Revista de Sistemas de Informação da FSMA
n. 30 (2022) pp. 15-27

SISTEMAS DE INFORMAÇÃO

http://www.fsma.edu.br/si/sistemas.html

TRILHA PRINCIPAL

Uma Revisão do Conceito de Nuvens de Rótulos

Geraldo Xexéo, *Doutor em Engenharia de Sistemas e Computação, COPPE/UFRJ*,

***Resumo*—Este artigo descreve o trabalho realizado no Laboratório de Informações Não Estruturadas, LINE, da COPPE/UFRJ com nuvens de rótulos. Apresenta um trabalho teórico e prático, incluindo um modelo conceitual de nuvens de rótulos, propostas originais como a nuvem de rótulos diferenciais, e seu processo de criação, além de um módulo Python para criação de vários tipos de nuvens de rótulos e algumas aplicações práticas do conceito pelo grupo.**

***Palavras-chave*—Nuvem de rótulos, Nuvem de Palavras, Lista ponderada.**

A Review of the Tag Clouds Concept

***Abstract*—This article describes the work carried out at the Laboratory of Unstructured Information, LINE, at COPPE/UFRJ concerning tag clouds. It presents theoretical and practical work, including a conceptual model of tag clouds, original proposals such as the differential tag cloud, and its creation process, as well as a Python module for creating various types of tag clouds and practical applications of these concepts.**

***Index Terms*—Tag Clouds, Word Clouds, Weighted List.**

I. Introdução

ESTE artigo apresenta uma revisão do trabalho teórico e aplicado do autor e seus alunos na área de "Nuvem de Rótulos", incluindo uma reescrita da definição formal de tipos de nuvens propostos por Xexeo, Morgado e Fiuza [1] e Morgado [2], as **nuvens de rótulo de sumário** e as **nuvens de rótulo diferenciais**. Este trabalho é realizado dentro do contexto do LINE - Laboratório de Informação Não Estruturada[1] da UFRJ.

Nuvens de rótulos podem ser explicadas como diagramas formados por símbolos, normalmente palavras, que têm como intenção indicar o assunto de um documento ou coleção de documentos. Usualmente, a importância de cada rótulo no documento é indicada por características da fonte adotada por cada palavra no diagrama, principalmente tamanho e cor. A Figura 1 exemplifica o conceito, sendo uma nuvem de rótulos criada a partir dos resumos da 29ª edição da Revista de Sistemas de Informação da da Faculdade Salesiana Maria Auxiliadora usando uma ferramenta disponível na Internet [2].

É difícil precisar o primeiro uso de uma nuvem de rótulos. A Wikipedia cita o site Flickr, e seu co-fundador Stewart Butterfield, como a principal influência em sua divulgação, e sites como Del.icio.us e Technorati como seus popularizadores. Butterfield teria se apoiado em um implementação de Jim Flanagan, hoje apenas disponível na Way Back Machine[3] [3].

Fig. 1: Uma nuvem de palavras gerada a partir de resumos de artigos da 29ª edição da Revista de Sistemas de Informação da da Faculdade Salesiana Maria Auxiliadora. Quanto maior a fonte, isto é, o tamanho das letras, mais frequente é a palavra. Nesse exemplo, todas as outras características são apenas decorativas.

Algumas nuvens apareceram antes na literatura, como no livro "Microserfs" de Douglas Coupland, de 1995, e a capa alemã do livro de Giles Deleuze e Félix Guatari "Tausend Plateaus. Kapitalismus und Schizophrenie", de 1992. A mais antiga referência impressa, porém, é uma página da revista Littérature, de 1923, com o título ERUTARETTIL, que lista a importância de alguns autores surrealistas, criada por André Breton e Robert Desnos [4] e apresentada na Figura 2. Não se tem notícia se houve influência dessas imagens na proposta de Flanagan.

Inicialmente, nuvens de rótulos foram usadas por sites, como o Flickr, Del.icio.us e Technorati, para classificar postagens ou coleções de postagens. Esses rótulos poderiam ser explicitamente dados por seres humanos, o que mais tarde fica conhecido como *folksonomies*, ou obtidos por algoritmos. Mais tarde, dada a suposta facilidade que as nuvens de rótulo parecem dar à análise de significado de um *corpus*, isso é, um documento ou conjunto de documentos determinado, passaram a ser usadas para outras atividades, como a análise política [5], ou a análise da relevância das acomodações no Airbnb [6].

É inegável o sucesso das nuvens de rótulo em algumas aplicações, principalmente quando é necessário ter uma

Autor correspondente: Geraldo Xexéo, *xexeo@ufrj.br*

[1] http://line.cos.ufrj.br

[2] https://www.wordclouds.com/

[3] https://web.archive.org/web/20041204231120/http://twiki.tensegrity.net/bin/view/Main/SearchReferralZeitgeist

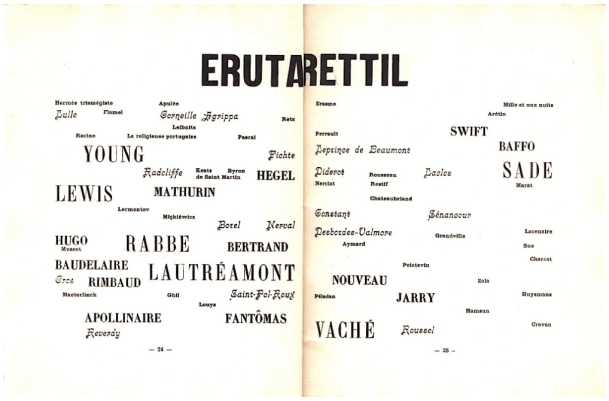


Fig. 2: Primeira imagem impressa conhecida de uma nuvem de palavras. Fonte: Ombre Blanches [4]

compreensão abstrata de uma quantidade grande de texto. Porém existem críticas, e seu uso como ferramenta de navegação em sites praticamente se encerrou. Segundo [7], as nuvens de rótulo quebram quase todas as regras de ouro do design. Se por um lado isso parece surpreendente, também permite que sejam estudadas melhorias, ou quem sabe descobertas novas formas do ser humano perceber a informação, que afetariam o estudo do design. Já Nielsen [8] diz "Nuvens de rótulo são usadas demais. Apesar de parecerem bonitas, elas usam espaço de tela de forma ineficiente e muitos usuários não sabem como usá-las".

As principais ideias tratadas por esse artigo são como entender formalmente a geração de uma nuvem de rótulos de forma automática e para que elas podem ser usadas. Apesar de existirem muitos algoritmos para gerá-las, Xexeo, Morgado e Fiuza [1] propuseram uma modelo conceitual que explica, de forma geral, como são geradas. Esse modelo é apresentado na Seção III. A partir desse modelo esses autores também propuseram duas formas diferentes de criar nuvens de rótulo, as nuvens de rótulos diferenciais, que mostram como um documento se diferencia dentre um conjunto de documentos descrito por uma nuvem de rótulo de sumário. Isso foi explorado, mais tarde, por Morgado [2] em sua dissertação de mestrado.

Para exemplificar esses conceitos, são apresentadas imagens originais de Morgado [2]. A Figura 3 mostra uma nuvem de rótulos de sumário para todos os documentos retornados em uma busca sobre a palavra Jaguar. É possível notar as palavras "cars", "animals" e "games", que se referem a três tipos diferentes de jaguares: o carro, o animal e o console de videogame. Apesar de existirem formas específicas para gerar nuvens de rótulo de resumo, elas podem, na prática, serem vistas como nuvens de rótulos de vários documentos.

Já uma nuvem de rótulo diferencial exige não só um documento, mas uma coleção de referência, pois ela mostra como um documento (ou documentos) se diferem de uma coleção. A Figura 4 mostra como um documento específico que trata o animal jaguar se diferencia da coleção completa (e sua nuvem de rótulos).

Fig. 3: Nuvem de rótulos sumário para o conjunto de documentos gerados por uma busca sobre a palavra Jaguar. Fonte:Morgado [2]

Fig. 4: Nuvem de rótulos diferencial para um documento sobre o animal jaguar, em referência ao conjunto de documentos gerado por uma busca pela palavra jaguar. Fonte:Morgado [2]

A. *Um pequena questão de nomenclatura*

O termo "nuvem de rótulos", adotado neste texto, resume uma variedade de termos usados na literatura e no dia a dia. Os mais comuns, além do que adotamos, são "nuvens de palavras", ou as versões em inglês, "*word clouds*" ou "*tag clouds*". Alguns autores, buscando uma forma supostamente mais científica, usam o nome "*weighted list*", ou "lista ponderada", porém esse nome se confunde com um conceito pouco conhecido de estrutura de dados. Outro nome que vem sendo usado é "*wordle*", que seriam uma versão das *tag clouds* que "dão atenção cuidadosa a tipografia, cor e composição" [9], mas que na prática resultam na mesma coisa.

