Processamento de Linguagem Natural na Educação Superior: Comparando Automaticamente Currículos de Cursos de Computação

Bruno Henrique Rasteiro, Rafael Augusto Monteiro, Thiago Alexandre Salgueiro Pardo
Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
bruno.rasteiro@usp.br, rafael.augusto.monteiro@usp.br, taspardo@icmc.usp.br

1. Introdução

A Computação é uma área de enorme importância atualmente, principalmente devido às demandas crescentes da sociedade, o que impacta na qualidade esperada dos cursos de formação no ensino superior brasileiro. Devido às diferentes necessidades do mercado e da própria academia, esses cursos apresentam enfoques variados, sendo que há 3 que são amplamente difundidos e reconhecidos: Bacharelado em Ciências de Computação (BCC), Engenharia de Computação (EC) e Bacharelado em Sistemas de Informação (BSI). Nesse cenário, segundo portal e-MEC[1] do governo federal, há, atualmente, mais de 300 cursos presenciais desses sendo oferecidos gratuitamente por diversas Instituições de Ensino Superior (IES) públicas.

A grande gama e as diferentes vertentes de disciplinas oferecidas nas IES tornam complexa a comparação entre os cursos. A complexidade aumenta quando as comparações são feitas por alunos do ensino médio, que não possuem conhecimento específico na área e dificilmente entendem os conteúdos apresentados nos currículos. Portanto, torna-se interessante uma ferramenta que facilite a compreensão da estrutura curricular dos cursos e permita uma comparação visual entre os cursos oferecidos em diferentes IES, auxiliando os alunos na escolha de suas carreiras. Esse tipo de funcionalidade também auxilia diretamente gestores da educação superior, pois permite a comparação entre cursos de forma mais rápida e eficiente, subsidiando decisões educacionais e a melhoria dos cursos.

Nesse artigo, é proposta a criação de uma aplicação que permite, por meio do uso de conceitos e técnicas da área de Processamento de Linguagem Natural (PLN), processar automaticamente os dados relativos a diferentes grades curriculares da área de computação e apresentar uma descrição e comparação gráfica dos resultados obtidos, com base em currículos de referência estabelecidos por associações de computação, como ACM, IEEE e SBC, facilitando a compreensão das diferenças e semelhanças entre os cursos.

O artigo apresenta seções explicando como são formulados os cursos de computação, o que é a área de PLN e como ela pode ser utilizada no problema descrito, e como é o sistema proposto e os resultados que produz.

1. emec.mec.gov.br

2. Materiais e Métodos

2.1. Currículos e Cursos de Computação

Os cursos de computação são divididos em dois grupos, aqueles que possuem a computação como atividade fim e os que têm a computação como atividade meio. No primeiro grupo está o Bacharelado em Ciências de Computação e a Engenharia de Computação; no segundo, encontra-se o Bacharelado em Sistemas de Informação[2].

Para contribuir com a padronização dos cursos de computação, foram desenvolvidos currículos de referência, como o currículo tradicional da SBC (Sociedade Brasileira de Computação) de 2005[3] e o *Computer Science Curricula* de 2013[4] da ACM em parceria com a IEEE. Com o advento desses currículos, as universidades ganharam um modelo de elaboração e adaptação de suas graduações, podendo padronizar em certa medida o ensino superior de computação. Neste trabalho, o currículo de referência da SBC foi tomado como base, dada sua importância na evolução da área no Brasil. Esse documento agrupa as disciplinas dos cursos em 6 núcleos, sendo eles: Fundamentos da Computação, Tecnologia da Computação, Matemática, Contexto Social e Profissional, Eletrônica e Ciências Básicas.

Para a escolha dos parâmetros comparativos e testes do sistema desenvolvido, adotamos os três cursos de computação oferecidos pelo Instituto de Ciências Matemáticas e de Computação (ICMC) da Universidade de São Paulo (USP), instituto ao qual os autores deste artigo estão filiados. Além dos três, para teste do sistema desenvolvido, utilizamos os melhores cursos de Ciências de Computação do Brasil, seguindo como critério o posicionamento no Ranking Universitário da Folha (RUF)[5].

