuldades em relação a avaliar resultados pois os conjuntos de dados não possuem rótulos para conferência da coerência alcançada. A avaliação realizada poderia ser feita por humanos, entretanto, é uma tarefa muito cara e custosa para se aplicar (BLEI, 2012).

Nesse contexto, o trabalho de RöDER; BOTH; HINNEBURG (2015) implementa métricas já propostas por outros autores a fim de verificar a coerência. Dentre as métricas propostas, são destacadas $PMI, NPMI, C_{UCI}, C_{NPMI}$ e C_v baseados em *sliding window* e *context window*.

A técnica de *sliding window* consiste em dividir uma sequência de caracteres em subconjunto de palavras consecutivas de tamanho N. Em que se desloca para qualquer direção sobre as palavras. Assim, considerando um conjunto de palavras $C = \{w_1, w_2, w_3, w_4, w_5, w_6\}$ com uma *sliding window* de tamanho 3, poderá conter a janela $W = \{w_2, w_3, w_4\}$ e deslizar para $W = \{w_3, w_4, w_5\}$. No *context window*, cria-se um subconjunto com *N* palavras consecutivas, no qual antecedem ou sucedem uma determinada palavra. Considerando uma janela de tamanho 2 sobre w_4 a janela se apresentará como $W = \{w_2, w_3, w_4, w_5, w_6\}$.

2.4.1 *Pointwise Mutual Information* (PMI)

Pointwise Mutual Information (PMI) é uma abordagem utilizada para mensurar a associatividade entre duas palavras. O método P(w_i, w_j) considera a frequência/probabilidade de se observar as palavras w_i e w_j ocorreram na mesma janela de palavras. Logo P(w_i) e P(w_j) as probabilidades de w_i e w_j ocorrerem individualmente. A fórmula é dada a seguir:

$$PMI(w_i, w_j) = \log\left(\frac{P(w_i, w_j) + \epsilon}{P(w_i) \cdot P(w_j)}\right)$$

A constante ϵ é utilizada para evitar a ocorrência de logaritmo de zero.

2.4.2 *Normalized Pointwise Mutual Information* (NPMI)

NPMI (*Normalized Pointwise Mutual Information*) é uma variação da PMI que normaliza o valor obtido para o intervalo [-1,1]. Onde 1 indicará completa ocorrência entre as palavras, -1 nenhuma co-ocorrência completa e 0 significa independência entre as duas palavras.

2.4.3 *UCI Coherence*

A métrica $UCICoherence$ utiliza uma $sliding\ window$ de tamanho 10 e aplica o PMI para calcular a coerência sobre todos os pares das $N{-}top\ words$ de um tópico, definida pela seguinte fórmula:

$$C_{UCI} = \frac{2}{N \cdot (N-1)} \cdot \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} PMI(w_i, w_j)$$

2.4.4 *NPMI Coherence*

Esta métrica é uma versão que utiliza NPMI invés de PMI e funciona de maneira semelhante, definida pela seguinte fórmula:

$$NPMICoherence = \frac{2}{N \cdot (N-1)} \cdot \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} NPMI(w_i, w_j)$$

2.4.5 C_v *Coherence*

Essa métrica utiliza uma variação da NPMI e *sliding window*, para calcular a co-ocorrência de cada palavra do tópico com todas as palavras deste tópico. O resultado é um conjunto de vetores, um para cada palavra. C_v utiliza como medida de similaridade, o cosinus,

para o vetor de cada palavra, com o vetor resultante da soma dos vetores de cada palavra do tópico. A média aritmética dessas semelhanças resulta na coerência.

No estudo de RöDER; BOTH; HINNEBURG (2015) é apresentado um conjunto de experimentos avaliando o desempenho das métricas citadas. Dentre todas as métricas, C_v demonstrou a melhor medida de coerência. A métrica C_v será utilizada neste trabalho para verificar o quanto representativo é um conjunto de tópicos.

3 REGISTROS ELETRÔNICOS DE SAÚDE

Desde o início do processo de informatização na área de saúde, os registros eletrônicos sobre a situação de saúde dos pacientes tem grande relevância. O registro eletrônico de saúde tem o conceito de um repositório de informações processáveis sobre o cuidado em saúde do indivíduo, armazenadas e transmitidas de forma segura e acessível por múltiplos profissionais. Tem como principal objetivo oferecer apoio a cuidados de saúde de qualidade, eficazes, seguros e integrados, ao longo de toda a vida do paciente (PANITZ, 2014).

3.1 Prontuário do paciente

O prontuário do paciente passou a ser incorporado como uma prática médica regular no final do século XVIII. Os registros médicos já eram realizados, entretanto não existia um padrão definido. Apresentavam relatos de casos e fatos considerado relevantes aos olhos dos médicos (PANITZ, 2014).

A partir do seu surgimento, o prontuário do paciente se tornou um instrumento importante para a medicina, com fundamentação a observação, classificação, sinais e sintomas verificados, agregando deste modo, o registro sistemático de todos os episódios ocorridos com o paciente durante o atendimento (PANITZ, 2014).

O preenchimento do prontuário é realizado por todas as categorias de profissionais, que fornecem assistência à saúde diretamente aos pacientes, tais como, médico, enfermeiro e assistente social. O prontuário do paciente é definido como o documento único constituído de um conjunto de informações provenientes de outros documentos (PANITZ, 2014).

Estes documentos podem representar a evolução de enfermagem, anotações, exames laboratoriais, sinais e imagens registradas, gerados a partir de fatos, acontecimentos e situações sobre a saúde do paciente e a assistência prestada. O prontuário possibilita a comunicação entre membros da equipe e a continuidade da assistência prestada ao paciente.

Conforme PANITZ (2014), em geral, a estrutura do prontuário do paciente e os dados que o compõem estão dispostos na seguinte forma:

- Identificação do paciente: dados gerais de identificação do paciente atendido nas unidades de saúde, como documentos de identificação oficial;
- Anamnese: entrevista realizada por um profissional de saúde com o paciente ou famili-

ares, que tem o objetivo de coletar informações para um diagnóstico inicial. Os dados mais frequentes são a queixa principal, história da doença atual, epidemiologia de doenças sexualmente transmissíveis, imunização, antecedentes de endemias, acidentes ou violência, antecedentes fisiológicos, antecedentes médicos patológicos, hábitos de vida e antecedentes familiares;

- Exame Físico: medições realizadas para verificar a condição geral do paciente no ato do atendimento. Entre os dados estão o peso, altura, impressão geral, pulso, temperatura, tensão arterial e exame segmentar;

- Diagnóstico: processo analítico ao exame de uma doença ou de um quadro clínico para chegar a uma conclusão. Entre os dados estão a hipótese diagnóstica, lista de problemas e exames complementares;

- Conduta: ações que o profissional de saúde exercerá sobre o caso em específico. Entre os dados estão o plano terapêutico, encaminhamentos e educacional;

- Identificação: espaço dedicado para identificar o profissional de saúde que realiza o atendimento, como assinatura.

Ao final, a estrutura de dados do prontuário se consolida com as informações de evolução clínica, pedido de parecer, prescrição médica, resumo de alta-óbito-transferência, folha de cirurgia e folha de anestesia. Bem como evolução da enfermagem, fisioterapia, serviço social, psicologia e terapia ocupacional.

3.1.1 Prontuário eletrônico do paciente

O conceito de prontuário eletrônico do paciente está estreitamente relacionado com o registro eletrônico de saúde. É o prontuário do paciente que antes estava localizado na central de arquivos médicos em papel e agora disponível no meio digital, possibilitando acesso simultâneo por profissionais de saúde.

O prontuário eletrônico do paciente apresenta diversas vantagens, entre elas a facilidade na recuperação de informação de interesse e disponibilidade de acesso remoto em qualquer local. Além disso, facilita e agiliza o acesso aos dados de atendimentos prévios, intervenções realizadas e história clínica do paciente (PANITZ, 2014).

3.2 Registro eletrônico de saúde x Prontuário eletrônico do paciente

Embora o prontuário eletrônico do paciente e registros eletrônicos de saúde compartilhem a noção fundamental, que é digitalização dos dados e armazenamento das informações sobre os processos de cuidado e saúde dos indivíduos, o conceito de registro eletrônicos de saúde tem amplitude consideravelmente maior em nível de integração e potencial no uso das informações em saúde (PANITZ, 2014).

O registro eletrônico de saúde é uma base de dados que permite que usuários autorizados externos à instituição tenham acesso aos registros médicos dos pacientes produzidos em outras unidades hospitalares. Logo, demonstra a evolução de prontuário eletrônico para o conceito de registro eletrônico, que é o compartilhamento de informações sobre saúde de um ou mais indivíduos, inter ou de multi-instituições, dentro de uma região (município, estado ou país), ou ainda, entre um grupo de hospitais (CMF, 2012).

Com isso, PANITZ (2014) aponta duas características fundamentais do registros eletrônico de saúde. A primeira é a abrangência institucional, que extrapola a instituição onde os dados do prontuário eletrônico são produzidos, para ser um ambiente de integração e compartilhamento de informações interinstitucional. A segunda característica, é o seu grau de dependência das informações produzidas no contexto de cada instituição de saúde isoladamente por meio do prontuário eletrônico do paciente.

Isto implica dizer que a existência de um registro eletrônico de saúde somente é possível se houver a implantação de registros eletrônicos sobre o processo de cuidado à saúde dos pacientes (prontuário do paciente) e engajamento das instituições de saúde a fim de compartilhar esses dados.

3.3 Atendimento em Unidades de Terapia Intensiva (UTI)

A Unidade de Terapia Intensiva (UTI) é o local dentro da unidade hospitalar destinado ao atendimento em sistemas de vigilância contínua a pacientes graves ou de risco, potencialmente recuperáveis. Um paciente é considerado grave quando apresenta instabilidade de algum dos seus sistemas orgânicos, devido a alterações agudas ou crônicas. Já o paciente considerado de risco é aquele que tem alguma condição potencialmente determinante de instabilidade (CREMESP, 1995).

Como o número de leitos em uma UTI é limitado, existem diversos critérios para admis-

são e alta de pacientes. Estes critérios permitem a utilização de modo racional destes leitos de alto custo para pacientes que necessitem quando seu estado de saúde exigir. Além de níveis de prioridade, como, o menor para pacientes menos grave, com alta probabilidade de recuperação, e prioridade maior, quando em fase terminal (CREMESP, 1995).