Existe uma leve diferença semântica entre "palavra" e "rótulo". Idealmente, quando é usado o termo "palavra" se supõe que as palavras que aparecem na nuvem foram tiradas diretamente do texto. Já no caso do termo "rótulo", se supõe que as palavras que aparecem na nuvem foram obtidas de certa forma a partir do texto, mas não necessariamente pertencem ao texto, podendo apenas ser rótulos calculados de alguma forma ou fornecidos por seres humanos. Nielsen [8] explicita essa diferença, dizendo que nuvens de palavras falam dos termos mais frequentes encontrados em um *corpus*. Em seu exemplo ele mostra como a comparação visual de duas nuvens de palavras pode levar a compreensão das diferenças e semelhanças entre dois documentos, mesmo que afirme que um resumo de um parágrafo seria mais útil.

Essa diferenciação, porém, é indiferente perante a teoria desenvolvida neste artigo. Assim, para normalizar a nomenclatura, será usado sempre o termo "nuvem de rótulos", já que rótulos podem ser palavras, mas nem todo rótulo é necessariamente uma palavra. Sempre que possível, os termos em inglês serão traduzidos.

B. Apresentação do artigo

Esse artigo se inicia com essa introdução, que explica informalmente o conceito de nuvens de rótulos, como apareceram e discute seus vários nomes. Na seção II é dada uma visão geral das nuvens de rótulo, discutindo seu conceito e formas de apresentação. Na seção III é apresentado um modelo formal para criação de nuvens de rótulos. A seguir, a seção IV discute algumas técnicas para criação de nuvens de rótulos, dentro do quadro deste modelo. Algumas aplicações são descritas na seção V, seguida de uma introdução a biblioteca GTAGLIB na seção VI. Finalmente as conclusões são apresentadas na seção VII

II. Visão Geral das Nuvens de Rótulos

SEGUNDO Rivadeneira, Gruen, Muller et al. [10], "Nuvens de rótulos são representações visuais de um conjunto de palavras, tipicamente um conjunto de rótulos selecionados por algum método racional, na qual atributos do texto como tamanho, estilo, ou cor são usados para representar características dos termos associados".

Apesar de seres humanos poderem criar nuvens de palavras, o método racional normalmente é um algoritmo, composto de vários passos típicos do processamento de linguagem natural: obtenção do arquivo original, extração do texto, limpeza de caracteres indesejados e outras formas de pré-processamento, extração das palavras mais significativas ou de conceitos, e a organização das palavras na forma de um diagrama.

Uma característica adicional importante das nuvens de palavras é que elas se referem a alguma coisa, um objeto, normalmente um documento ou uma coleção de documentos. Essa relação pode ser feita de forma explícita, por exemplo por meio de títulos, legendas ou URLs, ou implícita, por exemplo por meio da localização da nuvem de palavras dentro do próprio documento. Ou seja, **nuvens de rótulos são descrições de objetos**.

Rótulos são normalmente palavras, mas podem ser qualquer símbolo associado a um objeto, como uma imagem, um ícone, uma sequência de palavras (conhecidas como n-gramas), radicais, ou lemas.

O objetivo de cada rótulo é representar um conceito, que é ligado, de alguma forma, ao objeto sendo descrito pela nuvem. Marinchev [11] propõe que para cada objeto existe um conjunto de conceitos abstratos associados a ele. Um campo semântico é "o conjunto de conceitos conectados a um objeto foco, de tal forma que é independente das pessoas que atribuíram os rótulos (rotuladores originais) e é possível que outras pessoas tenham o entendimento destes conceitos" [11].

Uma nuvem de rótulos pode então ser interpretada como uma representação visual de um campo semântico de um objeto. Assim, a criação de uma nuvem de rótulos pode ser entendida como um processo de três passos [1], [11]:

1) Entendimento do objeto foco e dos conceitos a ele aplicáveis;
2) Compreensão do campo semântico associado ao objeto, e
3) Transcrição do campo semântico em uma nuvem de rótulos.

É importante notar que há uma dificuldade semântica entre cada passo, o que permite a introdução de ruídos. Pode ser que no primeiro passo o entendimento não seja completo ou correto, pode haver um problema na tradução desses conceitos em uma visão completa e finalmente pode haver também uma dificuldade de transformar os conceitos em rótulos. Por exemplo, o conceito de "saudade" não possui uma tradução em uma palavra apenas em outras línguas, enquanto os alemães possuem uma palavra, "schadenfreude", para o conceito de sentir prazer com a desgraça alheia. Portanto, em determinadas línguas, pode não ser fácil encontrar um rótulo perfeitamente adequado em uma situação.

O processo de interpretação de uma nuvem de palavras também é sujeito a ruídos. Ele é composto de [11]:

1) Leitura dos rótulos da nuvem, dentro de um contexto;
2) Interpretação dos rótulos em conceitos;
3) (Re)Criação do campo semântico, e
4) Criação de uma imagem mental do que é o objeto foco associada ao campo semântico.

Toda esse processo é mediado por propriedades visuais. Essas propriedades influenciam a percepção do usuário sobre o objeto associado. Bateman, Gutwin e Nacenta [12] discutem as características visuais das nuvens de rótulos e como elas influenciam a atenção do usuário, informações que podem ser usadas pelo criador da nuvem para direcionar a atenção do usuário e, consequentemente, sua interpretação do documento. Segundo Bateman, Gutwin e Nacenta [12], as propriedades visuais das fontes mais importantes são: tamanho, peso e intensidade das cores. As cores parecem ser importantes como indicadores, apesar desses autores não terem conseguido descobrir quais cores atraem mais atenção. Além disso, a posição dos rótulos é importante na percepção, devendo os mais importantes estar no centro da nuvem.

Várias táticas de posicionamento também podem ser usadas para manipular a atenção, como a proximidade semântica dos rótulos [13], a distância do centro em nuvens circulares [14], a movimentação dos rótulos e o uso de 3D.

Tudo isso leva a compreensão de que a interpretação de uma nuvem de rótulos está sujeita a vários ruídos, tanto na sua criação, quanto na sua interpretação. Esses ruídos podem criar, alterar, ou eliminar símbolos na nuvem, o que por sua vez pode criar, alterar ou eliminar conceitos do campo semântico formado na mente do leitor. Atualmente esse processo é normalmente automatizado, por meio de

algoritmos de processamento de linguagem natural (NLP), aprendizado de máquina, ou outras estratégias e heurísticas.

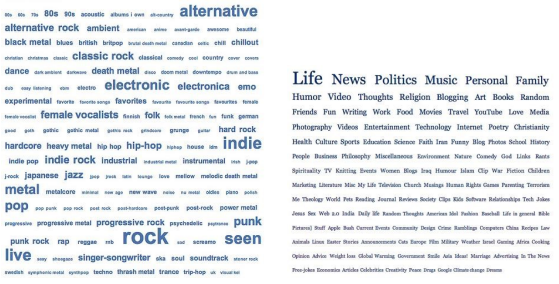
Neste artigo, mostramos um modelo conceitual geral que atende a qualquer forma de criação das nuvens de rótulos, o que permite entender os passos e estabelecer formas de avaliar o resultado.

A. *Características das Nuvens de Rótulos*

Como visto pelo exemplo da Figura 1, nuvens de rótulos são construídas pela disposição de símbolos, normalmente palavras, em um espaço, normalmente bi-dimensional. Esses rótulos podem ser colocadas de várias maneiras, o que é feito usualmente por motivos estéticos. No exemplo da Figura 1 a imagem de uma nuvem é usada para limitar o espaço, enquanto as palavras, que representam os rótulos, são colocadas em várias posições e rotações.

A praxe é que o tamanho do rótulo indique a importância do mesmo como representante do significado do objeto para o qual foi criada a nuvem. Outros atributos, como cor, centralidade da posição, etc. podem ser usados tanto com significado como por motivos decorativos. Na Figura 1, cor, tipo e peso da fonte e rotação são decorativos. O algoritmo usado também favorece colocar próximo ao centro as palavras mais importantes.

