Pesquisa de Conceitos em Microsoft Cognitive Search

José Diogo[1], José Henrique Mamede[2]

[1] Universidade Aberta jose.alexandre.diogo@gmail.com
[2] Universidade Aberta jose.mamede@uab.pt

Resumo

O processo de revisão sistemática de literatura em investigação continua a apresentar-se como um processo com um elevado custo de recursos humanos e de tempo. Com vista em otimizar este processo pretende-se estudar a performance da ferramenta de pesquisa Cognitive Search da Microsoft que contem funcionalidades de inteligência artificial (IA). Neste trabalho foi implementada uma solução de pesquisa, i.e., parametrização do serviço de pesquisa, que produz uma classificação de relevância dos artigos científicos. Uma análise qualitativa aos artigos científicos foi efetuada para analisar a performance da solução de pesquisa e habilidades de inteligência artificial da ferramenta. O tema da revisão sistemática é "how is artificial intelligence (AI) being used in Higher Education (HE) today, involving tree dimensions: learning with AI, learning about AI and learning for AI".

palavras-chave: revisão sistemática, pesquisa conceitos, inteligência artificial, Microsoft Cognitive Search

Title: Concepts search using Microsoft Cognitive Search

Abstract: The **s**ystematic review process of research literature continues to be a very time and human resource expensive process. With the objective of optimizing this process we intend to study the performance of Microsoft Cognitive Search service which contains artificial intelligence capabilities. In this work the search service tool was configured and parameterized (search solution) to produce a classification ranking of the research articles. These were manually analysed to infer on the performance of the search solution. The topic of the systematic review is "how is artificial intelligence (AI) being used in Higher Education (HE) today, involving tree dimensions: learning with AI, learning about AI and learning for AI".

keywords: Systematic Review, concepts search, artificial intelligence, Microsoft Cognitive Search

1. Introdução

Neste trabalho realizamos uma análise à potencialidade da utilização da ferramenta Cognitive Search da Microsoft, integrando funcionalidades e capacidades de inteligência artificial (IA) na seleção de conceitos, a partir de um conjunto vasto, pré-definido, de artigos científicos.

Atualmente o processo de revisão sistemática de literatura em processos de investigação, de forma a encontrar informação relevante sobre um tema particular implica um processo manual de revisão de milhares de documentos por parte do investigador, traduzindo-se numa operação morosa e, ainda que se apliquem metodologias científicas de suporte a essa atividade, acaba sempre por traduzir uma certa subjetividade.

A automatização do processo de revisão, objetivo deste trabalho, tem como objetivo reduzir o tempo e carga de trabalho do investigador durante esse processo de revisão. Se os resultados obtidos puderem ser comparáveis aos resultados obtidos pelo processo manual de análise do mesmo conjunto de documentos, então poderemos estar perante uma nova forma, mais rápida e menos trabalhosa de realizar esta atividade.
A Microsoft tem disponível uma ferramenta de pesquisa, a Cognitive Search, que é um serviço de pesquisa na cloud ideal para pesquisa de documentos, estruturados e não estruturados, de qualquer tipo e em larga escala e que permite incorporar funcionalidades e capacidade da IA à pesquisa. Desta forma, explorámos as capacidades e potencialidades efetivas da ferramenta MS Cognitive Search no processo de revisão sistemática de literatura em processos de investigação científica.

O objetivo geral do trabalho foi implementar uma ou várias soluções de pesquisa de conceitos utilizando a ferramenta Microsoft Cognitive Search que permita fazer uma triagem a milhares de artigos científicos e selecionar os mais relevantes sobre a investigação do tema “how is AI being used in Higher Education (HE) today, involving tree dimensions: learning with AI, learning about AI and learning for AI”. Para esta implementação foi efetuado o desenvolvimento adicional de programas que permitam a real utilização deste serviço disponibilizado pela Microsoft, incorporando a sua utilização neste contexto.

A base de artigos científicos utilizados foi obtida das seguintes fontes: ACM Digital Library (http://portal.acom.org); Google Scholar (http://scholar.google.com); IEE Digital Library (http://ieeexplore.ieee.org); ISI Web of Science (http:// www.isiknowledge.com).