2. Há também outros cursos, como Engenharia de Software e Licenciatura em Computação, que são relativamente menos difundidos e/ou são oferecidos em outras modalidades (como cursos sequenciais/tecnológicos), e, portanto, não são abordados aqui.

3. www.sbc.org.br/documentos-da-sbc/category/131-curriculos-de-referencia

4. www.acm.org/education/curricula-recommendations

5. ruf.folha.uol.com.br

2.2. Processamento de Linguagem Natural

PLN é uma subárea da Inteligência Artificial que visa habilitar as máquinas a lidar com a linguagem humana, realizando, por exemplo, tarefas como tradução automática, sumarização de textos e correção gramatical (Jurafsky e Martin, 2009). Para tanto, faz-se uso de diversos recursos e ferramentas linguístico-computacionais, como léxicos, gramáticas e analisadores textuais de níveis variados.

Neste trabalho, a tarefa de comparar currículos foi feita utilizando-se métricas de similaridade lexical normalmente usadas em PLN. Uma das mais simples é a *Word Overlap*, que se resume na razão do somatório de palavras repetidas pelo número total de palavras. O resultado indica a percentagem de similaridade entre os documentos comparados.

A métrica utilizada neste estudo é mais sofisticada, chamada de similaridade do cosseno, proposta por Salton e McGill (1987). Ela consiste em montar uma representação vetorial para os documentos e, em seguida, calcular o valor do cosseno do ângulo formado por eles. Quanto mais alto o resultado, mais similares são os vetores. O vetor de cada documento é formado de maneira que cada posição i representa uma palavra, sendo o conteúdo da posição i o número de ocorrências da palavra em questão naquele documento. Esta forma de representação é denominada de *bag of words*. O cálculo da medida do cosseno é feito pela fórmula a seguir (Manning e Schütze, 2003), considerando os vetores v e w de dois documentos que se deseja comparar e que as palavras vão da posição $i = 0$ até N dos vetores.

$$cosseno(\vec{v}, \vec{w}) = \frac{\vec{v} \cdot \vec{w}}{|\vec{v}||\vec{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2} \sqrt{\sum_{i=1}^{N} w_i^2}}$$

Para a tarefa de encontrar a similaridade entre ementas, é interessante remover da comparação algumas palavras que não acrescentam diferenças significativas no cálculo de similaridade de conteúdo das ementas. Portanto, antes de realizar a medida do cosseno, foi feita a remoção das *stopwords*, que incluem as palavras de classe fechada, como preposições, pronomes e artigos.

As métricas foram testadas na comparação das disciplinas equivalentes dos cursos de Ciências de Computação do ICMC-USP e do Instituto de Matemática e Estatística (IME) da USP, e do ICMC-USP e da Universidade Federal do Rio Grande do Sul (UFRGS). A Tabela 1 apresenta os resultados obtidos.

Disciplina	*Word Overlap*	Cosseno
Estatística	0.31	0.49
Alg. e Estruturas de Dados	0.20	0.41
Bases de Dados	0.21	0.35
Cálculo Numérico	abaixo de 0.20	0.35

Tabela 1: Resultados obtidos pelas métricas de similaridade *Word Overlap* e Cosseno para a comparação ICMC-USP e UFRGS.

A métrica do cosseno apresentou valores de similaridade mais expressivos para disciplinas equivalentes classificadas manualmente, tornando-se a escolhida para o sistema proposto nas sessões seguintes.

3. Um Sistema de Comparação de Currículos

O sistema proposto realiza duas tarefas: (i) analisa currículos e classifica suas disciplinas de acordo com um currículo de referência e (ii) compara currículos de diferentes cursos. Para a tarefa (i), o sistema toma o currículo da SBC como referência e utiliza as disciplinas de seus núcleos como base para analisar um ou mais currículos de cursos de computação. Para a tarefa (ii), avalia-se a similaridade entre as ementas das disciplinas de dois cursos diferentes. Por fim, são gerados gráficos que ressaltam de forma visual a estruturação dos cursos de acordo com a referência fornecida, assim como gráficos contrastivos entre os cursos, mostrando as similaridades e diferenças entre as disciplinas de suas grades curriculares.