Outras nuvens de rótulos seguem estratégias diferentes. Uma abordagem é listar as palavras sequencialmente, sem variar muitos elementos das fontes, de modo a oferecer menos distratores. A Figura 5a mostra uma nuvem ordenada em ordem alfabética, enquanto a Figura 5b mostra uma nuvem ordenada por importância do rótulo.



(a) Ordem alfabética. (b) Ordem de importância.

Fig. 5: Duas maneiras de construir uma nuvem de rótulos.

Outra forma de organizar é por semelhança conceitual entre os rótulos, como a nuvem da Figura 6. Para isso, no caso da criação automática, é necessário um processamento mais sofisticado, como o uso de *Latent Dirichlet Analysis* (LDA) [2].

III. Um Modelo Formal para Nuvens de Rótulo

UM modelo formal para nuvens de rótulo foi proposto por Xexeo, Morgado e Fiuza [1] e descrito de detalhadamente na dissertação de mestrado de Morgado [2]. Nesta seção ele é reapresentado buscando uma melhor ordem para a compreensão de conceitos.

Fig. 6: Nuvem ordenada por importância do rótulo.

O modelo começa com definições muito gerais, que servirão de base para a construção do resto do modelo, partindo do conceito primitivo de objetos, aos quais podem ser associados atributos. Cada atributo descreve uma propriedade do objeto, de acordo com um conjunto de valores. O conjunto de atributos define, implicitamente, a classe do objeto. Mapas permitem associar objetos a atributos, atributos a conjunto de valores e, um valor a cada atributo de um objeto. Esse modelo é descrito em UML na Figura 7.

A Subseção III-A descreve esse modelo, por meio de conjuntos e funções, introduzindo restrições, podendo ser considerada um meta-modelo de referência. A seguir, na Subseção III-B são introduzidos os conceitos de recursos, o que queremos representar nas nuvens de rótulos, e suas coleções, contextos, e as funções necessárias para trabalhar com eles. O modelo, porém, se baseia na ideia que essa representação existe por causa da existência de campos semânticos, e isso é tratado na Subseção III-C. Estas definições são então finalmente usadas para gerar rótulos (Subseção III-D), e criar nuvens de rótulos (Subseção III-E).

Esse modelo tenta deixar claro que **nuvens de rótulos são representações físicas de campos semânticos, onde valores associados aos conceitos, como relevância, são traduzidos para valores associados aos rótulos, como peso da fonte ou posição no diagrama.**

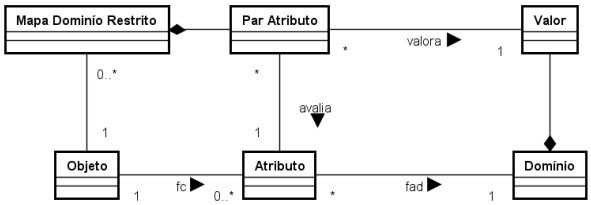


Fig. 7: Modelo UML que descreve como são construídos os atributos de um objeto

A. *Definições iniciais do modelo*

Um **objeto**, representado por o, é um conceito primitivo, semelhante a definição de elemento na Teoria Ingênua de Conjuntos [15]. O conjunto de objetos é o conjunto universo U. Na Computação, objetos são geralmente a classe raiz da qual derivam todas as outras, tanto em ontologias quanto em linguagens de programação.

Como no modelo relacional [16], um **domínio** é um conjunto de valores, denotado como V_i. O conjunto de todos os domínios é indicado pela letra V. Valores de um domínio

podem ser indicados pela letra v com índices, como em v_{ij}, que indica o j-ésimo valor do i-ésimo domínio.

Um **atributo**, representado por a, de um objeto o, é uma propriedade que descreve o. O conjunto de todos atributos possíveis é denotado por A. Um conjunto de atributos, isto é, um subconjunto de A, é representado por A_i. Um atributo de A_i é representado por a_i, e o atributo a_i para um objeto o_j é representado por a_{ij}.

Para associar um domínio a um atributo, é definida uma **função de atribuição de domínio**, representada por f_{ad}. Atributos só podem possuir um domínio, mas vários atributos podem possuir o mesmo domínio.

$$f_{ad} : A \rightarrow V \tag{1}$$

Um **par atributo**, é um par ordenado (a_i, v_{ij}) que indica um atributo a seu valor.

$$(a_i, v_{ij}) \in A \times V_i, \text{ onde } f_{ad}(a_i) = v_{ij} \tag{2}$$

Para que um objeto ser descrito, é preciso que ele possua atributos e que os atributos possuam valor. Para isso é necessário existir um **conjunto paratributo**, ou um **mapa de domínio restrito**, representado por m, que é um conjunto de pares atributos onde todo primeiro elemento de um par ordenado é único entre todos os componentes do mapa.

$$\begin{aligned} m = \{&(a_i, v_{ij})| \\ &(((a_i, v_{ij}) \in A \times V_i) \wedge (f_{ad}(a_i) = V_i)) \wedge \\ &((\text{ se } (a_i, v_{ij} \in m) \wedge (a_i, v_{ik} \in m)) \\ &\text{ então } (a_i = a_k) \Rightarrow (v_{ij} = v_{ik}))\} \end{aligned} \tag{3}$$

O conjunto de todos os mapas possíveis é denotado por M.

Uma **função de classificação**, representada por f_c, gera um conjunto de atributos $A_i, A_i \subset A$, a partir de um objeto o. A ideia aqui é que essa função, reconhecendo um objeto, gera todos os atributos que o descrevem, mas não seus valores. Ela representa a abstração de classificação dos seres humanos[4]. A notação $\wp(A)$ significa o conjunto potência de A, indicando que um objeto pode ter quaisquer quantidade de atributos

$$f_c : U \rightarrow \wp(A) \tag{4}$$

$$f_c(o) = \{a_{ij} | a_{ij} \text{ é uma propriedade de } o\} \tag{5}$$

Uma função de atribuição de mapa é uma função f_{am} que, dado um objeto, e um conjunto de atributos $A_i, A_i \subset A$, gera um mapa no qual para cada atributo de A_i há um par atributo correspondente. Esse mapa é, na prática, a atribuição de valores a cada atributo de cada objeto.

$$f_{am} : U \times \wp(A) \rightarrow M \tag{6}$$

$$\begin{aligned} f_{am}(o, A_i) = \{&(a_i, v_{ij}) | \forall a_{ij} \in A_i \Rightarrow \\ &(f_{ad}(a_{ij} = V_i) \wedge ()(a_i, v_{ij}) \in f_{am}(o, A_i)))\} \end{aligned} \tag{7}$$

Essas duas funções permitem classificar um objeto, isto é, determinar seus atributos, e dar valores aos mesmos, ambos de forma dinâmica.

[4]É possível observar que apesar do modelo falar de atributos, nada no modelo evita que o atributo seja uma função, já que funções também podem ser valores, e na prática ser equivalente a um método na orientação a objeto

B. Introduzindo recursos e contextos

Como nuvens de rótulos são criadas na Internet, o modelo formal restringe os objetos a recursos. A definição de **recurso** é tomada de do RFC 3986 [17]: qualquer conceito abstrato ou entidade física que pode ser identificado unicamente. Essa restrição é quase que nula, porém conceitualmente existem, ou podem existir, entidades ou conceitos que não podem ser identificados unicamente. Nesse modelo formal, um **recurso**, representado por r, é também qualquer objeto que possa ser descrito, ao menos parcialmente, por rótulos [2]. Um **contexto**, representado por w, é uma coleção de recursos. O conjunto de todos os contextos tem a notação W, inspirado na *World Wide Web*.

