Ciência da Computação / Sistemas de Informação

Exploração de técnicas de processamento de linguagem natural para identificação de tópicos

Harisson de Carvalho Alvarenga - 4º módulo de Ciência da Computação, UFLA, bolsista PIBIC/CNPq

Marluce Rodrigues Pereira - Orientador DCA, UFLA. - Orientador(a)

Paula C. F. Cardoso - Coorientador Departamento de Computação, UFPA

Resumo

O uso de computadores por deficientes visuais demanda melhorias de acessibilidade que pesquisadores buscam encontrar soluções para esse desafio. Os leitores de tela representam um avanço, pois permitem que os usuários obtenham informações do computador por meio de feedback sonoro, possibilitando a leitura de textos. Enquanto leitores sem deficiência visual utilizam técnicas de skimming para localizar rapidamente os assuntos de um texto, deficientes visuais precisam ouvi-lo inteiro para identificar o tema. Técnicas de identificação automática de tópicos podem amenizar esse problema, diminuindo o tempo necessário para que deficientes visuais encontrem a informação de seu interesse. Este trabalho investigou técnicas de segmentação e rotulação tópica automática de textos, com o intuito de criar cabeçalhos mais informativos para leitores de tela, facilitando assim a compreensão dos assuntos pelos deficientes visuais de maneira mais rápida. Para isso, foram utilizadas técnicas de processamento de linguagem natural e word embeddings. Para segmentar, duas abordagens foram testadas. A primeira baseou-se na frequência de palavras, utilizando a técnica de bag of words para encontrar similaridades entre blocos de texto. A segunda converteu o texto em embeddings de palavras, utilizando o modelo word2vec CBOW de 50 dimensões do repositório NILC. A distância do cosseno entre os embedding médios de cada frase foi calculada para realizar a segmentação. Na etapa de rotulação, utilizou-se o modelo keyBERT para extrair palavras-chave representativas para serem cabeçalhos de cada segmento topical. Para avaliar a segmentação, utilizou-se o córpus CSTNews, que possui 50 conjuntos de textos jornalísticos segmentados manualmente em tópicos. A medida windowdiff foi utilizada para comparar a segmentação automática com as do corpus. A medida atribui valores próximos de 0 para boas segmentações e valores próximos de 1 para segmentações inadequadas. Os resultados mostraram que a primeira abordagem apresentou problemas significativos, com a maioria das distâncias calculadas próximas de 0, indicando excesso de segmentos. A segunda abordagem, suavizou esse problema criando segmentos mais extensos, mas ainda apresentou baixa precisão. Assim, os resultados foram pouco satisfatórios, com valores próximos a 1. Para trabalhos futuros, planeja-se explorar outras abordagens de processamento de linguagem natural para aprimorar a etapa de segmentação.

Palavras-Chave: Acessibilidade, Segmentação tópica, PLN.
Instituição de Fomento: CNPq

Link do pitch: https://youtu.be/I5Az9blGU0g