A entrada para o sistema é composta simplesmente por arquivos texto com os dados de cada curso, sendo que o arquivo de um curso contém a listagem de disciplinas, com seus títulos, ementas e número de créditos-aula. Esses arquivos foram montados manualmente, coletando-se as informações dos cursos diretamente de suas páginas institucionais. As informações do currículo de referência da SBC também são fornecidas da mesma forma. Futuramente, pode-se automatizar essa etapa de *crawling* das informações dos cursos, que ainda não foi feita por não ser o foco desse estudo.

Tendo-se os dados de entrada, para a tarefa de analisar as disciplinas de um curso, é verificada a similaridade entre as ementas de suas disciplinas e as ementas das disciplinas do currículo de referência. A disciplina será classificada de acordo com a disciplina do currículo de referência que apresentar maior similaridade com ela. Para a tarefa de encontrar similaridades e diferenças entre dois cursos, é realizada a comparação entre as ementas das disciplinas de cada núcleo dos cursos. Aquelas que apresentarem um valor de similaridade (de acordo com a medida do cosseno) maior que um limite pré-definido são consideradas disciplinas equivalentes, enquanto as demais são consideradas disciplinas não equivalentes. Foi escolhido um valor mínimo limite (*threshold*) empiricamente, com base na acurácia de casos verificados manualmente. Adotou-se um *threshold* de 20%.

Com os dados em mãos, é possível gerar gráficos descritivos e comparativos entre os cursos utilizando as bibliotecas *matplotlib* e *matplotlib-venn*, da linguagem *python*. As Figuras 1, 2, 3 e 4 mostram exemplos de gráficos gerados pelo sistema.

Pode-se ver, na Figura 1, a distribuição de créditos-aula do curso de Ciências de Computação do ICMC em função dos núcleos do currículo da SBC. Nota-se claramente que a maior parte dos créditos concentra-se nos núcleos de Fundamentos de Computação e de Tecnologias de Computação. É interessante notar que o curso tem em sua grade disciplinas

de Eletrônica, mas isso não é mostrado no gráfico devido a falha de detecção desta similaridade pela medida utilizada. Esse tipo de situação é comum em PLN, pois, dada a variedade e riqueza da língua, é muito difícil atingir 100% de acurácia nas análises.

Na Figura 2, vê-se a comparação da distribuição dos créditos nos cursos de Ciências de Computação do ICMC e da UFRGS. Nesse gráfico contrastivo, ficam claras algumas diferenças, como o maior investimento do ICMC em disciplinas de Fundamentos de Computação e o maior investimento da UFRGS em disciplinas de Tecnologias de Computação.

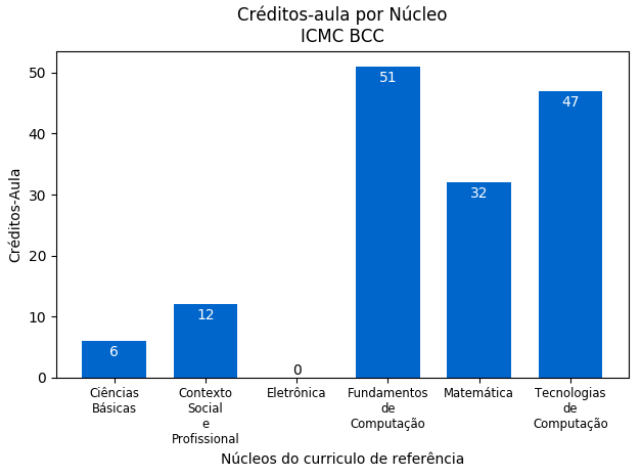

Figura 1: Distribuição da carga horária do BCC-ICMC, de acordo com os núcleos definidos pelo currículo de referência da SBC. É notável a maior carga horária em disciplinas de fundamentos e tecnologias da computação.