Da mesma forma como foi feito nas nuvens de rótulos, deseja-se representar recursos, e o modelo basicamente repete as definições do meta-modelo, mas restringindo os objetos classificáveis ao conjunto de recursos. Assim, ele se inicia por definir uma *função de classificação de recursos*, f_{cr}, e depois uma função de atribuição de mapa de recurso, f_{amr}.

$$f_{cr} : W \rightarrow \wp(A) \tag{8}$$

$$f_{cr}(r) = \{a_{ij} | a_{ij} \text{ é uma propriedade de } r\} \tag{9}$$

$$f_{amr} : W \times \wp(A) \rightarrow M \tag{10}$$

$$\begin{aligned} f_{amr}(r, A_i) = \{&(a_i, v_{ij}) | \forall a_{ij} \in A_i \Rightarrow \\ &((a_i, v_{ij}) \in f_{amr}(o, A_i) \wedge f_{ad}(a_{ij} = V_i))\} \end{aligned} \tag{11}$$

Tendo essas definições, é possível agora buscar **representações de recursos**. A ideia básica é que um recurso possui atributos valorados, e pode ser representado de alguma forma, a partir desses atributos e valores. Assim **representação de recursos**, ou mapa de recurso, representado como MR, é uma função de atribuição de mapa que tem como parâmetros um recurso e uma função de classificação de recursos, isto é, dado um recurso e seus atributos, retorna um mapa com os valores do recurso. Esse mapa é uma representação abstrata do recurso.

$$MR(r) = f_{amr}(r, f_{cr}(r)) \tag{12}$$

C. Criando campos semânticos

Um campo semântico, representado por CS, seguindo a definição de Marinchev [11], é um subconjunto do conjunto de todos os conceitos abstratos, representado pelo conjunto C. Supondo um predicado lógico, ou seja, uma função que responde verdadeiro ou falso, $D(c, r)$ que indica se um conceito descreve de alguma forma um recurso, um campo semântico para um recurso r pode ser definido como:

$$CS(r) = \{c_i | c_i \in C \wedge D(c_i, r)\} \tag{13}$$

Cada conceito, para um recurso específico, permite que seja descrito por um conjunto de atributos. Agora é possível ver como o meta-modelo pode ser usado para descrever tanto recursos como conceitos, pois ambos são objetos. Um

função de classificação de conceito, representada por f_{cc}, é definida por:

$$f_{cc} : W \times C \to \wp(A) \tag{14}$$
$$f_{cc}(r, c) = \{\text{atributos de } c \text{ quando descreve } r\} \tag{15}$$

E, para uma função de classificação de conceito, existe uma **função de atribuição de mapa para um conceito**:

$$f_{amc} : W \times C \times \wp(A) \to M \tag{16}$$
$$\begin{aligned} f_{amc}(r, c, A_i) = \{(a_{ij}, v_{ij}) \mid \forall a_{ij} \in A_i \Rightarrow \\ ((a_{ij}, v_{ij}) \in f_{amc}(r, c, A_i) \wedge \\ f_{ad}(a_{ij}) = V_i))\} \end{aligned} \tag{17}$$

Com isso é possível criar um **campo semântico avaliado** para um recurso r, que é um conjunto de pares ordenados no qual o primeiro elemento é um conceito c_i aplicável ao recurso r e o segundo elemento é um conjunto par atributo composto dos atributos induzidos por c_i em r.

$$\begin{aligned} CSA(r) = \{(c_i, m_i) | c_i \in CS(r) \wedge \\ m_i = f_{amc}(r, c_i, f_{cc}(r, c_i))\} \end{aligned} \tag{18}$$

O campos semânticos e os campos semânticos avaliados, que devem ser gerados, respectivamente, por uma **função geradora de campos semânticos**, f_{gcs} e por uma **função geradora de campos semânticos avaliados**, f_{gcsa}.

$$f_{gcs} : W \times \wp(W) \to C \tag{19}$$
$$f_{gcs}(r, w) = \{c_i \mid c_i \in C \wedge \text{representa } w(c_i, r)\} \tag{20}$$
$$f_{gcs}(r, w) = CS_w(r) \tag{21}$$

$$f_{gcsa} : W \times \wp(W) \to C \times M \tag{22}$$
$$\begin{aligned} f_{gcsa}(r; w) = \{(c_i, m_i) \mid c_i \in CS_w(r) \\ \wedge m_i = f_{amc}(r, c_i f_{cc}(r, c_i))\} \end{aligned} \tag{23}$$

Neste ponto é importante entender o que é possível: não só associar conceitos aos recursos, mas também associar atributos e valores aos conceitos, e também aos recursos. O modelo da Figura 8 mostra essas relações.

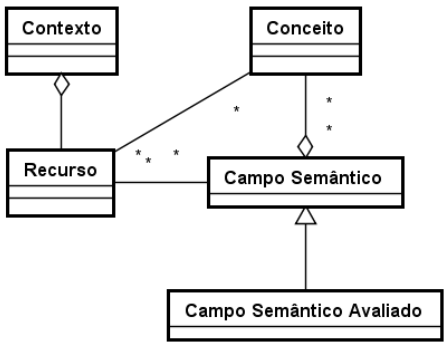


Fig. 8: Campos semânticos e recursos em UML.

D. Gerando Rótulos

Os conceitos de um campo semântico devem ser transformados em rótulos a serem usados nas nuvens, com vários atributos. Para isso são necessárias mais algumas definições. Rótulos são normalmente palavras únicas ou n-gramas, mas podem na verdade ser qualquer símbolo.

Uma **função de classificação de rótulo** é uma função f_{ct}, que, dado um recurso r e um rótulo t (do inglês *tag*) gera um conjunto de atributos $A_i, A_i \subset A$, que pode ser utilizado para representar os atributos da rótulo t quando se referem ao recurso r. O conjunto de todos os rótulos é descrito pela letra T.

$$f_{ct} : W \times T \to \wp(A) \tag{24}$$
$$= \{a_{ij} | a_{ij} \text{ é um atributo de } t \text{ quando se refere a } r\} \tag{25}$$

Uma **função de atribuição de mapa para uma rótulo** t, é uma função f_{amt} que, dados um rótulo t, um recurso r, e um conjunto de atributos $A_i, A_i \subset A$, gera um mapa no qual para cada atributo de A_i há um par atributo correspondente o descrevendo.

$$f_{amt} : W \times T \times \wp(A) \to M \tag{26}$$
$$\begin{aligned} f_{amt}(r, t, A_i) = \{(a_{ij}, v_{ij}) \mid \forall a_{ij} \in A_i \Rightarrow \\ ((a_{ij}, v_{ij}) \in f_{amt}(r, t, A_i) \wedge \\ (f_{ad}(a_{ij}) = V_i))\} \end{aligned} \tag{27}$$

Uma **representação de rótulo** ou um **mapa de rótulo** para uma rótulo t, $MT(t)$, é um mapa:

$$MT(t) = f_{amt}(r, t, f_{ct}(r, t)) \tag{28}$$

Mapas de rótulo agem como as representações dos rótulos. Por exemplo, o mapa de uma rótulo pode ser formado por tuplas que representam suas propriedades na nuvem de rótulos.

Um **gerador de conjunto de rótulos** é uma função f_{gct} que dado um contexto w, um recurso específico $r, r \subset w$, gera um conjunto de rótulos $CT(r)$ que representa o grupo de rótulos que podem ser consideradas, sob alguma razão, serem aplicadas a r no contexto w.

$$f_{gct} : W \times \wp(W) \to T \tag{29}$$
$$\begin{aligned} f_{gct}(r, w) = \{t_i | \exists c_i \in CS_w(r) \wedge \\ \text{representa}_w(t_j, c_j)\} = CT_w(r) \end{aligned} \tag{30}$$

Dado um conjunto de rótulos $TC(r)$, uma **nuvem de rótulos abstrata** $TCA(r)$ é um conjunto de tuplas

$$TCA(r) = \{(t_j, m_i)\}, \tag{31}$$

E. Criando Nuvens de Rótulos

A partir das definições do modelo, é possível entender que para criar uma nuvem de rótulo é necessário seguir o seguinte processo:

1) Escolher um recurso, normalmente um documento, dentro de um contexto;

2) Criar um campo semântico avaliado para o objeto, utilizando para isso uma função gerado de campos semânticos avaliados ;
3) Utilizar o campo semântico avaliado para definir uma nuvem de rótulos abstrata para o objeto, por meio de um gerador de conjunto de rótulos, e
4) Gerar uma representação visual para a nuvem de rótulos abstrata, a partir de uma mapeamento de atributos a propriedades dos rótulos.

F. Um modelo para nuvens de rótulo de sumário

Uma **nuvem de rótulos de sumário** é uma nuvem que representa todos os recursos existentes de um determinado contexto. Ela pouco se difere de uma nuvem de rótulos de um conjunto maior, mas tem seu papel teórico para a definição de nuvens de rótulo diferenciais, já que é preciso que essa nuvem maior seja criada de alguma forma que dê sentido as nuvens diferenciais.