A “string” de pesquisa de referência utilizada é:
(“Artificial Intelligence” AND “Higher Education” AND (“Learning” OR “Teaching”))
Espera-se que a solução criada seja capaz de fazer uma seleção de documentos relevantes para o tema indicado comparável àquela que um investigador faria.

2. Serviço de Pesquisa

Como referido anteriormente foi utilizada a ferramenta *Azure cognitive search* que é um serviço de pesquisa na *cloud* da Microsoft com competências de IA pré treinadas / construídas que podem ser aplicadas aos dados extraídos de documentos. O *Azure cognitive search* possui também um motor de pesquisa que com capacidades de IA incorporadas: o subsistema para pesquisa semântica.

Em termos arquiteturais o serviço de pesquisa encontra-se entre o armazenamento de dados e a aplicação de cliente que envia pedidos de consulta. Para se criar uma solução de pesquisa é então necessário criar / definir os 3 principais componentes: Armazenamento de ficheiros; *Azure cognitive search* (serviço de pesquisa); Aplicação de pesquisa.

O serviço de armazenamento utilizado foi o armazenamento de *blobs* do *Azure* (Myers, 2021) que é a solução de armazenamento adequada a dados não estruturados (pdf's) que permite ser utilizada diretamente pelo serviço de pesquisa. A aplicação criada permite configurar o serviço de pesquisa e cada um dos seus componentes e executar pedidos de pesquisa. Em baixo resume-se cada um dos componentes principais do serviço de pesquisa:
-O **índice** é o conjunto de dados estruturados que vai ser pesquisável pelo motor de busca. Um índice é definido por um esquema de índice (Steen, Deng & Schiavon, 2021) que define a forma como os dados dos documentos estão organizados no índice, i.e., define os diferentes campos do índice e suas características (por exemplo: nome ficheiro, conteúdo, output habilidade cognitiva).
- O **indexante** é responsável por extrair dados dos ficheiros armazenados e colocá-los de forma estruturada no índice de pesquisa (Steen, Deng & Cortez, 2021). É no processo de indexação que são executadas as capacidades cognitivas.
-**Habilidades/ferramentas cognitivas** (Enriquecimento em IA): Estas ferramentas atuam sobre a informação extraída dos dados "originais" e geram nova informação que pode ser adicionada ao índice de pesquisa: habilidades de processamento de imagem e competências de processamento de linguagem natural (Microsoft, 2021).
-**Motor de pesquisa** com opção de utilização de pesquisa por "similaridade" e "semântica". A pesquisa por "similaridade" é a pesquisa usualmente utilizada em algoritmos de pesquisa que se baseiam no método TD/IDF ("term frequency-inverse document frequency") (Hariharan, 2021) que se baseia na frequência do termo em cada documento e na sua raridade por todos os documentos no índice (Steen, 2021b). Por sua vez a pesquisa semântica utiliza modelos pré treinados de IA de compreensão linguística e relevância semântica para classificação dos documentos pesquisados (Steen , 2021a].

3. A Solução Desenvolvida

O serviço providencia APIs e bibliotecas para várias linguagens de programação que permitem utilizar o serviço para submeter pesquisas, definir índices, indexantes, habilidades cognitivas, etc. Neste projeto utilizou-se a plataforma .NET com as bibliotecas do SDK (*software development kit*) do serviço.

A aplicação é uma aplicação em consola que permite configurar automaticamente o índice, o indexante e respetivo conjunto de habilidades cognitivas e que também permite fazer pesquisas de semelhança e pesquisas semânticas.

O projeto utiliza como base 3 serviços do *Azure*: uma conta de armazenamento; um serviço de pesquisa e um serviço de habilidades cognitivas. Os documentos são importados ao sistema de armazenamento diretamente através do portal do *Azure*. O serviço de pesquisa e o serviço de habilidades cognitivas, depois de criados, são configurados diretamente pela aplicação.

Da análise às capacidades cognitivas disponíveis foi percetível que face ao objetivo proposto apenas a habilidade de recolher "key phrases" se encontrava dentro do âmbito do projeto. O trabalho focou-se, então, em testar a pesquisa por similaridade integrando a habilidade cognitiva "key phrases" e perceber o potencial na seleção de artigos na pesquisa de conceitos. Este trabalho baseou-se na seguinte frase de pesquisa:
+"artificial intelligence"+"higher education"+(learning|teaching).