De acordo com NASCIMENTO; ALVES; MATTOS (2014), os procedimentos realizados em UTI são destinados a prestação de assistência especializada a pacientes em estado críticos, com foco em cuidados qualificados por meio de controle rigoroso dos parâmetros vitais e assistência contínua.

Inicialmente realiza-se a admissão do paciente, e para isso deve seguir um protocolo de admissão, onde deve constar: solicitação de vaga ou leito disponível pela equipe de origem para o médico responsável pelo plantão, e repasse de informações do enfermeiro da unidade de origem para o enfermeiro da UTI. Com isto são realizados os registros de enfermagem, dos quais seguem uma sistematização na assistência de enfermagem implementada, como anamnese, exame físico, definição de diagnósticos e planejamento de enfermagem (NASCIMENTO; ALVES; MATTOS, 2014).

Conforme o estado de saúde do paciente, os próximos procedimentos são escolhidos quando cabíveis de forma singular. Dentre os procedimentos possíveis, pode se destacar a monitoração da pressão intracraniana, intubação traqueal, ventilação mecânica ou suporte ventilatório, monitoração cardíaca e dosimetria de pulso, cateter venoso periférico e central, entre outros.

Todos os procedimentos realizados, bem como a evolução clínica, devem ser registrados pelos profissionais, a fim do compartilhamento e leitura das informações do paciente, de modo imprescindível (BACKES; ERDMANN; BüSCHER, 2015).

3.4 Base de Dados

A coleção de documentos com registros eletrônicos de saúde é proveniente do setor de UTI, especificamente retirados da base de dados *Medical Information Mart for Intensive Care III* (MIMIC III)[1].

MIMIC é uma grande base de dados, disponível gratuitamente, que inclui dados relacionados à saúde de 53.423 internações hospitalares distintas. Foram contabilizados 38.597 pacientes adultos com idade acima de 16 anos e 7.870 pacientes recém-nascidos. Estes pacientes

[1] https://mimic.physionet.org

permaneceram em UTIs do Centro Médico Beth Israel Deaconess em Boston, Massachussets entre 2001 e 2012.

A base de dados é mantida pelo Instituto Tecnológico de Massachussets (MIT). MIT processa todas as informações de identificação do paciente e remove qualquer dado que possa identificar um paciente. Além disso, MIT requer uma formação de proteção de dados do paciente para qualquer pessoa que solicite acesso à base de dados MIMIC.

MIMIC contém mais de 40 gigabytes de dados disponíveis, que representam 26 tabelas relacionais. Nesta seção, serão apresentadas apenas as tabelas utilizadas para extrair os textos utilizados neste trabalho.

A Figura 3.1 traz uma visão geral da base de dados, do processo de captação e formação do conjunto de dados. Os tipos diferentes de dados disponíveis, são apresentadas a seguir:

- **Bilíngue**: são dados codificados, como a Classificação Internacional de Doenças (CID) e entre outros;
- **Descritivo**: detalhes demográficos, datas de admissão e saída;
- **Dicionário**: traduções para identificadores, por exemplo, CID códigos com rótulos associados, a partir de referência cruzada;
- **Intervenções**: procedimentos, tais como diálises, imagiologia e entre outros;
- **Laboratório**: análise sanguínea, hematologia e outros, bem como resultados de testes;
- **Medicação**: registros de administração de medicamentos intravenosos e ordens de medicação;
- **Notas**: notas de texto realizado por profissionais de saúde, como notas de progresso e resumos de alta hospitalar;
- **Fisiológicos**: avaliações, por exemplo, a frequência cardíaca, a pressão sanguínea, taxa respiratória;
- **Relatórios**: relatórios de texto livre de eletrocardiograma e de imagem.

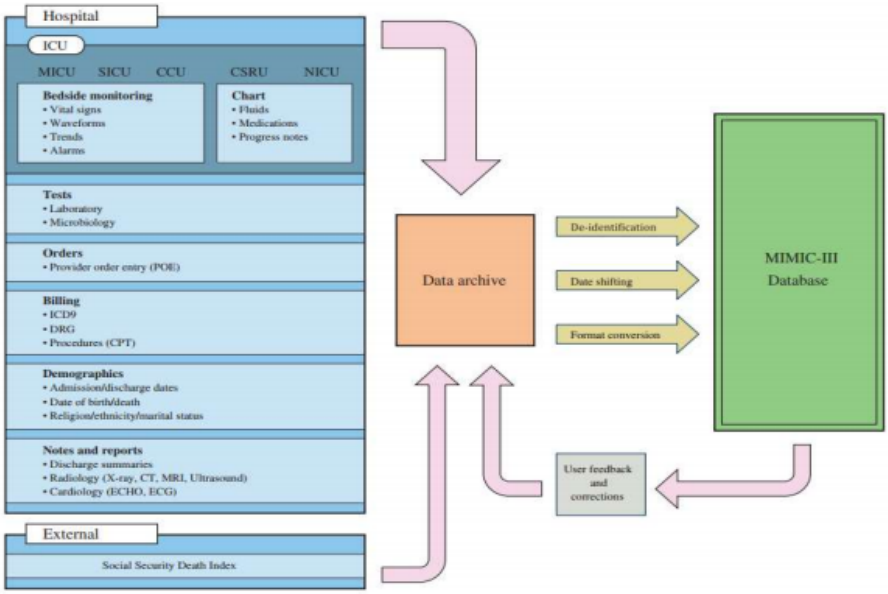


Figura 3.1: Visão geral da base de dados de cuidados intensivos MIMIC III (JOHNSON et al., 2016).

As tabelas são ligadas por identificadores que usualmente contém o sufixo ID como o campo `SUBJECT_ID`, identificador único para tabela `PATIENT`. A Figura 3.2 apresenta todas as ligações da base de dados utilizando um diagrama de classe de relacionamento entre as entidades.

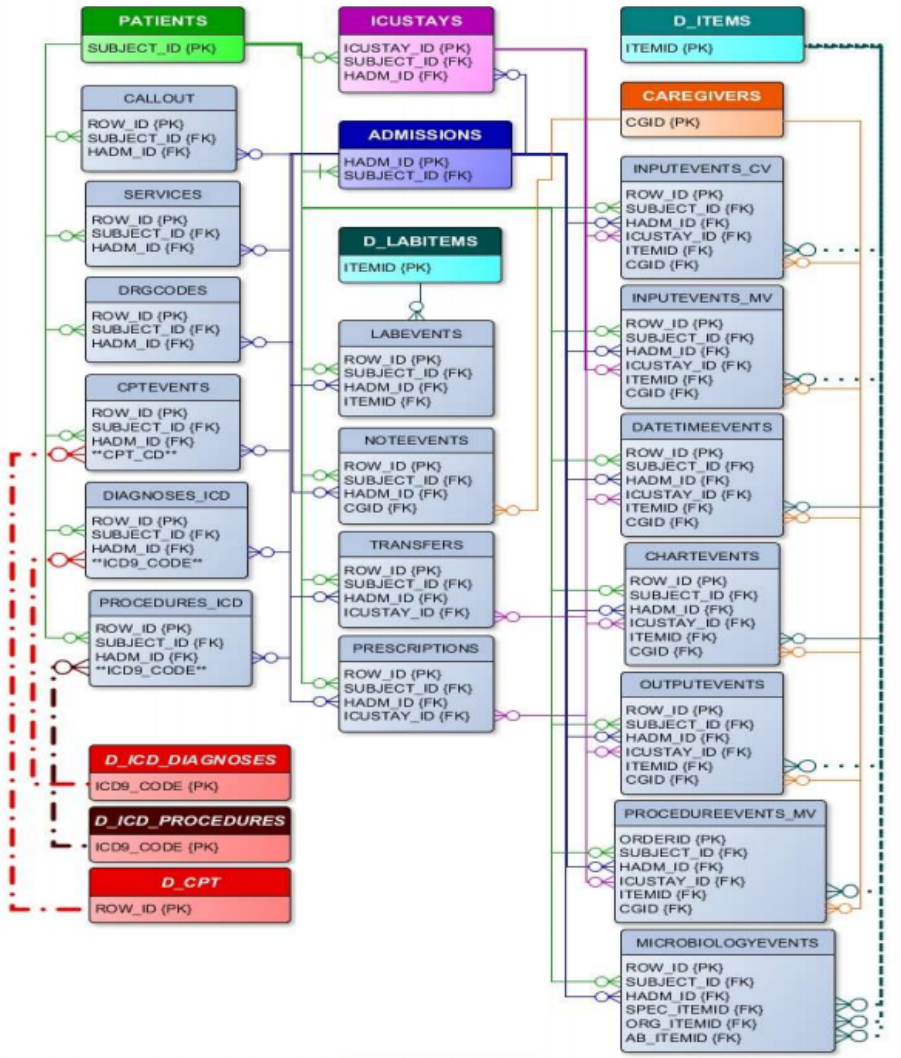


Figura 3.2: Diagrama de Classes relacionamento entidade (MILES, 2017).

Uma das tabelas pertencentes à base de dados que é utilizada na formação do conjunto de dados deste trabalho é a tabela `ADMISSIONS`, que representa internação do paciente. Cada internação é associada a um único identificador, representado pelo atributo `HADM_ID`. Em `ADMISSIONS` há o atributo `DIAGNOSIS`, o qual descreve o diagnóstico preliminar em texto livre para o paciente na internação hospitalar. O diagnóstico é geralmente atribuído pelo médico de plantão e não usa uma ontologia sistemática, ou seja, descrevem sinais/sintomas e possíveis suspeitas. Finalmente, o atributo `HOSPITAL_EXPIRE_FLAG` que indica se o paciente evoluiu a óbito ou alta hospitalar durante a internação. Contém os valores discretos 1 para indicar morte e 0 alta hospitalar.

Cinco das tabelas apresentadas na Figura 3.2 correspondem ao dicionário de dados do MIMIC III: `D_CPT`, `D_ICD_DIAGNOSES`, `D_ICD_PROCEDURES`, `D_ITEMS` e `D_LABITEMS`. Essas tabelas são utilizadas para referência cruzada com outras tabelas, pois

contém as descrições para os códigos.

Dentre estes dicionários, o D_ICD_DIAGNOSES define os códigos da Classificação Internacional de Doenças Versão 9 (CID-9) para diagnósticos. Composto pelos atributos SHORT_TITLE e LONG_TITLE, que fornecem a descrição para cada código. Códigos são atribuídos no final da estadia do paciente e ficam contidos na tabela DIAGNOSES_ICD, na qual são identificados o pacientes, a internação e o código do diagnóstico. Um paciente tem pelo menos 1 ou *N* diagnósticos, partir disto, a tabela possui a coluna SEQ_NUM que fornece a ordem que os diagnósticos são elencados ao paciente e ordenados por prioridade.