Uma nuvem de rótulo de união (UTC) é na prática uma união de conjuntos:

$$UTC(w) = \bigcup_{r \in w} TCA(r) \tag{32}$$

Essa nuvem então representa todos os conceitos possíveis de todos os recursos, mas não o resumo da coleção de recursos como um todo. Para isso, é necessário criar uma **função de sumarização** é uma função $f_s(t)$, que dado um rótulo t sabe decidir se ele deve ou não aparecer na nuvem de rótulo de sumário.

$$f_s(t) \to \mathbb{B}, \tag{33}$$

onde $\mathbb{B}$ é o conjunto dos booleanos $\{F, V\}$.

Sendo assim, uma **nuvem de rótulo de sumário** é definida como:

$$STC = \{t | (t \in UTC(w)) \wedge (f_s(t) = V)\} \tag{34}$$

Portanto, resumidamente esta nuvem de rótulos representa o conjunto de rótulos que estão relacionados a pelo menos a algum dos recursos do contexto e, ao mesmo tempo, atendem às condições especificadas por $f_s(t)$. Geralmente, esta função utilizará algum valor limite como filtro para selecionar somente os rótulos que pertencem a nuvem de rótulos de sumário.

G. Um modelo para nuvens de rótulo diferenciais

Finalmente é possível definir formalmente o conceito de **nuvem de rótulos diferencial**. Seguindo o esquema das subseções anteriores, pretende-se fornecer um modelo que possa ser utilizado por um processo para gerá-las.

O interesse é representar as diferenças de um recurso específico em relação aos outros integrantes da coleção.

O conceito de diferença é usado no sentido de realçar, destacar ou evidenciar as características ou conceitos particulares de um recurso que não estejam sendo abordados pelos demais recursos neste mesmo contexto. Portanto, uma $DTC(r, w)$ é definida como uma nuvem de rótulos, que fornece uma representação visual dos conceitos presentes em um específico recurso r, e que de alguma forma, se destacam ou se diferenciam dos demais conceitos presentes nos outros recursos, que são componentes de um determinado contexto w.

Para começar, é apresentada a função de diferenciação que é uma função $f_d(t, r, w)$, que irá fornecer uma condição para definir se o rótulo t deverá ser utilizado para representar um conceito que existe no recurso r que está sendo analisado mas não aparece na STC.

$$f_d : T \times W \times C \to \mathbb{B} \tag{35}$$

O modelo aqui apresentado cria um arcabouço de conjuntos e funções que guiam, como um template, as técnicas de construção de nuvens de rótulos, descritas na próxima seção.

Desta forma, é possível definir uma **nuvem de rótulo diferencial**, como:

$$DTC(r, w) = \{t | (t \in CRC(w) \vee t \in TCA(r)) \wedge f_d(t, r, w) = V\} \tag{36}$$

Logo, na $DTC(r, w)$ são encontrados rótulos que descrevem o documento individualmente, mas não são necessariamente consideradas na nuvem de rótulos de sumário de resumo do conjunto. Sendo assim, a **principal função de uma nuvem de rótulos diferencial é representar características individuais e importantes de cada recurso** e por consequência auxiliar aos usuários no momento em que estes realizam uma busca, em certo contexto, a encontrar os recursos que eles desejam.

IV. Principais Técnicas Utilizadas

As principais técnicas utilizadas na criação de nuvens de rótulos são típicas das áreas de Busca e Recuperação da Informação [18], Mineração de Texto, e Processamento de Linguagem Natural [19]. Um processo para gerar uma nuvem de rótulos pode ser dividido em 7 passos:

1) Recuperação do documento, ou coleção de documentos, que pode ser tanto apenas a leitura de um arquivo conhecido como a recuperação automática de uma série de documentos na web por meio de *crawlers* [20];
2) Leitura do arquivo e extração do texto, muitas vezes exigindo o uso de bibliotecas especializadas, como o pacote `pdfminer.six` [21], no caso de arquivos PDF;
3) Pré-processamento do texto [19], que exige passos como:
   a) eliminação de símbolos indesejados no processamento, como TABs, CRs e LFs, acentos e variações aplicadas às letras, como diacríticos, ou criação de representações alternativas, de modo a corrigir possíveis erros;
   b) eliminação ou tradução de códigos que representam imagens, que não pertencem a língua sendo processada, como emoticons, ou que, em

sequência, possuem um significado específico, como ":-)" para um rosto sorrindo;

c) eliminação de comandos ou marcações que não pertencem ao texto, como marcações HTML ou XML, normalmente usando pacotes como *Beautiful Soup* [22];

d) transformação para minúsculas ou maísculas, conhecida como *case-folding*, se e como desejado;

e) tokenização, i.e., separação do documento em tokens, ou palavras, e possível geração de n-gramas [23];

f) correção de erros de digitação ou de ortografia [24], [25];

g) tratamento linguístico, em busca da descoberta da classe gramatical, função sintática, significados, etc.

h) eliminação de palavras indesejadas na nuvem, tanto *stopwords*, aquelas que não trazem isoladamente diferenciação do significado do texto, quanto palavras que aparecem sempre em uma coleção [18], [23], como o nome do site onde as palavras foram obtidas;

i) eliminação das variações de gênero, número, tempo, grau, etc, e ainda, em alguns algumas vezes, a lematização, i.e., buscar a forma padrão da língua, possivelmente por meio de tabelas, ou a radicalização (*stemming*), i.e., remoção de sufixos e prefixos Baeza-Yates e Ribeiro-Neto [18], Jurafsky e Martin [19], Manning, Raghavan e Schütze [23] e Porter [26].

4) Geração de estruturas de dados intermediárias, na forma de índices, ou estruturas chave-valor, como listas invertidas, matrizes termo-documento ou ainda baseados na semântica latente, modelos de linguagem, etc; incluindo preocupações com n-gramas [18], [23];
5) Extração da informação desejada, na forma de uma lista ordenada de rótulos e propriedades, geralmente apenas uma lista palavra-peso [2];
6) Mapeamento das propriedades associadas aos rótulos um forma de representação dos mesmos, geralmente apenas o tamanho da fonte, e
7) Geração final do diagrama, com posicionamento, embelezamento estético, etc. [2].

Apesar do pré-processamento de texto ser composto de tarefas simples, ou pelo menos amplamente conhecidas, ele é crítico na qualidade das nuvens de rótulo, principalmente quando é necessário levar mais em conta o conceito que as palavras representam do que propriamente as palavras. A conflação de palavras, por meio da eliminação de suas variações, como por exemplo, a conflação pela variação em número pode garantir que a forma plural e singular sejam representadas de maneira unificada na nuvem, juntando palavras como "nuvem" e "nuvens". Por outro lado, pode não ser interessante fazer a conflação em gênero, para poder separar palavras como "cidadão" e "cidadã" em um estudo de gênero. A eliminação de *stopwords*, que hoje é pouco preocupante em outras técnicas de Processamento de Linguagem Natural, também é tipicamente essencial, para evitar que a nuvem fique repleta de palavras como "que" e "mas". Por outro lado ela deve também ser feita com cuidado, já que palavras que constam de listas de *stopwords* podem ser importantes na compreensão do texto, como o uso de pronomes pessoais e possessivos.

Muito do que é feito no passo 4 parte de técnicas básicas, como a simples contagem de frequência, no caso de um único documento. Técnicas mais avançadas buscam conceitos em vez das palavras retiradas diretamente do texto, como a Análise por Semântica Latente (LSA) [27], ou a Análise Latente de Dirichlet (LDA) [28], usadas no passo 4, são na prática técnicas de redução de dimensionalidade, que tentam fazer com que palavras encontradas no texto se juntem em grupos, permitindo usar dessas palavras como rótulos. Essa unificação de palavras em um conceito também pode ser conseguido por meio de tesauros, como o *WordNet* [29].

Um *thesaurus* é uma lista de palavras com significados semelhantes, sendo que a semelhança pode ocorrer por sinonímia, hiperonímia, ou outras formas. O *WordNet* foi criado na Universidade de Princeton e se tornou um padrão *de facto* para a língua inglesa, e uma referência de formato adotada por várias outras iniciativas em outras línguas. No WordNet, as palavras estão reunidas em conjuntos de um mesmo significado conhecidos como *synset*. Iniciativas semelhantes em português incluem o OpenWordNet-PT [30].

Recentemente, há um interesse em "semantic word clouds", que buscam fazer uma representação que aproxima mais os rótulos que têm significado similar [31], [32], como mostrado na Figura 6. Esses modelos usam as técnicas citadas nos parágrafos anteriores ou similares.