O estudo incidiu na comparação dos resultados obtidos para a frase de pesquisa descrita em cima utilizando o campo "key phrases" e o campo "conteúdo".
Em baixo é mostrado o menu inicial da aplicação e as várias opções. As opções 1 a 7 permitem configurar os vários componentes do sistema de pesquisa. A opção 1 faz a ligação da base de dados ao serviço de pesquisa. A opção dois define as habilidades cognitivas ("key phrases"). A opção 3 gera o esquema de índice. A opção 4 cria um mapa de sinónimos aos vários campos, i.e., a pesquisa sobre um determinado termo irá incluir também os seus sinónimos (Steen, 2021c). A opção 5 gera perfis de pontuação que permite que certos campos tenham mais relevância na pesquisa que outros. Esta opção não foi utilizada nos resultados. A opção 6 é utilizada quando o sistema esta configurado com as opções anteriores e executa o processo de indexação. Após a indexação estar completa, as opções 8,9 e 0 podem ser executadas para efetuar pesquisas. A opção 8 executa pesquisas por similaridade. A opção 9 executa pesquisas semânticas (fora do âmbito deste projeto). A opção 0 executa automaticamente pesquisas sobre "key phrases" e conteúdo sobre a frase de pesquisa deste projeto e guarda os dados em um documento ".csv". Esta última opção foi criada para gerar os dados a serem estudados neste trabalho.

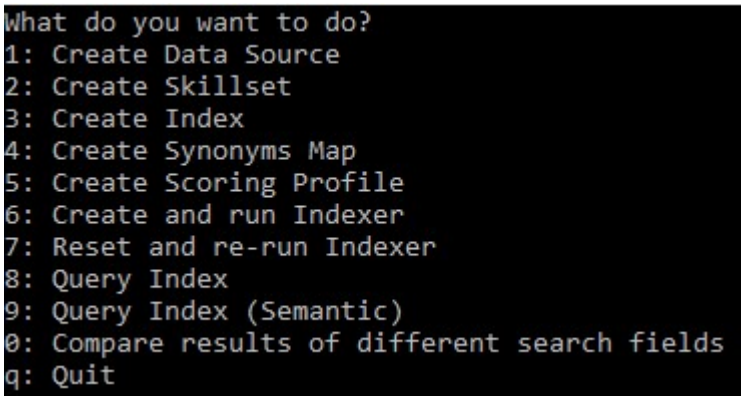

Figura 1. Menu da aplicação em consola

4. Resultados

Das habilidades cognitivas disponíveis pela ferramenta "Azure Cognitive Search" a maioria das habilidades cognitivas não se adequam à seleção de artigos científicos, tendo sido apenas identificada uma habilidade com potencial: "key phrases". Esta habilidade foi então alvo de estudo neste trabalho para perceber os benefícios que possa ter na pesquisa de conceitos em artigos científicos.

4.1. Key Phrases

Como referido anteriormente este trabalho focou-se na comparação entre pesquisas no campo conteúdo e pesquisas no campo "key phrases". Por forma a entender o que são "key phrases" foram extraídas e analisadas. Na tabela em baixo mostram-se alguns exemplos dos termos extraídos de um dos documentos estudados:

Tabela 1. Exemplos de "key frases" extraídas de um documento

Key Phrases			
MOOC student dropout prediction phenomenon	artificial intelligence	gap	TABLE II
high school dropout topic	course materials	I.	NLP
effective machine learning solutions	past year	time	Wang
higher education settings	Computer Science	Figure	Coleman
2018 IEEE Global Engineering Education Conference	Southeast Norway	ML application	29th
Norwegian University	learning experience	NTNU	A. Kurti
Institutional characteristics	MOOCs	January	Universidade Aberta

Em média cada foram extraídos 681 "key phrases" de cada documento. De notar que a habilidade cognitiva não cria palavras nem conceitos novos, ou seja, todas as frases/palavras encontram-se no documento. Os exemplos em cima foram escolhidos para tentar mostrar a diversidade de tipos de "key phrases", mas não representam a proporção encontrada no documento. Verifica-se então que muitas das "key phrases" são um resumo dos tópicos abordados, mas também se encontram muitas instituições, nomes de pessoas e abreviaturas. Também se encontram números de secção do documento. Verifica-se uma grande tendência de palavras que contêm letras maiúsculas. O algoritmo percorre todo o documento e inclui termos em rodapés, como é o caso da palavra-chave "Universidade Aberta".