NOTEEVENTS mantém todas as notas para o paciente. Possui o atributo CATEGORY que define o tipo da nota salva, por exemplo, em CATEGORY o valor *Discharge* indica uma nota de alta. O atributo TEXT é propriamente a nota, em texto livre, que apresenta todas as ações e resultados sobre o paciente realizadas em um determinado evento. Por Exemplo, em uma nota *Discharge* conteria o resumo de alta-hospitalar.

PROCEDURES_ICD contém os procedimentos realizados durante a internação do paciente. Fornece o código da CID-9 para o procedimento especificado, que pode ser associado à tabela D_ICD_PROCEDURES para determinar qual procedimento é registrado para o paciente. A D_ICD_PROCEDURES contém os atributos SHORT_TITLE e LONG_TITLE que fornecem a definição para o código de procedimento realizado.

Assim, as coleções de documentos utilizados neste trabalho é formado pelos atributos apresentados acima, em que cada documento representa uma internação hospitalar. Os detalhes das coleções de documentos geradas são apresentados no Capítulo 5.

4 TRABALHOS RELACIONADOS

A utilização de coleções de documentos para extração de tópicos pode ser encontrada em diversos trabalhos na literatura. BLEI (2012) apresenta uma revisão da literatura com os trabalhos mais relevantes. Este estudo tem como foco a análise de registros eletrônicos de saúde, de modo que foram selecionados estudos similares, com aplicação de modelagem de tópicos sobre registros gerados na área de saúde.

4.1 *Discovering Associations Among Diagnosis Groups Using Topic Modeling*

LI et al. (2013) realizou um estudo aplicando o algoritmo de modelagem de tópicos *LDA* para agrupar grupos de códigos de diagnósticos, especificamente códigos CID, de pacientes de *Rochester Epidemiology Projects* (REP). REP é uma infraestrutura de pesquisa única em que os registros médicos de pessoas residentes em Olmsted County, Minnesota, foram associados e arquivados.

Inicialmente foram geradas distribuições de tópicos para registros médicos selecionados em uma determinada população. A distribuição de tópicos sobre grupos de códigos de diagnósticos tem o intuito de mensurar a ligação de um grupo de doença com um tópico específico.

Com um total de 4.644 pacientes com grupos de diagnósticos, o *LDA* foi utilizado para gerar distribuições de tópicos, variando de 20 a 147 tópicos.

No estudo de LI et al. (2013) foram identificados que 20 tópicos estão conectados à grupos de doenças. No entanto, foi observado que o mesmo grupo de diagnóstico pode se enquadrar em mais de um tópico diferente.

Os resultados e análises sobre os tópicos foram separados em termos de relação de doenças e de grupos de pacientes. Para relação de doenças, os resultados indicaram que a modelagem de tópicos pode gerar tópicos estatisticamente significativos a esse grupo e identificaram doenças que compartilham alguns pontos comuns.

Em termos de agrupamento de paciente, apresentou divergência na contagem real de grupos de diagnóstico para a proporção correspondente para cada tópico. Esta diferença mostra que alguns grupos de diagnóstico têm mais tópicos do que outros, ou seja, para algumas doenças, os pacientes têm que passar por mais avaliações. Assim, as análises de distribuição de tópico podem revelar a natureza sutil de doenças.

LI et al. (2013) concluiu que o *LDA* tem o potencial de ampla aplicação em epidemio-

logia, bem como, em outros estudos biomédicos, devido a não ser supervisionado e de grande poder interpretativo.

4.2 *Redundancy-Aware Topic Modeling for Patient Record Notes*

No estudo de COGEN et al. (2013) é apresentada uma variação do algoritmo *LDA*, o Red-LDA. O objetivo do Red-LDA é tratar a situação de redundância em anotações clínicas. Esta variação leva em conta o fato de que as palavras são copiadas de um documento de origem nos registros de pacientes.

COGEN et al. (2013) afirma que anotações clínicas em um determinado registro do paciente contém muita redundância. Em grande parte devido ao hábito de documentação, copiando anotações de notas anteriores e colando em uma nova nota de registro.

LDA considera que cada documento é produzida por uma mistura de tópicos e cada palavra é produzido por um tópico. Ao contrário de *LDA*, *Red-LDA* assume que cada documento pertence a um conjunto pré-definido (prontuário do paciente) e que algumas de suas palavras são amostradas a partir de outro documento, o documento de origem para o registro. Em cada conjunto de documentos, um é o documento de origem e os outros são documentos que contém redundâncias copiados do documento original.

Para avaliar o modelo, foi realizada uma comparação com outros métodos de modelagem de tópicos, de acordo com duas métricas quantitativas para avaliação de modelagem de tópico, *log-likelihood* e *topic coherence* e uma análise qualitativa dos temas gerados.

Os resultados indicaram que *Red-LDA* produz temas que são mais coerentes. No estudo de COGEN et al. (2013) também é destacado que *Red-LDA* permite criar modelos superiores para características de registros eletrônicos de saúde, em comparação a outros modelos.

4.3 *Survival Topic Models for Predicting Outcomes for Trauma Patients*

No estudo de ZHANG; JIANG; PETZOLD (2017) é proposto um novo modelo chamado *Survival Topic Model* (*SVTM*). *SVTM* gera tópicos sobre pacientes utilizando dados como medições, notas e informação de óbito/alta, para formar o conjunto de dados. Prevê a probabilidade de óbito/alta em função do tempo. Assim, aplica o modelo para previsão de resultados de pacientes com trauma.

O intuito é que cada paciente tenha uma distribuição latente de condições de doença, ou

seja, tópicos. *SVTM* tem como inspiração o *LDA*, mas ao invés de uma distribuição sobre os tópicos, cada paciente é representado por uma distribuição latente de condição de doença.

O modelo consiste de três partes: um submodelo das notas, um submodelo das medições e um submodelo da sobrevivência. Para o submodelo das notas a coleção de documentos é formada por notas de médicos e enfermeiros, que gera inicialmente a distribuição de tópicos.

O submodelo de medições tem a coleção formada com dados de medições e testes realizados no hospital. A cada medição do conjunto de dados, os mesmos são incluídos na distribuição de tópicos resultantes do submodelo de notas. O estudo de ZHANG; JIANG; PETZOLD (2017) justifica então, que o conjunto de tópicos gerados representa melhor a natureza da condição do paciente.

O submodelo de sobrevivência tem o conjunto formado por dados demográficos de data e histórico de data, que se aplica o modelo de regressão de Cox, que inclui relações lineares com tópicos. O estudo de ZHANG; JIANG; PETZOLD (2017) assume então que os tópicos podem ser amostrados a partir de uma distribuição normal multivariada. Com isso, as relações entre tópicos podem ser aprendidas a partir da matriz de covariância. Este estudo também demonstrou a aplicação sobre o conjunto de dados MIMIC III. Os resultados geraram 25 tópicos, em dados sobre pacientes que sofreram trauma, no qual identificaram interessantes tópicos analisando as *top 7-words* de cada tópico.

Foi possível verificar diferentes tipos de traumas a partir dos tópicos sobre as notas. Dentre estes tipos, foi visto que estavam altamente correlacionadas uma com as outras, o que pode ser explicado, pois os traumas correlacionadas estavam localizados próximos uns dos outros no corpo humano. Assim, pacientes poderiam apresentar múltiplos traumas, e consequentemente, várias avaliações realizadas para estes.

4.4 *Using topic modeling to infer the emotional state of people living with Parkinson's disease.*

O estudo de VALENTI et al. (2019) utiliza modelagem de tópicos para inferir o estado emocional de pessoas que vivem com a doença de Parkinson. Se utilizou de dois modelos para este propósito: *LDA* e *Linguistic Inquiry and Word Count* (*LIWC*) a fim de avaliar os desempenhos.

A coleção de documentos é proveniente de entrevistas durante um ensaio clínico aleatório realizado anteriormente. Os dados incluíam respostas às perguntas abertas sobre aconteci-

mentos da vida diária no passado recente, os quais os participantes tinham experimentado como particularmente frustrantes ou agradáveis.

A coleção de documentos foi dividida da seguinte forma: 90% em um conjunto de treino e 10% para o conjunto de teste. O passo seguinte foi a geração de tópicos com K tópicos inicializado com os seguintes valores: 4,16,24,34,44,50,64 e 91. Para cada conjunto de tópicos gerados, cada um deles ganhou um rótulo, com valores, frustrante = 0 ou agradável = 1.

A partir disso, com um conjunto de par-valor (Tópicos, Rótulos) foi utilizado o modelo de classificação regressão logística. Os modelos foram avaliados utilizando métricas padrão de desempenho de aprendizagem de máquina, tais como, a precisão, recall, e F_1-score.

Os resultados mostraram que o *LDA* é adequado quando a contagem de palavras em um documento é aproximadamente o da sentença média. Nesse caso, o modelo *LDA* corretamente contribui na predição correta da categoria 86% do tempo e *LIWC* apenas e 29% do tempo.

5 PROJETO DE EXPERIMENTO

Este capítulo descreve as etapas necessárias para execução do experimento. A seguinte ordem de execução dos experimentos é apresentada: configuração dos ambientes, geração das coleções de documentos (óbito e alta), pré-processamento das coleções e extração de tópicos. As aplicações desenvolvidas estão disponíveis na plataforma de hospedagem de código-fonte GitHub [2]. De forma detalhada cada etapa é descrita a seguir.

5.1 Configuração do ambiente

O ambiente para geração da coleção de documentos foi baseado em um notebook DELL Inspiron 5566 com o processador Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz, 2701 Mhz, 2 Núcleos, 4 Processadores Lógicos, 4GB de memória RAM e disco SSD de 256GB com sistema operacional Windows 10. Para as etapas posteriores, o ambiente utilizado foi um servidor Linux com o processador Intel(R) Xeon(R) CPU E7- 4850 @ 2.00GHz 1064.444 Mhz, 10 Núcleos, 80 Processadores, 126 GB de memória RAM e 6 terabytes de disco.

5.2 Geração das coleções de documentos (Óbito e Alta)

A base de dados se divide em 26 tabelas no formato CSV que ao total somam 45GB de dados. Para gerenciamento e visualização dos dados, houve a necessidade de utilizar um sistema gerenciador de banco de dados objeto relacional.