Morgado [2] mostra uma instância do processo de criação de nuvens de rótulo por meio de um diagrama de atividades UML, adaptado na Figura 9. Além disso, Morgado [2] propõe algoritmos para duas abordagens para criação de nuvens de rótulos abstratas.

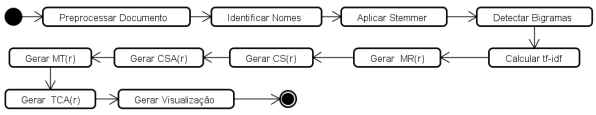


Fig. 9: Uma instância de um processo de geração de nuvens de rótulo. Adaptado de: Morgado [2].

A medida de importância de um termo em um texto usa normalmente uma variação do **tf-idf**, acrônimo em inglês para indicar que estão sendo consideradas simultaneamente a frequência do termo no documento e frequência inversa do termo em todos os documentos. Essa é uma medida usada na busca e recuperação da informação que indica a importância relativa de um termo em um documento considerando como ele é usado em toda coleção de documentos, onde termos frequentes em todos os documentos são considerados menos importantes. Ape-

sar de existirem muitas fórmulas, que produzem valores diferentes, para o conceito, uma das mais citadas é [18]:

$$TF - IDF = f_i \log \frac{N}{n_i}, \tag{37}$$

onde f_i é a frequência do termo i no documento em questão, N o número de documentos na coleção e n_i o número de documentos da coleção onde o termo i aparece.

Para implementação dessas técnicas, existem pacotes bastante completos oferendo uma grande variedade de técnicas de Processamento de Linguagem Natural, como o NLTK [33] para Python e Weka para Java [34]. Existem também pacotes específicos para a geração de nuvens de rótulos, como o `wordclouds` [35].

V. Aplicações

ESTA seção exemplifica o uso de nuvens de rótulos. É importante lembrar que o objetivo geral de uma nuvem de rótulos é representar, de uma forma facilmente compreendida, um objeto, normalmente um documento, ou um conjunto de objetos. Outro uso comum é na diferenciação entre duas nuvens, sendo nesse caso usadas as nuvem diferenciais e de resumo.

Dessa forma, havendo duas coleções de documentos, é possível compará-los das seguintes formas: pela comparação visual de duas nuvens de rótulos, uma para cada coleção, ou pela comparação de três nuvens de rótulos: uma que mostra o que todos os documentos tem em comum, a nuvem de rótulo sumário, e uma para cada coleção de documentos, que mostra como esse subconjunto se diferencia do todo.

Paiva e Pinho [6] apresentam uma análise visual, na forma de nuvens de rótulos, de projetos de lei da Câmara de Deputados, sendo que essa análise visual é criada por meio de um algoritmo de LSA [27]. Essa é uma forma sofisticada de fazer a visualização por nuvens de rótulos, pois busca mostrar não as palavras encontradas no texto por sua frequência, mas sim por seu significado. Nesse trabalho, foram analisados apenas Projetos de Lei propostos por deputados, enre 1934 e 2022, do site Dados Abertos da Câmara[5]. Para exemplificar, mostramos ns Figuras 10 e 11 como comparar os tópicos mais relevantes nos períodos de 2000 a 2019 e ao período da pandemia, 2020 a 2022. É possível notar o aparecimento da palavra pandemia, a diminuição da relevância do termo trânsito, e o aparecimento dos termos penal e crime de forma bastante relevante, possivelmente indicando medidas de segurança pública. Nesse caso são comparadas nuvens de rótulos tradicionais.

Dias e Paiva [5] apresentam outra aplicação de nuvens de rótulos tradicionais, que mostra quais os cômodos mais comentados em relação aos aluguéis no *Airbnb*. Nesse caso, fazem também uma análise de sentimentos, permitindo a separação dos cômodos mais elogiados dos mais criticados. Essa tipo de trabalho tem grande interesse da indústria. Uma rápida análise visual das nuvens geradas a

[5] https://dadosabertos.camara.leg.br

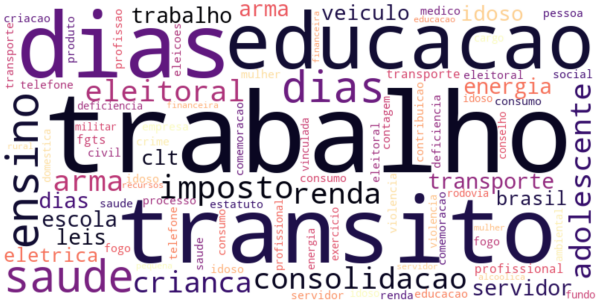


Fig. 10: Relevância de termos para o período 2000-2019. Fonte: Paiva e Pinho [6]

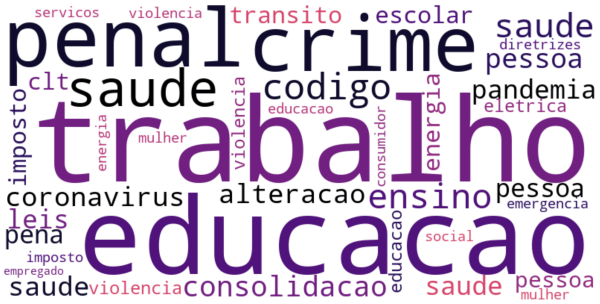


Fig. 11: Relevância de termos para o período 2020-2022. Fonte: Paiva e Pinho [6]

partir do texto das revisões em inglês no site da empresa mostra que enquanto banheiros e cozinhas recebem a maior quantidade de opinões positivas, banheiros recebem a maior quantidade de opiniões negativas. As palavras que aparecem na nuvem foram determinadas por uma lista criada pelos autores, o que também é uma opção para casos onde se busca informação específica. A análise de sentimentos foi feita a partir da atribuição de polaridades a partir da biblioteca `textblob` [36] e as nuvens com a biblioteca `wordcloud` [35].

Já Boechat e Kuchpil [37], apresentaram análise das publicações de Computação e como elas vêm se alterando ao longo do tempo na UFRJ, desta vez baseados em nuvens de rótulos diferenciais. Para isso o autor gerou um nuvem de rótulos de sumário a partir dos títulos de todas as publicações entre 2002 e 2022, apresentada na Figura 12 e ainda a nuvem de rótulos diferencial para cada três anos. Como exemplo, as Figuras 13 e 14 permitem avaliar o que fica constante e o que tem mudado pesquisa nos últimos 20 anos, e indica o compromisso com a inovação, que aparece fortemente nos dois casos. Pode-se observar que a palavra "innovation" não aparece nos anos intermediários como relevante.

Esses exemplos mostram diferentes tipos de análise que podem ser feitas: comparando itens no tempo, pela atuação de partidos políticos, pelo interesse específico em negócios, etc. Muitos usos podem ser imaginados. Kui, Liu, Liu et al. [38], por exemplo, afirmam que nuvens de rótulos, no contexto educational, são uma ferramenta para predição e recomendação de recursos de aprendizado.

Fig. 12: Nuvem de rótulos de sumário para os títulos das publicações entre 2002 e 2022. Fonte: Boechat e Kuchpil [37]

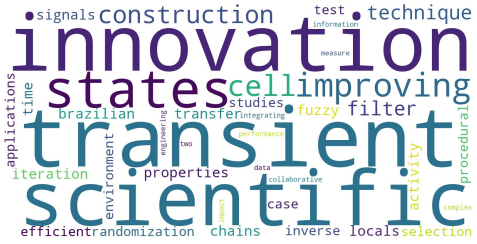


Fig. 13: Nuvem de rótulos de diferencial para os títulos das publicações entre 2002 a 2004. Boechat e Kuchpil [37]

A. *Trabalhos Correlatos*

Outros grupos de trabalho investigaram o uso ou aplicaram diretamente nuvens de rótulos. Não é o foco deste trabalho fazer uma revisão ou uma comparação, porém alguns trabalhos são interessantes tanto por fornecer indicações sobre o uso das nuvens de rótulos, como por servir como inspiração para aplicações.

Em termos de uso de nuvens como ferramenta de recuperação da informação, Sinclair e Cardew-Hall [39] concluiram que a busca por meio de nuvens de rótulos ser útil como apoio a navegação e busca por serendipidade mas, por outro lado, a interface de busca é melhor para os usuários com objetivos claros. Quanto à visualização, já citamos anteriormente os trabalhos de Viégas e Wattenberg [7] e Viégas, Wattenberg e Feinberg [9], que estudaram como tornar nuvens de rótulo mais corretas em relação as regras de design, propondo a ferramenta *wordle*.