4.2. Análise preliminar a um conjunto de dados reduzido

Os primeiros resultados estudados utilizando a aplicação em consola foram obtidos de 12 documentos inicialmente fornecidos. O estudo focou-se na comparação entre pesquisas no campo conteúdo e pesquisas no campo "key phrases". Estes resultados preliminares demonstraram melhorias na utilização de "key phrases" na seleção de documentos. Dos 12 documentos 2 foram excluídos na pesquisa em "key phrases" que não eram relevantes, e o documento mais relevante (da análise manual) passou de 3º no ranking de relevância da pesquisa em conteúdo para 1º na pesquisa em "key phrases".
Esta análise inicial permitiu, no entanto, perceber algumas limitações que também foram verificadas na segunda fase de análise de resultados, apontando-se em baixo as principais:

- A utilização de abreviaturas no mapa de sinónimos (IA e ML) podem influenciar a pontuação de forma inapropriada caso estas abreviaturas sejam utilizadas nos artigos com outro propósito. Um dos artigos obteve uma pontuação elevada pois continha o termo ML extensivamente, mas como abreviatura para "Modern Language" ao invés de "Machine Learning". Este artigo não foi selecionado quando o campo de pesquisa era sobre apenas "key phrases"
- Muitos dos documentos contêm os termos pesquisados em secções não relacionadas com o conteúdo como referências bibliográficas e títulos e rodapés. A título de exemplo é comum o termo "university" ser encontrado porque o nome da universidade que publicou o documento está incluído no documento (em títulos e rodapés).
- Como era expectável o algoritmo também não tem como identificar se os documentos são efetivamente artigos científicos, e, portanto, essa seleção tem que ser feita previamente.

4.3. Análise pesquisa em conteúdo vs "key frases" ao total de documentos

Numa segunda fase foram analisados os resultados obtidos por 798 documentos. A análise dos mesmos não permitiu concluir que existe uma vantagem clara em utilizar o campo de pesquisa "key phrases" na pesquisa de conceitos.

Comparados para ambas as pesquisas os gráficos de pontuação de relevância normalizada (pela pontuação de relevância mais elevada) vs ranking, verifica-se uma semelhança clara entre as duas.

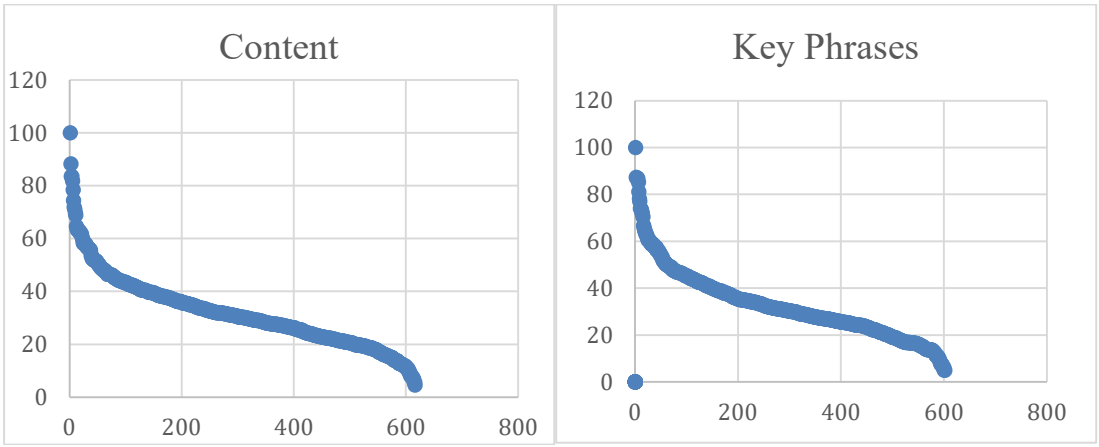

Figura 2. Gráficos Pontuação de relevância vs ranking para pesquisa em conteúdo (esquerda) e pesquisa em "key frases" (direita)