Para isto foi utilizado o *PostgreSQL*, um sistema de banco de dados relacional de objetos, de código aberto com desenvolvimento ativo. No *PostgreSQL* ocorreu a carga de dados das tabelas, resultando em uma instância com tamanho 46.5GB de memória.

Inicialmente, para extração dos documentos e consequentemente a construção das coleções, houve a seleção de tabelas e atributos para formar um documento. Foram escolhidas tabelas que continham atributos em formato de texto e apresentavam relevância nos dados observados.

Um documento é composto pelos seguintes dados: da tabela `ADMSSION` foram retirados do atributo `DIAGNOSIS`, contendo o diagnóstico inicial da internação. Da tabela `D_ICD_DIAGNOSIS` todos os diagnósticos elencados durante a internação informados no atri-

[2] https://github.com/Ivairpuerari/TCC-Ivair

buto LONG_TITLE. Da Tabela D_ICD_PROCEDURES foi extraído o atributo LONG_TITLE contendo os procedimentos realizados. Por fim, as notas de eventos durante a internação foram retirados da tabela NOTEEVENTS atributo TEXT.

Para extração destas informações e geração das coleções foi desenvolvido um programa em *Python* (versão 3) *collection.py*, com o intuito de realizar a integração com o *PostgreSQL* pela busca dos dados e separação das internações em óbito e alta. A busca foi realizada individualmente em cada tabela, selecionado o dado e a identificação única da internação. As consultas implementadas com os dados selecionados são demonstrados na Figura 5.1.

```
SELECT hadm_id,diagnosis
FROM admissions
WHERE admissions.hospital_expire_flag = {}

SELECT admissions.hadm_id, d_icd_diagnoses.long_title
FROM admissions
JOIN diagnoses_icd
    ON diagnoses_icd.hadm_id = admissions.hadm_id
JOIN d_icd_diagnoses
    ON d_icd_diagnoses.icd9_code = diagnoses_icd.icd9_code
WHERE admissions.hospital_expire_flag = {}

SELECT admissions.hadm_id, d_icd_procedures.long_title
FROM admissions
JOIN procedures_icd
    ON procedures_icd.hadm_id = admissions.hadm_id
JOIN d_icd_procedures
    ON d_icd_procedures.icd9_code = procedures_icd.icd9_code
WHERE admissions.hospital_expire_flag = {}

SELECT admissions.hadm_id, noteevents.text
FROM admissions
JOIN noteevents
    ON noteevents.hadm_id = admissions.hadm_id
WHERE admissions.hospital_expire_flag = {}
```

Figura 5.1: Consultas SQL para busca na base de dados.

Para selecionar apenas internações que evoluíram a óbito ou alta, o filtro utilizado nas consultas foi o atributo HOSPITAL_EXPIRE_FLAG, no qual o valor 1 representa óbitos e 0 para altas.

Com os dados retornados, a estrutura de dados utilizada foi o *Dictionary* disponível no *Python*, que tem a representação de uma coleção de *{Key, Value}*. No experimento cada *Dictionary* representa uma coleção de documentos para óbito ou alta, com a *Key* sendo a identificação única para cada internação e *Value* contendo os dados hospitalares retornados das consultas, concatenados por um espaço em branco. Assim, é representado um documento para uma única internação.

A Figura 5.2 apresenta a função responsável pela rotina descrita acima. Ao final, cada *Dictionary* alta e óbito, representa a coleção de documentos. A partir destes dois dicionários, foram gerados dois arquivos TXT referentes às coleções de documentos óbito e alta.

```
def insertAdmissionLife(sql):
    sql = sql.format('0')
    print(sql)
    rs = db.prepare(sql)
    for row in rs:
        line = str(row[1]).replace('\n', ' ')
        try:
            admissionLife[row[0]] += ' '+line
        except:
            admissionLife[row[0]] = line
```

Figura 5.2: Função que gera um documento e coleção de documentos.

O arquivo referente à coleção de óbitos é composto por 6.051 documentos, com a média de 10.931 palavras por documento e ao total, um tamanho 500MB da coleção. Já a coleção de altas possui 53.954 documentos, com a média de 7.174 palavras por documento e 2,77GB de tamanho. Percebe-se que na coleção a maioria dos registros são relacionados à alta, representando quase 90% dos documentos das coleções.

O prontuário do paciente como definido anteriormente é formado pela concatenação de todos os documentos gerados durante a internação. O número médio de documentos utilizados para gerar um documento na coleção foi de 30 documentos para a coleção correspondente a altas e 43 para a coleção de óbitos.

STATUS EPILEPTICUS Grand mal status Hodgkin's disease, unspecified type, unspecified site, extranodal and solid organ sites Postinflammatory pulmonary fibrosis Pneumonia, organism unspecified Unspecified essential hypertension Macular degeneration (senile), unspecified Antineoplastic and immunosuppressive drugs causing adverse effects in therapeutic use Other generalized ischemic cerebrovascular disease Open biopsy of brain Continuous invasive mechanical ventilation for 96 consecutive hours or more Spinal tap Venous catheterization, not elsewhere classified Enteral infusion of concentrated nutritional substances Venous catheterization, not elsewhere classified Admission Date: [**2108-8-22**] Discharge Date: [**2108-8-30**] Date of Birth: [**2036-5-17**] Sex: M Service: Neurology HISTORY OF PRESENT ILLNESS:

Figura 5.3: Fragmento de um documento pertencente à coleção de documentos.

5.3 Pré-processamento das coleções

Um documento tem sua maior parte composto por notas sobre os eventos realizados durante a internação no formato de texto livre. Conforme pode ser visto na Figura 5.3, não existe um padrão na escrita, apresentando muitas pontuações e símbolos no meio do texto. Para melhor desempenho e qualidade nos tópicos, faz-se necessário o pré-processamento das coleções.

Com este intuito a coleção passou pelas seguintes etapas de pré-processamento:

- Remoção de todas as pontuações e símbolos especiais;
- Padronização das palavras em formato minúscula;
- Processo *lemmatization* de agrupar as diferentes formas flexionadas de uma palavra para que possam ser analisadas como um único item. Para isso foi utilizada a biblioteca *Natural Language Toolkit (NLTK)* [3], especificamente a função *WordNetLemmatizer*;
- Processo de *Stemmatização* para remover afixos morfológicos das palavras. Disponível na biblioteca *NLTK* a função *PorterStemmer*;
- Remoção de todos os dígitos contidos nas coleções;
- Remoção de palavras com menos de três letras;
- Remoção das palavras consideradas *Stopwords*. Inicialmente foram adicionadas palavras comuns à língua inglesa e como foi utilizado o *stopwords* da biblioteca *nltk.corpus* somente foi necessário configurar o idioma utilizado, ou seja para *English*. Posteriormente,

[3] https://www.nltk.org/

foram adicionadas palavras especificas sobre o assunto consideradas não relevantes, visualizadas a partir dos tópicos gerados. As *stopwords* não comuns, isto é, palavras diferentes de *and*, *with* e *to*, que foram percebidas e removidas são apresentas na Figura 5.4.

size	bypass	lastnam	system
placement	updat	sign	review
function	infant	refills	cultur
action	deni	move	subcutaneous
find	good	site	home
like	swallow	code	patient
final	deliveri	dose	beam
clip	take	unit	during
tablet	call	studi	mgdl

Figura 5.4: StopWords utilizadas no pré-processamento.

O pré-processamento provoca uma diminuição expressiva no tamanho do documento. A Figura 5.5 apresenta o mesmo fragmento do documento da Figura 5.3 no qual fica perceptível o resultado do pré-processamento sobre o documento.

statu epilepticu grand statu hodgkin diseas unspecifi type unspecifi site extranod solid organ site postinflammatori pulmonari fibrosi pneumonia organ unspecifi unspecifi essenti hypertens macular degener senil unspecifi antineoplast immunosuppress drug caus advers effect therapeut gener ischem cerebrovascular diseas open biopsi brain continu invas mechan ventil consecut hour spinal venou catheter elsewher classifi enter infus concentr nutrit substanc venou catheter elsewher classifi admiss discharg birth servic neurolog histori present known lastnam yearold gentleman histori

Figura 5.5: Fragmento de um documento pertencente à coleção de documentos pós-processado.

Após o pré-processamento, a coleção de óbitos possui os mesmos 6.051 documentos, mas com média de palavras 5.352 por documento e tamanho de 219MB de dados. A coleção de altas resultou em 53.954 documentos com média de 3.384 palavras por documento e 1,23GB de dados.

5.4 Extração dos tópicos

Com as coleções de documentos processados a próxima etapa foi extrair os tópicos. Para esta tarefa o modelo escolhido para geração de tópicos foi o *LDA*. Para tal, faz se uso da biblioteca *Gensim*[4]. *Gensim* dispõe do modelo em duas abordagens, *LDAModel* e *LDAMulticore*, sendo que o *LDAMulticore* utiliza de mais núcleos da CPU para paralelizar e acelerar o treinamento do modelo. Desta forma a abordagem escolhida para aplicação foi o *LDAMulticore*.

Inicialmente foi realizada a tokenização dos documentos em que cada documento é dividido em *tokens*, ou seja, cada documento é dividido em palavras. Por exemplo um documento $D = \{\text{renal chronic blood}\}$ se torna $D' = \{\text{'renal', 'chronic', 'blood'}\}$.

A partir disso foram gerados dois arquivos referentes ao dicionário e *corpus* da coleção. O dicionário possui o registro das palavras contidas na coleção de documentos, atribuindo o identificador único para a palavra e o quantidade de ocorrências daquela palavra na coleção.

Na construção do dicionário é aplicado um filtro para que apenas contenha palavras que estejam presentes na maioria dos documentos. Assim, previne-se uma possível distorção de resultados por influência de palavras que estejam em apenas alguns ou em todos documentos. Nos parâmetros para o filtro definiu-se que o limite inferior seria de 10% da quantidade de documentos da coleção e limite superior em 80%, mantendo-se apenas as primeiras 10.000 palavras mais frequentes.

O *corpus* é a combinação de todos os documentos de texto, com representação matricial entre os documentos e termos, resultando na frequência de termos por documento. A Figura

[4] https://radimrehurek.com/gensim/

5.6 é um exemplo da representação da matriz documento-termo para uma coleção de três documentos com três palavras com frequências diferentes para cada documento.

DOCUMENT TERM MATRIX			
	RENAL	BLOOD	CHRONIC
DOC 1	1	2	0
DOC 2	0	1	1
DOC 3	2	0	1

Figura 5.6: Representação matriz documento-termo.