Quanto a usos específicos, Egler, Costa e Gonçalves [40] usam nuvens de palavras para trazer significado à descrição dos assuntos tratados por grupos políticos de direita brasileiro nas redes sociais, usando a cor para identificar o tipo de mídia onde as palavras foram encontradas. O uso de nuvens de rótulos para avaliar simultaneamente uma grande quantidade de documentos é uma das suas melhores aplicações.

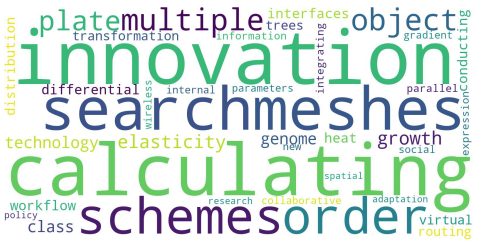


Fig. 14: Nuvem de rótulos de diferencial para os títulos das publicações entre 2020 a 2022. Boechat e Kuchpil [37]

Outra aplicação interessante foi proposta por Al-Msie'deen [41]: o uso de nuvens de rótulos para visualização de código fonte orientado a objeto, criando nuvens para identificadores em geral e também divididos em pacotes, classes, métodos e atributos. Ela mostra que atributos específicos podem ser obtidos dos conceitos, no caso o tipo de identificador, e representados na nuvem. Se a práticas de dar nomes significativos as entidades do programa for seguida, a visualização dessas nuvens ajuda então a entender o código e localizar entidades, caso contrário ela mostra que a prática não foi seguida e o código precisa ser refatorado.

Alguns artigos também mostram variações possíveis. Momin, Kays e Sadri [42] apresenta Redes de Bigramas, uma representação gráfica onde palavras que se relacionam para a par são mostradas como um grafo, onde as palavras são os vértices e os relacionamentos são os arcos.

Li, Dong e Yuan [43] mostram algumas alternativas para incluir nuvens de rótulos em mapas como estratégia de visualização de que pontos de interesse existem em uma região. Assim, cada área do mapa é na verdade uma nuvem de rótulos diferente.

Outra variante é uma nuvem de dados, *data cloud*. Nessa variante, os rótulos são algum campo específico de uma base de dados, como o nome de um país, e os atributos são outros campos, como usar a população para o tamanho da fonte e a cor para o continente.

Esses exemplos mostram que existem ainda caminhos para inovação, principalmente na área de aplicação.

VI. GTAGLIB

GTAGLIB é uma biblioteca em Python, descrita em Mattos [44], capaz de gerar nuvens de rótulos diferenciais e de sumário. Sua implementação foi um projeto final de cadeira, e foi motivado pela percepção do aluno da ausência de uma ferramenta específica para realizar a criação desses tipos de nuvem, o que foi feito de forma ad-hoc por outras pessoas que trabalharam com esses dois conceitos no LINE.

A biblioteca implementa dois métodos, sendo que o primeiro usa os termos mais relevantes no documento e o segundo usa, além desses termos, sinônimos presentes no WordNet [29]. Além disso, o primeiro parte das nuvens diferenciais e depois gera a nuvem sumário, enquanto o segundo gera a nuvem sumário antes de gerar as diferenciais.

O método 1 segue os seguintes passos:

1) tokenização;
2) uso do stemmer de Porter [26];
3) remoção das *stopwords*;
4) filtragem para aceitar apenas substantivos, por meio de POS-Tagging [19], que rotula a classe gramatical da palavra;
5) geração de bigramas;
6) cálculo do TF-IDF de cada palavra;
7) escolha dos termos mais relevantes como campo semântico;
8) criação da nuvem abstrata a partir do campo semântico;

9) criação da nuvem diferencial (são criadas primeiro);
10) criação da matriz termo-documento;
11) aplicação do LSA;
12) aplicação do Coeficiente de Correlação de Pearson na matriz reconstruída para obter uma matriz de correlação de termos;
13) escolha de um termo raiz;
14) escolha dos termos mais correlacionados com o termo raiz, e
15) contrução da matriz de sumário.

O método 2 repete o processo até o passo 7, e segue:

8) expansão do campo semântico com os sinônimos das palavras contidas no próprio campo, usando o Wordnet;
9) criar a nuvem abstrata de cada documento;
10) ordenar a união dos termos das nuvens;
11) escolher os primeiro quarenta termos como os de maior relevância, e
12) para cada documento, construir a nuvem diferencial com os termos não presentes no sumário, colorindo os termos expandidos em outra cor.

A Listagem 1 apresenta um exemplo de uso.

Listagem 1: Fragmento de código com exemplo de uso da GTAGLIB, partindo do sumário para a diferencial

```
from gtaglib.generator import TagGenerator

set_summary, differential = TagGenerator(
    semantic_field_size=40,
    stemmer="porter",
    generate_bigrams=True,
    use_tfidf=True
).generate_tag_cloud(dataset,1,root="step",
    outputdir="myoutputdir")
```

A biblioteca pode ser encontrada em https://github.com/GuilhermeCaeiro/gtaglib. Sua classe principal é a `TagGenerator`. Para gerar a nuvem de rótulos abstratas podem ser usados os métodos `generate`, ou `generate_tag_clouds`. Para gerar as imagens, a `gtaglib` atualmente usa a biblioteca `wordcloud` internamente.

VII. Conclusões

NESTE artigo foi feita uma revisão do trabalho do grupo do LINE com nuvens de rótulos. Após a introdução foi feita uma revisão informal do assunto, seguida de uma reestruturação do modelo previamente apresentado por Xexeo, Morgado e Fiuza [1] e [2]. A esse modelo se seguiram as explicações para técnicas de criação de nuvens de rótulos. Depois foram apresentadas algumas aplicações do grupo e trabalhos correlatos, sendo que a penúltima seção apresenta uma biblioteca para gerar nuvens de dados abstratas, sumário e diferencial. Com essa revisão, o interessado em nuvens de rótulos pode encontrar orientações para trabalhos teóricos e aplicados na área.

Como trabalhos futuros, é possível sugerir três áreas: a área de nuvens de rótulos semânticos, isto é, a inserção de mais conhecimento sobre o tópico nas nuvens; a melhoria dos algoritmos de visualização; a compreensão do efeito de usar nuvens de rótulos como representação de documento e, por último, utilizações não triviais e variações de formato dessas nuvens.

Referências

[1] G. Xexeo, F. Morgado e P. Fiuza, "Differential Tag Clouds: Highlighting Particular Features in Documents," em *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 03*, sér. WI-IAT '09, USA: IEEE Computer Society, 2009, pp. 129–132.

[2] F. F. Morgado, "Representação de Documentos Através de Nuvens de Termos," Dissertação de Mestrado, Rio de Janeiro, 2010.

[3] Wikipedia. "Tag Cloud." (30 de out. de 2022), endereço: https://en.wikipedia.org/wiki/Tag_cloud#:~:text=The%20first%20tag%20clouds%20on,designer%20Stewart%20Butterfield%20in%202004. (acesso em 30/10/2022).

[4] Ombre Blanches. "Cloud on Title." (10 de mai. de 2012), endereço: https://ombresblanches.wordpress.com/2012/05/10/cloud-on-title/ (acesso em 30/10/2022).

[5] C. C. Dias e T. M. Paiva, "Análise Semântica Latente (LSA) Aplicada a Projetos de Lei," em *Explorações em Mineração de Texto*, ES-782/22, G. Xexéo, ed., Relatório Técnico, 2022, pp. 17–24. endereço: https://www.cos.ufrj.br/index.php/pt-BR/publicacoes-pesquisa/details/15/3060.

[6] B. D. de Paiva e J. P. L. Pinho, "Most Relevant Rooms in an AirbnbAccommodation: A QuantitativeAnalyses on Reviews," em *Explorações em Mineração de Texto*, ES-782/22, G. Xexéo, ed., Relatório Técnico, ago. de 2022, pp. 5–16. endereço: https://www.cos.ufrj.br/index.php/pt-BR/publicacoes-pesquisa/details/15/3060.

[7] F. B. Viégas e M. Wattenberg, "TIMELINES Tag Clouds and the Case for Vernacular Visualization," *Interactions*, v. 15, n. 4, pp. 49–52, jul. de 2008, ISSN: 1072-5520.

[8] J. Nielsen. "Tag Cloud Examples," Nielsen Normam Group. (23 de mar. de 2009), endereço: https://www.nngroup.com/articles/tag-cloud-examples/ (acesso em 02/11/2022).