Porém a pesquisa em "key phrases" provocou alterações significativas a nível da relevância relativa de forma significativa, i.e., alterou os rankings entre as pesquisas (50% de correspondência de documentos no top 30 e 70% no top 100). No entanto após uma análise manual de uma seleção de artigos não foi possível verificar uma melhoria qualitativa significativa na alteração de ranking dos documentos. De notar que a análise manual foi efetuada a um número limitado de documentos e idealmente uma análise a todos os documentos seria necessária para retirar melhores conclusões.

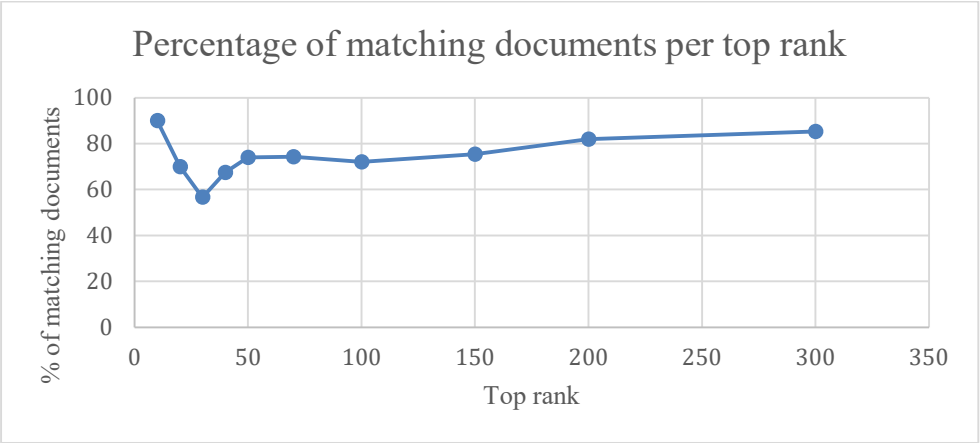

Figura 3. Percentagem de documentos coincidentes em ambas as pesquisas por grupo de documentos com maior relevância ("top rank")

A análise manual foi efetuada aos documentos pertencentes aos top 10 de ambas as pesquisas, aos 12 documentos excluídos pela pesquisa em "key phrases", a 24 documentos com uma redução significativa de relevância em pesquisa em "key phrases" e a 28 documentos que tiveram um aumento significativo de relevância em pesquisa em "key phrases".

Os documentos analisados foram classificados de acordo com a sua temática como:

1. Sobre IA na educação, mas não específico a ensino universitário:
2. Não são sobre IA;
3. Sobre IA na universidade, mas não relacionado com ensinar:
4. Apenas sobre IA, não sendo sobre educação nem universidades:
5. Relevante:
6. Artigos não específicos sobre os temas, mas incluem-nos:

Sendo as categorias 2,3 e 4 consideradas irrelevantes. A categoria 1 não é considerada irrelevantes pois o termo de pesquisa "university", apesar de não fazer parte do conteúdo, continua a encontrar-se presente em referências bibliográficas, etc, i.e., todos os termos relevantes à pesquisa encontram-se nos documentos.

Da análise do top 10 de ambas as pesquisas verificou-se que ambos partilham 9 documentos e a maioria são relevantes. A pesquisa em "key phrases" elimina 1 documento da categoria 1, no entanto não elimina 1 documento da categoria 4. que é irrelevante.

Tabela 2. Ranking dos documentos do top 10 em ambas as pesquisas

Ranking dos documentos da pesquisa em conteúdo.	Posição do documento no ranking da pesquisa em "key phrases"
1	1
2	5
3	3
4	4
5	9
6	6
7	2
8	"Eliminado"
9	7
10	8

Da análise dos artigos "eliminados", i.e, não devolvidos pela pesquisa em "key phrases" apenas 12 num total de 616 devolvidos pela pesquisa em conteúdo foram eliminados. Dos 12 apenas 4 eram irrelevantes sendo os restantes 8 do tipo 1.