Por fim, a aplicação do modelo *LDA* necessita além do dicionário e o corpus da coleção, informar valores para hiperparâmetros importantes, como: número de iterações, passos, número de tópicos (*K*) e α.

A cada geração de tópicos houve a avaliação por meio da métrica C_v (veja Capítulo 2). Esta foi a medida escolhida para ser possível encontrar os melhores hiperparâmetros e definir o modelo final para extração dos tópicos.

O número de iterações representa o máximo de iterações pelo corpus ao inferir a distribuição de tópicos de um *corpus* e foi definido em 1.000 iterações. O passos é o número de passos através do *corpus* durante o treinamento, o qual foi definido em 50.

Foram definidos 6 experimentos com os seguintes números de tópicos: 5, 10, 15, 25, 50 e 100. Para cada resultado de tópicos foi aplicada a métrica de coerência. A Figura 5.7 apresenta um gráfico com o desempenho de cada valor para o número de tópicos. Observando a figura, percebe-se que após *K* próximo de 20, a coerência dos tópicos diminui.

Deste modo, foram definidos testes com *K* tópicos no intervalo [8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23] que apresentaram maior valor de coerência. No experimento, para cada *K* tópicos presente no intervalo, houve a variação no valor do α em [0.01, 0.31, 0.61, 0.90999, *symmetric*, *asymmetric*] com o intuito de definir o valor do α posteriormente. Os valores *symmetric* e *asymmetric* estão definidos nas equações 5.1 e 5.2 respectivamente.

$$\alpha = \frac{1}{num\_topics} \tag{5.1}$$

$$\alpha = \frac{1}{\sqrt{num\_topics} + 1} \tag{5.2}$$

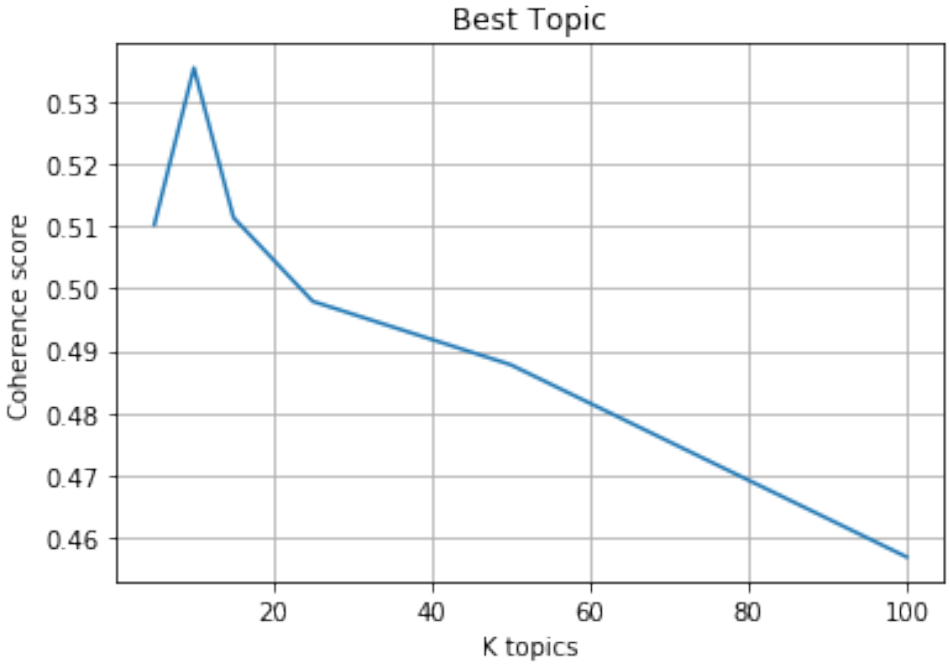


Figura 5.7: Gráfico de coerência de cada tópico inicial.

A Figura 5.8 apresenta os melhores resultados da combinação K e hiperparâmetros. Perceba que o melhor valor para K foi 11, assim, foi definido $K = 11$ para os experimentos deste trabalho. O próximo capítulo apresenta, então, os resultados finais dos experimentos.

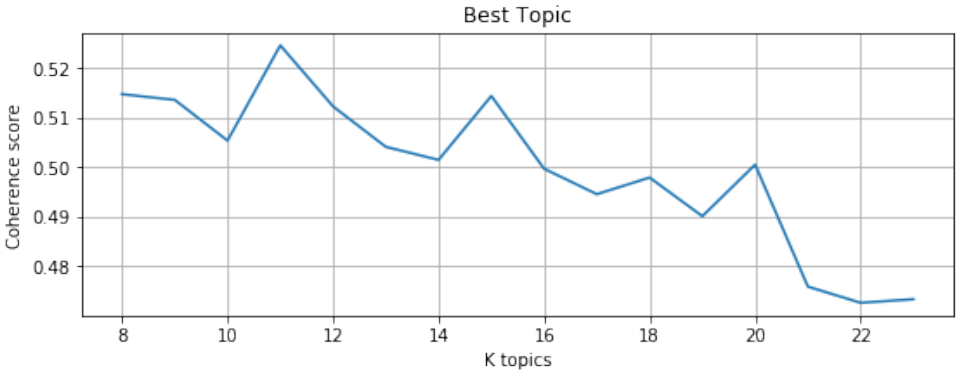


Figura 5.8: Gráfico de coerência de cada tópico.

6 EXPERIMENTO

A partir do projeto de experimento, foram obtidas as coleções de documentos utilizadas neste trabalho. Em seguida, realizou-se o pré-processamento sobre as coleções aplicando as técnicas apresentadas. Ao final, obteve-se em duas coleções de documentos referentes as internações que evoluíram a óbito e que obtiveram alta, baseados em registros eletrônicos de saúde.

Nas próximas seções são apresentados a configuração final dos hiperparâmetros para execução do modelo *LDA*, a extração dos tópicos para cada coleção, a rotulação dos tópicos e os resultados obtidos.

6.1 Execução

Com as coleções de documentos prontas, o próximo passo foi realizar a criação do dicionário e *corpus* para cada coleção. Os parâmetros para o filtro do dicionário foram definidos em 10.000 palavras mais frequentes e que estivessem presentes no mínimo em 10% e no máximo 80% dos documentos.

O dicionário da coleção de óbitos resultou em 1.945 palavras, enquanto o dicionário da coleção de altas teve 1.662 palavras. O *corpus* foi formado por 6.051 documentos para a coleção de óbitos e 53.954 para a coleção de altas.

Logo após, foi realizada a configuração final no modelo *LDA*, com os seguintes hiperparâmetros: *iterações* $= 1000$, $passos = 50$, $K = 11$. Para definir α, houve a aplicação do *LDA* para cada valor de α proposto no projeto de experimento. A Figura 6.1 apresenta as escalas de coerências para α definidos. Percebe-se que α entre 0.01 - 0.31 e *symmetric* apresentam a melhor coerência e aptos a serem escolhidos, neste trabalho foi utilizado $\alpha = 0.01$.

A execução do modelo *LDA* com os hiperparâmetros definidos ocorreu inúmeras vezes, até que um número considerável de palavras irrelevantes (*stopwords*) fossem removidas para cada coleção de documentos. Consequentemente, em termos de coerência o modelo apresentou melhora, com coerência em 55,22% para coleção de óbitos e 62,33% para a coleção de altas.

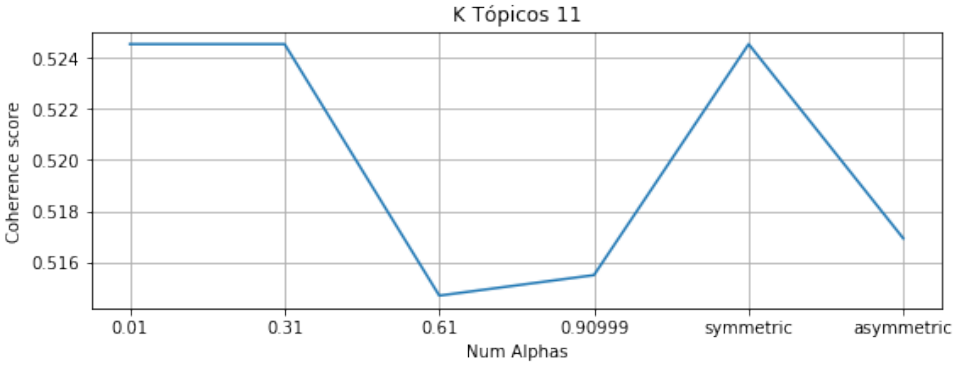


Figura 6.1: Gráfico de coerência de cada α.

6.2 Extração de tópicos

Após a execução do modelo *LDA* foram extraídos 11 tópicos para cada coleção de documentos óbito e alta. Todos os tópicos extraídos para cada coleção são apresentados no Apêndice A deste trabalho em formato de tabela, com as probabilidades das palavras em cada tópico.

Os *hot-topics* são os tópicos mais presentes em uma coleção de documentos e representam os assuntos mais frequentes entre os documentos. As avaliações sobre os tópicos foram realizadas a partir de uma análise das *top-10* palavras do tópico. Dos 11 tópicos de cada coleção, foram considerados *top-5* tópicos como *hot-topics*.

Para descobrir os *top-5* tópicos, foram verificados a quantidade de documentos que um tópico representa, ou seja, para cada tópico foram contabilizados as ocorrências de documentos, sendo que, cada documento tem uma probabilidade diferente de pertencer a um tópico.

Deste modo, documentos podem pertencer a um ou mais tópicos, com probabilidades altas ou baixas. Assim, foi utilizada a metodologia de buscar apenas tópicos que no mínimo contenham 50% de probabilidade de pertencer àquele documento.

A Figura 6.2 apresenta a quantidade de documentos para cada tópico e a probabilidade de documentos que aquele tópico pertence em no mínimo 50%, sobre os documentos da coleção de óbito, enquanto a Figura 6.3 exibe a quantidade de documentos para cada tópico da coleção de altas. Em ambas as figuras, os tópicos estão classificados em ordem decrescente.