[9] F. Viégas, M. Wattenberg e J. Feinberg, "Participatory Visualization with Wordle," *IEEE transactions on visualization and computer graphics*, v. 15, pp. 1137–44, nov. de 2009.

[10] A. W. Rivadeneira, D. M. Gruen, M. J. Muller e D. R. Millen, "Getting Our Head in the Clouds: Toward Evaluation Studies of Tagclouds," em *Proceedings of the SIGCHI Conference on Human Factors*

in Computing Systems, sér. CHI '07, San Jose, California, USA: Association for Computing Machinery, 2007, pp. 995–998.

[11] I. Marinchev, "Practical Semantic Web – Tagging and Tag Clouds," *Journal Cybernetics and Information Technologies*, v. 6, n. 3, pp. 33–39, 2006.

[12] S. Bateman, C. Gutwin e M. Nacenta, "Seeing Things in the Clouds: The Effect of Visual Features on Tag Cloud Selections," em *Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia*, sér. HT '08, Pittsburgh, PA, USA: Association for Computing Machinery, 2008, pp. 193–202.

[13] Y. Hassan-Montero e V. Herrero-Solana, "Improving Tag-Clouds as Visual Information Retrieval Interfaces," em *Proceedings of the 1st International Conference on Multidisciplinary Information Sciences and Technologies InSCiT*, 2006.

[14] K. Bielenberg e M. Zacher, "Groups in Social Software: Utilizing Tagging to Integrate Individual Contexts for Social Navigation," M.Sc. Thesis, 2006.

[15] K. Devlin, *The Joy of Sets: Fundamentals of Contemporary Set Theory* (Undergraduate Texts in Mathematics), 2ª ed. Berlin: Springer New York, 1993.

[16] C. J. Date, *An introduction to database systems*, Eighth. Boston, MA, USA: Pearson/Addison Wesley, 2004, pp. xxvii + 983 + 22.

[17] T. Berners-Lee, R. Fielding e L. Masinter, "Uniform Resource Identifier: Generic Syntax, RFC 3986," rel. técn., jan. de 2005.

[18] R. Baeza-Yates e B. Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology behind Search*. USA: Addison-Wesley Publishing Company, 2011.

[19] D. Jurafsky e J. H. Martin, *Speech and Language Processing*. 12 de jan. de 2022, draft. endereço: https://web.stanford.edu/~jurafsky/slp3/ (acesso em 03/11/2022).

[20] M. de Mattos Mayworm, "Um Crawler Peer-to-Peer Baseado em Agentes," M.Sc. Programa de Engenharia de Sistemas, COPPE/UFRJ, Rio de Janeiro, 28 de set. de 2007.

[21] Y. Shinyama, P. Guglielmetti e P. Marsman. "pdfminer.six." (2020), endereço: https://pdfminersix.readthedocs.io/en/latest/ (acesso em 04/11/2022).

[22] L. Richardson, "Beautiful soup documentation," *April*, 2007.

[23] C. D. Manning, P. Raghavan e H. Schütze, "Probabilistic information retrieval," em *Introduction to Information Retrieval*. Cambridge University Press, 2009, pp. 201–217.

[24] P. Norvig. "How to Write a Spelling Corrector." (2007), endereço: https://norvig.com/spell-correct.html (acesso em 19/01/2022).

[25] P. Norvig, "Natural Language Corpus Data," em *Beautiful Data: The Stories Behind Elegant Data Solutions*, T. Segaran e J. Hammerbacher, ed., 2009, pp. 219–242.

[26] M. F. Porter, "An Algorithm for Suffix Stripping," *Program*, v. 14, n. 3, pp. 130–137, jul. de 1980. endereço: https://tartarus.org/martin/PorterStemmer/def.txt (acesso em 22/03/2022).

[27] T. K. Landauer, D. S. McNamara, S. Dennis e W. Kintsch, *Handbook of Latent Semantic Analysis*, 1ª ed. Lawrence Erlbaum, fev. de 2007.

[28] D. M. Blei, A. Y. Ng e M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, v. 3, n. null, pp. 993–1022, mar. de 2003, ISSN: 1532-4435.

[29] C. Fellbaum, ed., *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998.

[30] V. de Paiva, A. Rademaker e G. de Melo, "OpenWordNet-PT: An Open Brazilian Wordnet for Reasoning," em *Proceedings of COLING 2012: Demonstration Papers*, Published also as Techreport http://hdl.handle.net/10438/10274, Mumbai, India: The COLING 2012 Organizing Committee, dez. de 2012, pp. 353–360. endereço: http://www.aclweb.org/anthology/C12-3044.

[31] J. Jo, B. Lee e J. Seo, "WordlePlus: Expanding Wordle's Use through Natural Interaction and Animation," *IEEE Computer Graphics and Applications*, v. 35, n. 6, pp. 20–28, 2015. DOI: 10.1109/MCG.2015.113.

[32] M. A. Hearst, E. Pedersen, L. Patil, E. Lee, P. Laskowski e S. Franconeri, "An Evaluation of Semantically Grouped Word Cloud Designs," *IEEE Transactions on Visualization and Computer Graphics*, v. 26, n. 9, pp. 2748–2761, 2020.

[33] S. Bird, E. Loper e E. Klein, *Natural Language Processing with Python*. O'Reilly Media Inc, 2009.

[34] T. W. Workbench, "Eibe Frank and Mark A. Hall and Ian H. Witten," em *Data Mining: Practical Machine Learning Tools and Techniques*, 4ª ed. Morgan Kaufmann, 2106, Online Appendix.

[35] A. Mueller. "WordCloud for Python documentation." (2020), endereço: https://amueller.github.io/word_cloud/index.html (acesso em 03/11/2022).

[36] S. Loria, "textblob Documentation," *Release 0.15*, v. 2, 2018.

[37] P. Boechat e P. Kuchpil, "Análise da evolução dos títulos de pesquisas em computação da UFRJ por meio de Differencial Tag Clouds," em *Explorações em Mineração de Texto*, ES-782/22, G. Xexéo, ed., Relatório Técnico, ago. de 2022, pp. 26–33. endereço: https://www.cos.ufrj.br/index.php/pt-BR/publicacoes-pesquisa/details/15/3060.

[38] X. Kui, N. Liu, Q. Liu, J. Liu, X. Zeng e C. Zhang, "A survey of visual analytics techniques for online education," *Visual Informatics*, 2022.

[39] J. Sinclair e M. Cardew-Hall, "The folksonomy tag cloud: when is it useful?" *Journal of Information Science*, v. 34, n. 1, pp. 15–29, 2008.

[40] T. T. C. Egler, T. Costa e P. P. Gonçalves, "A (In)Visibilidade Da Rede Tecnopolítica Bolsonarista," *AR@CneRevista Electrónica de Recursos*

en Internet Sobre Geografía Y Ciencias Sociales, v. XXV, n. 251, 1 de out. de 2021.

[41] R. Al-Msie'deen, "Tag Clouds for Object-Oriented Source Code Visualization," *Engineering, Technology & Applied Science Research*, v. 9, n. 3, pp. 4243–4248, 2019.

[42] K. Momin, H. M. I. Kays e A. M. Sadri. "Identifying Crisis Response Communities in Online Social Networks for Compound Disasters: The Case of Hurricane Laura and Covid-19." (out. de 2022), endereço: https://arxiv.org/abs/2210.14970 (acesso em 04/11/2022).

[43] C. Li, X. Dong e X. Yuan, "Metro-Wordle: An Interactive Visualization for Urban Text Distributions Based on Wordle," *Visual Informatics*, v. 2, n. 1, pp. 50–59, 2018, Proceedings of PacificVAST 2018, ISSN: 2468-502X.

[44] G. C. de Mattos, "GTAGLIB: a Library to Generate Tags and Tag Clouds," em *Explorações em Mineração de Texto*, ES-782/22, G. Xexéo, ed., Relatório Técnico, ago. de 2022, pp. 35–48. endereço: https://www.cos.ufrj.br/index.php/pt-BR/publicacoes-pesquisa/details/15/3060.

[45] G. Xexéo, "Explorações em Mineração de Texto," Programa de Engenharia de Sistemas - COPPE/UFRJ, rel. técn. ES-782/22, ago. de 2022, Relatório Técnico. endereço: https://www.cos.ufrj.br/index.php/pt-BR/publicacoes-pesquisa/details/15/3060.