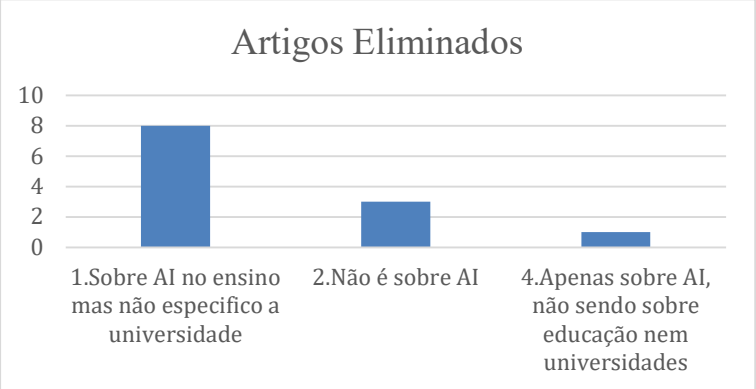

Figura 4. Artigos eliminados por pesquisa em "key frases"

Da análise dos artigos com relevância reduzida e aumentada pela pesquisa em "key phrases" foram identificados:

- 4 documentos relevantes com relevância aumentada face a 2 cuja relevância foi reduzida;
- Em relação a documentos considerados irrelevantes existiram 18 que tiveram a sua relevância aumentada e 14 que tiveram a sua relevância reduzida. No entanto 5 destes documentos são da categoria 2, e este aumento de relevância deveu-se à ambiguidade criada pelos termos abreviados do mapa de sinónimos, como por exemplo o termo ML que se encontra extensamente em um documento, mas com o significado "Mathematical Literacy";
- Em relação a documentos da categoria 1 existiram 4 documentos com relevância reduzida e também 4 documentos com relevância aumentada.

Tabela 3. Analise aos documentos com relevância reduzida em pesquisa "key frases"

Categoria	Qtd	%	Média Relevância Normalizada Pes. Cont	Média Relevância Normalizada Pes. Key Ph	Diferença de Relevância
1	4	16.7	51.8	37.4	14.4
2	5	20.8	37.1	26.3	10.8
3	4	16.7	41.5	26.3	15.3
4	5	20.8	46.4	32.3	14.1
5	2	8.3	48.3	35.7	12.6
6	4	16.7	44.3	29.5	14.7

Tabela 4. Analise aos documentos com relevância aumentada em pesquisa "key frases"

Categoria	Qtd	%	Média Relevância Normalizada Pes. Cont	Média Relevância Normalizada Pes. Key Ph	Diferença de Relevância
1	4	14.3	37.1	48.5	11.4
2	5	17.9	41.6	61.1	19.5
3	3	10.7	37.0	50.5	13.5
4	10	35.7	22.8	37.5	14.7
5	4	14.3	36.8	54.1	17.3
6	2	7.1	35.9	47.1	11.2

Como se pode verificar a única indicação de uma melhoria de resultados é num maior número de documentos relevantes com a relevância aumentada do que reduzida. Contudo essa diferença não é significativa. Assim não é possível concluir que a pesquisa em "key phrases" seja uma mais-valia.

Verificou-se também que alguns termos de pesquisa deste trabalho levaram a resultados indesejados:

- O termo "university" está contido na grande maioria dos documentos como em referências bibliográficas, ou nome da universidade do autor, etc, não sendo, no entanto, parte do conteúdo do documento;
- O termo learning encontra-se em conceitos de IA como "deep learning", existindo assim documentos que não estão relacionados com ensino que são devolvidos e com relevância significativa;

- Os termos abreviados no mapa de sinónimos podem estar contidos nos documentos, mas com outro significado. Isto teve um impacto mais significativo na pesquisa em “key phrases”.

5. Conclusões finais e trabalho futuro

Dado o tipo de tema dos documentos e a frase de pesquisa deste trabalho não foi observada uma vantagem clara e significativa na utilização da habilidade cognitiva “key phrases” face a uma pesquisa em conteúdo.