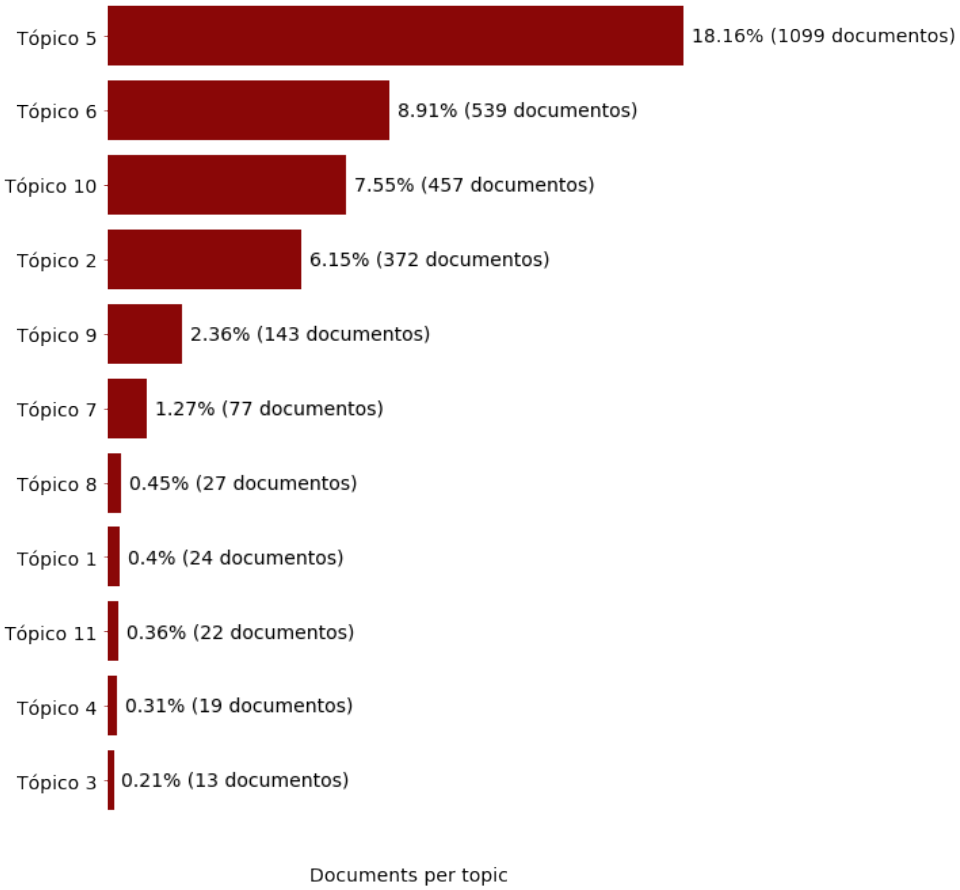


Figura 6.2: Gráfico de documentos por tópicos da coleção de óbitos.

Os *hot-topics* serão descritos a seguir em ordem decrescente e estão apresentados na Figura 6.4 para coleção de óbito e a Figura 6.5 para coleção de altas. Os tópicos estão em forma de nuvem de palavras, sendo que, as palavras com maior frequência e consequentemente relevância aparecem em destaque.

Conforme a Figura 6.2, os *hot-topics* da coleção de óbitos são compostos pelos tópicos 9, 2, 10, 6 e 5. O tópico 9 contém 143 dos documentos da coleção de óbitos, representando 2,3% de documentos na coleção. Já o tópico 2 representa 6,15% da coleção de documentos de óbitos em 372 documentos. O tópico 10 está presente em 457 documentos, no total de 7,55% da coleção, enquanto o tópico 6 em 539 documentos, representando 8,91% da coleção. Por fim o *top-1*, o tópico 5 com 1.099 documentos pertencentes, o qual representa 18,16% da coleção de óbitos.

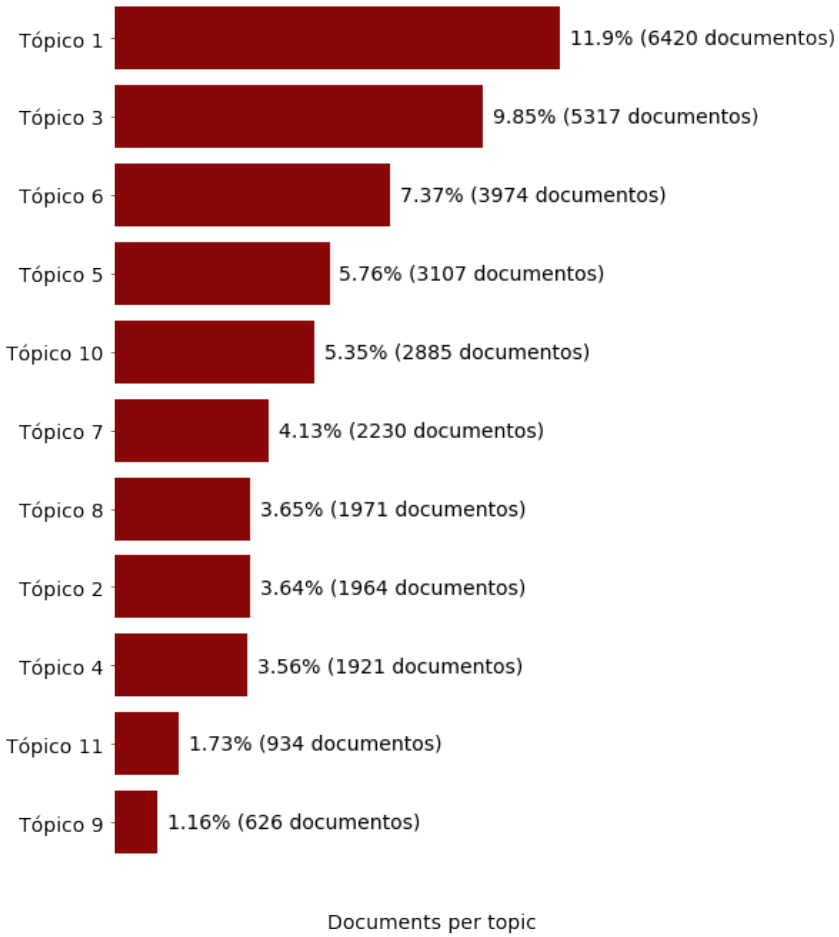


Figura 6.3: Gráfico de documentos por tópicos da coleção de altas.

Os *hot-topics* da coleção de altas são compostos pelos tópicos 10, 5, 6, 3 e 1. O tópico 10 representa 5,35% da coleção de altas, dos quais são 2.885 documentos. Já o tópico 5 está presente em 3.107 documentos e compõe 5,76% da coleção. O tópico 6 contém 3.974 documentos e representa 7,37% na coleção. Para o tópico 3 há 5.317 documentos e 9,85% da coleção. O *top-1* da coleção de alta é o tópico 1 com 6.420 documentos e 11,9% da coleção de altas.

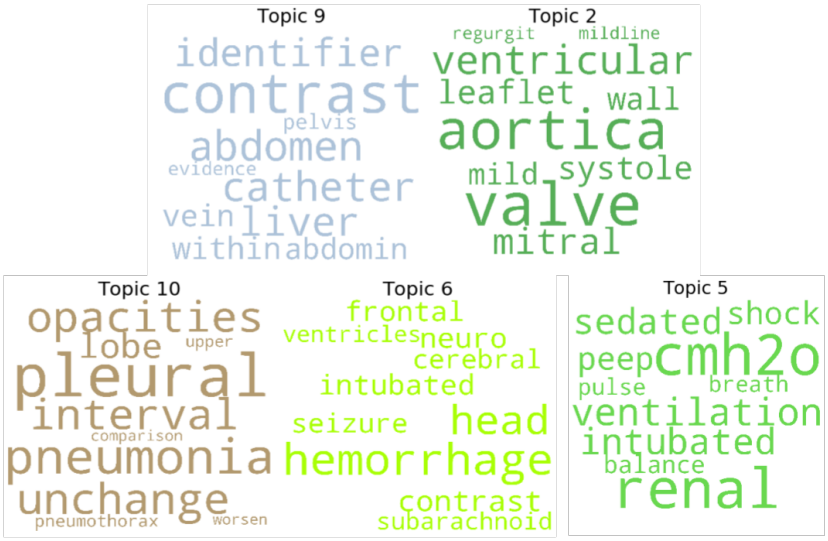


Figura 6.4: *Top-5* tópicos da coleção de óbitos.

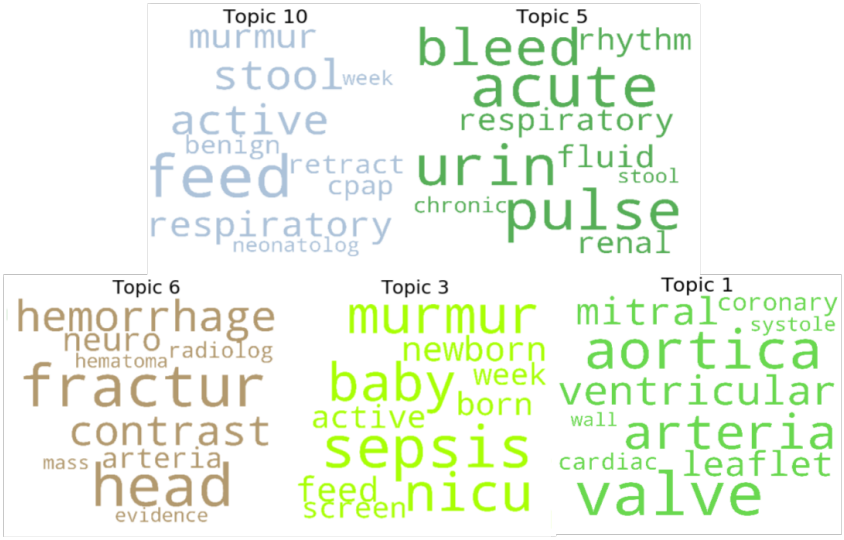


Figura 6.5: *Top-5* tópicos da coleção de altas.

Com a finalidade de rotular os tópicos houve a necessidade de buscar pessoas da área da saúde com vivência no ambiente hospitalar. Assim, a rotulação dos tópicos aconteceu através de discussões com estudantes da 8ª fase do curso de Enfermagem da Universidade Estadual de Santa Catarina (UDESC), que realizaram atividades teórico-prática dentro da UTI do Hospital Regional do Oeste.

Inicialmente foram repassadas as informações necessárias deste trabalho, apresentando um resumo. Após isto, foi realizada a criação de um formulário online para que individualmente fosse avaliado cada tópico para cada coleção de documentos. O formulário foi dividido em duas seções, uma para cada coleção. Cada seção apresentava os tópicos extraídos com as palavras em ordem probabilística de mais frequente para menos frequente, com o questionamento "Qual sistema corpóreo afetado/procedimento/complicação/tratamento corresponde este conjunto de palavras?".