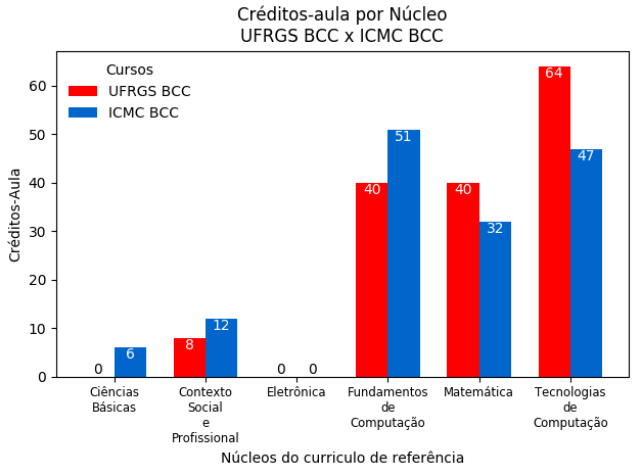

Figura 2: Comparação da distribuição das cargas horárias entre BCC-UFRGS e BCC-ICMC, de acordo com os núcleos definidos pelo currículo de referência da SBC. É possível notar que o BCC-ICMC possui maior ênfase em fundamentos, enquanto o BCC-UFRGS investe mais em tecnologias.

As Figuras 3 e 4 são complementares, mostrando um diagrama de Venn com similaridades e diferenças entre cursos de forma mais refinada, listando as disciplinas únicas e compartilhadas entre os cursos.

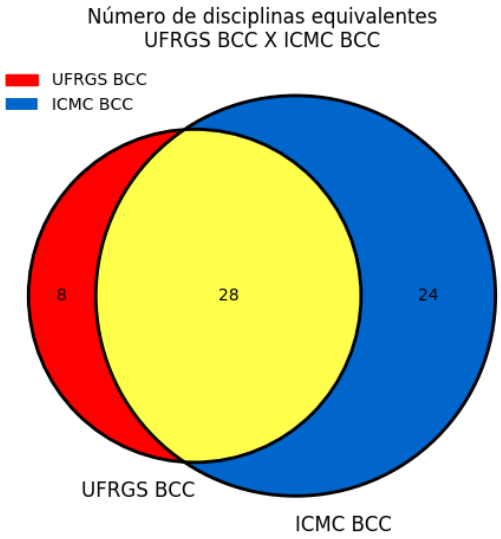

Figura 3: Comparação entre as disciplinas dos cursos BCC-ICMC e BCC-UFRGS. O diagrama exibe o número de disciplinas únicas (sem equivalentes) do BCC-UFRGS em vermelho, o número de disciplinas únicas (sem equivalentes) do BCC-ICMC em azul, e o número de disciplinas equivalentes entre os cursos em amarelo.

O sistema proposto e os gráficos gerados são parte de uma interface web de fácil uso, a qual poderá ser publicamente acessada futuramente, permitindo ao usuário (aluno do ensino médio ou gestor em IES) conduzir análises comparativas entre cursos.

4. Considerações Finais

Este artigo apresentou uma proposta de usar conceitos e técnicas de PLN para apoiar atividades relacionados ao ensino superior. Em particular, os cursos de Computação foram utilizados para teste do sistema proposto. Como aplicação futura, destaca-se a generalização do sistema para a comparação de currículos de qualquer área de interesse, assim como o desenvolvimento de um aplicativo para dispositivos móveis.

Agradecimentos

À Pró-Reitoria de Cultura e Extensão Universitária (PRCEU), pelo financiamento deste projeto através do Programa Unificado de Bolsas (PUB) da Universidade de São Paulo, e à FAPESP.

Referências

[1] Jurafsky, D. and Martin, J.H. (2009). *Speech and Language Processing*. Prentice Hall.

[2] Manning, C.D. and Schütze, H. (2003). *Foundations of Statistical Natural Language Processing*. The MIT Press.

[3] Salton, G. and McGill, M.J. (1987). *Introduction to Modern Information Retrieval*. McGraw Hill Computer Science Series.

Figura 4: Comparação entre as disciplinas dos cursos BCC-ICMC e BCC-UFRGS. O diagrama exibe à esquerda o nome das disciplinas únicas (sem equivalentes) do BCC-UFRGS, à direita o nome das disciplinas únicas (sem equivalentes) do BCC-ICMC, e ao centro o nome das disciplinas equivalentes entre os dois cursos. Este gráfico utiliza o mesmo esquema de cores da Figura 3, sendo um complemento para a visualização do gráfico anterior.