Era intenção deste trabalho utilizar o serviço de pesquisa semântica, sendo a aplicação preparada para ser utilizada com este serviço. Todavia tal não foi possível face à indisponibilidade do serviço de forma gratuita por parte da Microsoft à data do trabalho. Também existia a intenção de ter os resultados obtidos pela revisão sistemática manual do aluno de mestrado por forma a poder comparar os resultados e mais objetivamente responder aos objetivos propostos inicialmente. No entanto estes resultados manuais não foram obtidos durante o tempo de execução deste trabalho.

Sugere-se assim como trabalho futuro um estudo da pesquisa semântica bem com a utilização de uma revisão manual como base de comparação. Tendo resultados manuais, os parâmetros de pesquisa poderiam também ser alterados por forma a ajustar aos resultados finais. Este foi o propósito de adicionar o perfil de pontuação que aumenta a relevância do campo de pesquisa “key phrases” quando combinado com o campo “conteúdo”. Obviamente que estes parâmetros deveriam posteriormente ser testados em pesquisas sobre outros conceitos por forma reduzir o “over-fitting”.

A pesquisa semântica também tem a capacidade de responder a “queries” sob a forma de pergunta. Seria também interessante colocar as diretamente as perguntas definidas na introdução como “queries” e avaliar as respostas devolvidas pelo algoritmo.
Sugere-se também proceder à pesquisa retirando os termos que foram identificados como ambíguos e que influenciaram a pesquisa em “key phrases” de forma negativa: abreviaturas, o termo “university” e o termo “learning”.

Referências

Heidi Steen (a) (2021). Pesquisa semântica em Pesquisa Cognitiva de Azure. Obtido da web a 30 de Abril de 2021 em https://docs.microsoft.com/pt-pt/azure/search/semantic-search-overview

Heidi Steen (b) (2021). Criação de consultas na Pesquisa Cognitiva Azure. Obtido da web a 30 de Abril de 2021 em https://docs.microsoft.com/pt-pt/azure/search/search-query-create

Heidi Steen (c) (2021). Sinónimos em Pesquisa Cognitiva Azure. Obtido da web a 20 de Abril de 2021 em https://docs.microsoft.com/pt-pt/azure/search/search-synonyms

Heidi Steen, Sunny Deng, Agustin Schiavon (2021). Criação de índices de pesquisa na Pesquisa Cognitiva Azure. Obtido da web a 20 de Abril de 2021 em https://docs.microsoft.com/pt-pt/azure/search/search-what-is-an-index

Heidi Steen, Sunny Deng, Santiago Cortez (2021). Indexadores na Pesquisa Cognitiva do Azure. Obtido da web a 24 de Abril de 2021 em https://docs.microsoft.com/pt-pt/azure/search/search-indexer-overview

Microsoft (2021). Enriquecimento IA em Pesquisa Cognitiva Azure. Obtido da web a 24 de Abril de 2021 em https://docs.microsoft.com/pt-pt/azure/architecture/solution-ideas/articles/cognitive-search-with-skillsets

Puneet Hariharan (2021). Configure o algoritmo de classificação de semelhança na Pesquisa Cognitiva Azure. Obtido da web a 17 de Abril de 2021 em https://docs.microsoft.com/pt-pt/azure/architecture/solution-ideas/articles/cognitive-search-with-skillsets

Tamra Myers (2021). Armazenamento de BLOBS no Azure. Obtido da web a 15 de Abril de 2021, em https://docs.microsoft.com/pt-pt/azure/storage/blobs/storage-blobs-introduction

José Diogo, Engenheiro de Projeto Sénior na área de instalação de equipamentos "subsea". Obteve o grau Mestre em Engenharia Mecânica pelo Instituto Superior Técnico m 2012 e licenciou-se em Engenharia Informática pela Universidade Aberta em 2021.

Henrique S. Mamede, Professor Auxiliar com Agregação no Departamento de Ciências e Tecnologia (DCeT) da Universidade Aberta e investigador sénior no INESC TEC. Licenciado em Engenharia Informática pela COCITE, Mestre em Informática pela Universidade de Lisboa, Doutor em Sistemas e Tecnologias de Informação pela Universidade do Minho e Agregado em Ciência e Tecnologia Web pela Universidade de Trás-os-Montes e Alto Douro.

(esta página par está propositadamente em branco)