O total de oito estudantes participaram respondendo o formulário, com base nas respostas, foram geradas as tabelas de respostas, representados nas Figuras 6.7 da coleção de óbitos e 6.6 da coleção de altas. As tabelas apresentam os resultados obtidos pelo formulário de avaliação de tópicos.

Assuntos	Tópicos										
	Tópico 1	Tópico 2	Tópico 3	Tópico 4	Tópico 5	Tópico 6	Tópico 7	Tópico 8	Tópico 9	Tópico 10	Tópico 11
Sistema Cardíaco	8										
Diabetes		3									
Pós-operatório		1									
Pâncreas		2									
Hepático		1									
Sistema Hepático/pancreático		1									
Infecção			3								
Prematuridade			4							2	
Cirurgia			1								
Sistema Respiratório				7			8		8	2	8
Labio leporino				1							
Hemorragia					3						
Sistema Cardiovascular					1						
Sistema Renal					4			2			
Sistema Neurológico						5					
TCE						3					
Sistema Hepático								5			
Sistema Gastrointestinal								1			
Sistema Respiratório e Cardíaco										2	
Câncer										2	

Figura 6.6: Tabela de respostas da coleção de alta.

Assuntos	Tópicos										
	Tópico 1	Tópico 2	Tópico 3	Tópico 4	Tópico 5	Tópico 6	Tópico 7	Tópico 8	Tópico 9	Tópico 10	Tópico 11
Sistema Respiratório	2				6		5			6	
Infecção	3		1								3
Sistema Tecidual	1										
Trauma	1										
Queimado/ Politraumatismo	1										
Sistema Cardiovascular		8		7			1	3			1
Sistema Hepático			5						3		
Sistema urinário/renal											1
Sistema Renal			1				2	5	1	1	3
TCE				1							
Sistema Neurológico					1	8					
Respiratório associado a problema neurológico					1						
Sistema Gastrointestinal									2		
Passagem de cateter venoso									1		
Sistema Gastrointestinal/Hepático									1		
Cirurgia			1								
Derrame Pleural										1	

Figura 6.7: Tabela de respostas da coleção de óbito.

Os assuntos respondidos no formulário demonstraram como base em geral os sistemas corpóreos. Para a coleção de alta houve 20 assuntos diferentes para os tópicos da coleção, enquanto a coleção de óbitos apresentou 17 assuntos diferentes elencados. Em média 2 assuntos diferentes foram elencados para cada tópico da coleção de alta, já para a coleção de óbitos apresenta em média 3 assuntos para cada tópico.

De modo geral, os resultados obtidos pela avaliação dos tópicos através do formulário foram considerados condizentes, visto que estavam relacionados ao objetivo desse trabalho. Dessa forma, todos os tópicos extraídos a partir dos assuntos selecionados pelo formulário se mostraram aptos a serem utilizados para definição de assunto (rótulo) para cada tópico.

6.3 Resultados

Nesta seção é realizada a rotulação dos tópicos de cada coleção, ou seja, são identificados qual assunto é representado dentro de cada tópico e os assuntos mais presentes em cada coleção. Ao final, são apresentadas as intersecções e disjunções entre os tópicos pelas coleções.

6.3.1 Tópicos (Assuntos) da coleção de alta

Como já apresentado na Tabela 6.6, o tópico 1 foi interpretado através de seu conjunto de palavras, evidenciando o sistema cardíaco como assunto prioritário. Já o tópico 2 resultou

em diabetes, seguido do tópico 3 identificado o assunto prematuridade, os tópicos 4, 7, 9, 10 e 11 para sistema respiratório, o tópico 5 sistema renal, no tópico 6 com maior evidencia para sistema neurológico e por fim o tópico 8 sistema hepático.

Os assuntos mais presentes na coleção conforme os *hot-topics* são sistema respiratório, sistema renal, sistema neurológico, prematuridade e sistema cardíaco. É possível visualizar que o sistema respiratório está presente em 5 tópicos, totalizando 8.596 documentos, representando 15,93% da coleção de documentos contra 11,90% referente ao sistema cardíaco, sendo que, o assunto sistema cardíaco é *top-1* entre os *hot-tópics*. Resultando em sistema respiratório, o assunto mais presente na coleção de altas se levado em conta a soma de todos os documentos dos tópicos com o assunto sistema respiratório.

O assunto sistema respiratório pertence ao conjunto de tópicos 4, 7, 9, 10 e 11, sendo que, o tópico 10 contém o maior número de documentos deste conjunto com 2.885, ou seja, apresenta a maior influencia na soma de todos os documentos. Analisando os assuntos elencados na Tabela 6.6 para o tópico 10, apresenta prematuridade como um possível assunto, também sistema respiratório e sistema respiratório e cardíaco, ou seja, esses sistemas possivelmente podem estar correlacionados, devido aos assuntos do tópico 10, que levam a compreender como possível causa, um recém-nascido prematuro com sistema respiratório e cardiovascular imaturo.

Deste modo, poderíamos definir o sistema cardíaco e respiratório como os mais presentes na coleção de alta.

6.3.2 Tópicos (Assuntos) da coleção de óbitos

Com relação a coleção de óbitos, a Tabela 6.7 apresentou no tópico 1 infecção como o assunto relevante, enquanto que para os tópicos 2 e 4 o sistema cardiovascular, seguido de sistema hepático para os tópicos 3 e 9, sistema respiratório para os tópicos 5, 7 e 10, sistema neurológico para o tópico 6, sistema renal para os tópicos 8 e 11.

Os *hot-topics*, ou seja, assuntos mais presentes da coleção de óbitos foram sistema hepático, sistema cardiovascular, sistema neurológico e sistema respiratórios como *top-1*. Ao contrário da coleção de altas, houve uma distribuição regular dos assuntos sobre os tópicos, tal que, o sistema respiratório foi o assunto mais presente na coleção.

6.3.3 Intersecções e disjunções das coleções de documentos

Os assuntos encontrados nas coleções de documentos estão agrupados em dois conjuntos de tópicos. Os conjuntos de assuntos estão divididos em assuntos da coleção de altas e coleção de óbitos:

- Altas: Sistema Cardíaco, Diabete, Prematuridade, Sistema Respiratório, Sistema Renal, Sistema Neurológico e Sistema Hepático.
- Óbitos: Sistema Cardíaco, Infecção, Sistema Respiratório, Sistema Renal, Sistema Neurológico e Sistema Hepático.

A Figura 6.8 apresenta as disjunções e intersecções sobres as duas coleções. As intersecções das coleção de documentos acontecem por meio dos assuntos, sistema cardíaco, sistema respiratório, sistema renal, sistema neurológico e sistema hepático. As disjunções se mostram na coleção de altas nos assuntos diabete e prematuridade enquanto na coleção de óbitos o assunto infecção.

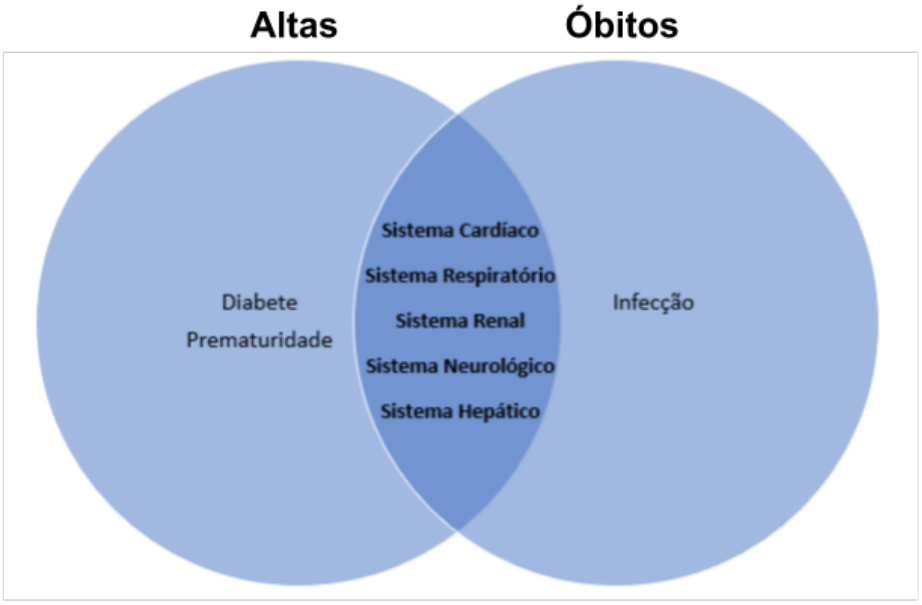


Figura 6.8: Intersecções e disjunções das coleções de documentos.

Através da disjunção é possível avaliarmos a infecção sendo um possível fator contribuinte importante na evolução de óbito de um paciente e que pode ser não interpretado como

a causa do óbito. Por exemplo, o tópico 11 da coleção de óbitos, que foi definido como assunto sistema renal pela interpretação das palavras que compõe o tópico, foi possível inferir que o paciente evidenciava doença renal crônica, fazendo o uso do medicamento lasix do qual é indicado para diminuição de edemas devido a distúrbios renais, promovendo a excreção de urina. Por outro lado, ainda neste tópico a palavra infecção e bacteremia podem evidenciar que a evolução á óbito não necessariamente está associado a doença renal crônica, mas sim, por uma complicação do quadro clínico.

7 CONCLUSÃO

Neste trabalho de conclusão de curso foi realizada uma análise exploratória sobre registros eletrônicos de saúde do setor de unidade de terapia intensiva. Os registros eletrônicos de saúde foram divididos em duas coleções de documentos: internações que obtiveram alta e internações que evoluíram a óbito. Como método de extração de informações foi utilizada modelagem de tópicos, mais especificamente o modelo *LDA*, para extração de tópicos (assuntos) sobre as coleções de documentos.

Para extração dos tópicos, o *LDA* utiliza da co-ocorrência de uma palavra dentro de uma coleção de documentos, permitindo o modelo alocar tópicos através de uma probabilidade utilizando distribuição de *Dirichlet*. A fim de viabilizar este processo, a cada coleção de documentos aplicam-se técnicas de pré-processamento que realizam a limpeza dos dados para posterior a aplicação do *LDA*.

Para definição do melhor modelo *LDA*, ou seja, a definição de hiperparâmetros utilizados na extração de tópicos, foi realizado uma avaliação por meio de métrica de coerência, assim, sendo possível a definição do número 11 como o número de tópicos que melhor descreve os assuntos abordados em cada coleção.

Através da análise das *top-10* palavras do tópico, como resultado, foi possível definir os assuntos abordados em cada coleção. Para a coleção de alta os principais assuntos são sistema respiratório, sistema renal, sistema neurológico, prematuridade e sistema cardíaco, enquanto para coleção de óbitos são sistema hepático, sistema cardiovascular, sistema neurológico e sistema respiratório. Foram analisadas as disjunções e intersecções dos assuntos definidos em cada coleção, e observado a infecção como importante fator contribuinte para evolução a óbito.

7.1 Trabalhos Futuros

Com o intuito de explorar a análise neste trabalho, podem ser realizados trabalhos referentes as internações de alta ou óbitos, compreendendo a evolução dos tópicos em relação ao tempo da internação a fim de inferir uma linha de tempo de tópicos que representam a internação do paciente.

Além disso, seria possível utilizar os documentos pertencentes aos tópicos contidos na intersecção com o objetivo de analisar os assuntos e entender os possíveis fatores que fazem dois documentos com o mesmo assunto pertencer a coleção de óbitos ou altas.

REFERÊNCIAS

BACKES, M. T. S.; ERDMANN, A. L.; BüSCHER, A. O ambiente vivo, dinâmico e complexo de cuidados em Unidade de Terapia Intensiva. **Rev. Latino-Am. Enfermagem**, [S.l.], jun 2015.

BLEI, D. M. Probabilistic topic models. **Communications of the ACM**, [S.l.], v.55, n.4, p.77–84, 2012.

CMF. **Conselho Federal de Medicina, Sociedade Brasileira de Informática em Saúde. Cartilha sobre Prontuário Eletrônico**. Accessed: 2019-10-27, http://portal.cfm.org.br/crmdigital/Cartilha_SBIS_CFM_Prontuario_Eletronico_fev_2012.pdf.

COGEN, R. et al. Redundancy-Aware Topic Modeling for Patient Record Notes. **Department of Biomedical Informatics, Columbia University**, [S.l.], 2013.

CREMESP. **Conselho Regional de Medicina do Estado de São Paulo. Resolução CREMESP nº 71, de 08 de novembro de 1995**. Accessed: 2019-10-27, http://www.medicinaintensiva.com.br/cremesp.htm.

JOHNSON, A. E. et al. MIMIC-III, a freely accessible critical care database. **Scientific Data)**, [S.l.], MAY 2016.

LI, D. C. et al. Discovering Associations Among Diagnosis Groups Using Topic Modeling. **Mayo Clinic, Rochester**, [S.l.], 2013.

MILES, J. Reconstruction of the MIMIC-III Database for Data Analytics. **Oswego State University of New York**, [S.l.], 2017.

NASCIMENTO, H. M. do; ALVES, J. S.; MATTOS, L. A. D. de. Humanização no acolhimento da família dos pacientes internados em Unidade de Terapia Intensiva. **Centro Universitário Católico Salesiano Auxilium (UNISALESIANO)**, [S.l.], p.30–39, JUL 2014.

PANITZ, L. M. Registro eletrônico de saúde e produção de informação da atenção à saúde no SUS. **Escola Nacional de Saúde Pública Sergio Arouca**, [S.l.], p.30–39, JUL 2014.

ROSE, S. Machine Learning for Prediction in Electronic Health Data. **JAMA Network Open**, [S.l.], v.1, n.4, p.e181404–, 2018.

RöDER, M.; BOTH, A.; HINNEBURG, A. Exploring the Space of Topic Coherence Measures. **ACM INTERNATIONAL CONFERENCE ON WEB SEARCH AND DATA MINING**, [S.l.], p.399–408, 2015.

STEYVERS, M.; GRIFFITHS, T. Probabilistic topic models. **Handbook of latent semantic analysis**, [S.l.], v.427, n.7, p.424–440, 2007.

VALENTI, A. P. et al. Using topic modeling to infer the emotional state of people living with Parkinson's disease. **Assistive Technology**, [S.l.], p.1–10, 2019.

WANG, H. et al. Finding Complex Biological Relationships in Recent PubMed Articles Using Bio-LDA. **Public Library of Science**, [S.l.], 2011.

ZHANG, Y.; JIANG, R.; PETZOLD, L. Survival Topic Models for Predicting Outcomes for Trauma Patients. **2017 IEEE 33rd International Conference on Data Engineering (ICDE)**, [S.l.], p.1497–1504, 2017.

APÊNDICES

APÊNDICE A – Tópicos extraídos

Tópico 1	Tópico 2	Tópico 3	Tópico 4
0.017*"skin"	0.031*"valve"	0.020*"liver"	0.009*"arrest"
0.013*"drain"	0.026*"aortica"	0.016*"bleed"	0.008*"transfer"
0.012*"wound"	0.021*"ventricular"	0.015*"renal"	0.007*"daily"
0.011*"draining"	0.017*"mitral"	0.011*"cirrhosi"	0.007*"comfort"
0.011*"fractur"	0.014*"leaflet"	0.009*"hepatic"	0.007*"pulse"
0.009*"open"	0.013*"systole"	0.008*"ascites"	0.006*"arrive"
0.009*"ventilation"	0.012*"wall"	0.008*"sepsis"	0.006*"rhythm"
0.008*"thick"	0.011*"mild"	0.008*"dialysis"	0.006*"unresponse"
0.008*"respiratory"	0.010*"regurgit"	0.008*"lactulos"	0.006*"received"
0.008*"suction"	0.010*"mildline"	0.007*"transplant"	0.006*"known"

Tabela A.1: Tópicos 1-4 extraídos da coleção de óbitos.

Tópico 5	Tópico 6	Tópico 7	Tópico 8
0.014*"respiratory"	0.030*"hemorrhage	0.012*"cmho"	0.012*"renal"
0.011*"ventilation"	0.025*"head"	0.011*"renal"	0.009*"chronic"
0.008*"wean"	0.011*"contrast"	0.011*"ventilation"	0.009*"abdomin"
0.008*"neuro"	0.011*"neuro"	0.010*"intubated"	0.008*"drain"
0.008*"secretion"	0.010*"intubated"	0.009*"sedated"	0.007*"rhythm"
0.007*"intubated"	0.010*"frontal"	0.008*"peep"	0.007*"breath"
0.007*"thick"	0.010*"seizure"	0.007*"shock"	0.007*"heparin"
0.007*"suction"	0.009*"cerebral"	0.006*"balance"	0.007*"afib"
0.007*"shift"	0.009*"subarachnoid"	0.006*"pulse"	0.006*"ventilation"
0.006*"urin"	0.009*"ventricles"	0.006*"breath"	0.006*"neurologic"

Tabela A.2: Tópicos 5-8 extraídos da coleção de óbitos.

Tópico 9	Tópico 10	Tópico 11
0.025*"contrast"	0.014*"pleural"	0.010*"lasix"
0.013*"abdomen"	0.013*"pneumonia"	0.010*"chronic"
0.012*"catheter"	0.010*"unchange"	0.009*"renal"
0.012*"liver"	0.010*"opacities"	0.007*"urin"
0.011*"identifier"	0.010*"interval"	0.007*"hypotension"
0.010*"within"	0.009*"lobe"	0.007*"bacteremia"
0.010*"vein"	0.008*"pneumothorax"	0.007*"pulse"
0.009*"abdomin"	0.007*"upper"	0.007*"infection"
0.009*"pelvis"	0.007*"comparison"	0.007*"transfer"
0.008*"evidence"	0.006*"worsen"	0.006*"skin"

Tabela A.3: Tópicos 9-11 extraídos da coleção de óbitos.

Tópico 1	Tópico 2	Tópico 3	Tópico 4
0.030*"valve"	0.012*"insulin"	0.016*"sepsis"	0.017*"respiratory"
0.024*"aortica"	0.010*"surgeri"	0.013*"baby"	0.011*"ventilation"
0.019*"arteria"	0.009*"wound"	0.013*"murmur"	0.010*"secretions"
0.018*"ventricular"	0.008*"post-op"	0.013*"nicu"	0.009*"wean"
0.016*"mitral"	0.007*"regular"	0.013*"newborn"	0.009*"neuro"
0.013*"leaflet"	0.007*"dilaudid"	0.013*"feed"	0.009*"thick"
0.012*"coronary"	0.007*"extreme"	0.011*"active"	0.008*"skin"
0.011*"cardiac"	0.007*"incision"	0.011*"born"	0.008*"urin"
0.011*"systole"	0.006*"diet"	0.010*"week"	0.008*"intubated"
0.010*"wall"	0.006*"intact"	0.010*"screen"	0.007*"suction"

Tabela A.4: Tópicos 1-4 extraídos da coleção de altas.

Tópico 5	Tópico 6	Tópico 7	Tópico 8
0.009*"acute"	0.022*"head"	0.020*"effusion"	0.017*"liver"
0.008*"urin"	0.022*"fractur"	0.016*"pleural"	0.015*"contrast"
0.008*"pulse"	0.019*"hemorrhage"	0.013*"tube"	0.012*"fluid"
0.007*"bleed"	0.018*"contrast"	0.012*"pulmonary"	0.011*"abdomen"
0.007*"respiratory"	0.010*"neuro"	0.011*"radiolog"	0.011*"abdominal"
0.007*"fluid"	0.010*"arteria"	0.011*"pneumonia"	0.010*"renal"
0.007*"rhythm"	0.010*"radiolog"	0.009*"lobe"	0.010*"vein"
0.007*"renal"	0.008*"evidence"	0.009*"interval"	0.009*"bleed"
0.007*"chronic"	0.008*"mass"	0.009*"lower"	0.009*"hepatic"
0.006*"stool"	0.008*"hematoma"	0.008*"opacities"	0.009*"radiolog"

Tabela A.5: Tópicos 5-8 extraídos da coleção de altas.

Tópico 9	Tópico 10	Tópico 11
0.034*"tube"	0.054*"feed"	0.015*"respiratory"
0.013*"drain"	0.022*"active"	0.014*"tube"
0.011*"draining"	0.022*"stool"	0.011*"acute"
0.011*"pleural"	0.015*"respiratory"	0.010*"ventilation"
0.010*"pneumothorax"	0.014*"murmur"	0.010*"fluid"
0.010*"effusion"	0.011*"retract"	0.009*"intubated"
0.010*"wean"	0.011*"cpap"	0.009*"failure"
0.009*"radiolog"	0.010*"benign"	0.008*"balance"
0.009*"respiratory"	0.010*"neonatolog"	0.008*"breath"
0.007*"neuro"	0.010*"week"	0.007*"nutritional"

Tabela A.6: Tópicos 9-11 extraídos da coleção de altas.