UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO
Departamento de Ciências de Computação

Disciplina SCC0275 - Ciências de Dados

Estudo sobre a relação da estrutura de um texto e a performance de digitação

docente

Profa. Roseli Aparecida Francelin Romero

Grupo 14

alunos

Rodrigo Anes Sena de Araújo - 9763064

Rodrigo Noventa Júnior - 9791243

I - Introdução

O grupo decidiu realizar o trabalho sobre uma base de dados de estatísticas de digitação, extraída do site https://www.keyhero.com/. A motivação deste estudo se dá pelo fato de que é cada vez mais importante para um profissional, atualmente, ser rápido e ter acurácia em sua digitação, melhorando seu desempenho e resultados. Além disso, tal estudo pode ser útil para diversos sites de digitação como o citado acima para classificar textos por grau de dificuldade, criar um sistema de sugestão de textos para perfis diferentes de usuários e construir um plano de treinamento em digitação com aumento gradativo de dificuldade. Algumas informações que o grupo quis estudar foram a relação entre a estrutura de um texto e o impacto no desempenho médio de sua digitação. Saber dessas informações pode ser útil para estimar o tempo de digitação de diversos textos. Saber como os diversos tipos de idiomas influenciam na velocidade de digitação e quais elementos gramaticais têm mais influência na digitação.

Foram usadas diversas técnicas para estudar os dados extraídos. Foram feitas plotagens diversas, como o box-plot, strip-plot, gráficos de distribuição e a estimativa de densidade kernel. Foram analisadas as matrizes de correlação de Pearson e Phik para os dados. Além da técnica de regressão linear com o intuito de mais análises e a criação de um modelo de predição da influência de um tipo de texto na sua respectiva velocidade e acurácia de digitação.

A seção II (Trabalhos relacionados) contextualiza este projeto em relação a outros projetos prévios relacionados com o tema, para embasar a motivação e interesse do trabalho.

A seção III (Material e métodos) detalha mais a etapa prática do projeto, desde a apresentação do dataset, a extração e pré processamento dos dados e o modelo utilizado para a predição.

A seção IV (Experimentos) contém os resultados do modelo de predição, uma tabela para estudo dos resultados, assim como comentários e conclusões a respeito.

A seção V (Conclusão) finaliza o artigo. Nela, serão resumidas as principais conclusões obtidas deste trabalho.

II - Trabalhos relacionados

O trabalho relacionado a velocidade de digitação no qual nos baseamos foi um artigo do site Medium, escrito pelo usuário Niraj Pandkar [1]. O autor analisa suas próprias estatísticas de digitação retiradas de sua conta no site Typeracer. Com uma base de mais 4000 estatísticas de corridas em um espaço de tempo de 3 anos. Ele analisou a influência do tamanho de um texto, a quantidade de pontuações e o número de letras maiúsculas na sua respectiva velocidade de digitação.

O autor também desenvolveu um sistema de predição de dificuldade para textos que extraia o tamanho do texto, quantidade de pontuações, quantidade de letras maiúsculas e um vetor de frequência das palavras e usava essas informações como entrada para uma regressão logística que classificava o texto em difícil, médio e fácil. A acurácia do classificador (49.8%) foi um pouco superior à acurácia de uma predição aleatória (33.3%).

O segundo artigo relacionado a este trabalho é Typing is writing: Linguistic properties modulate typing execution (2016) [2]. Esse artigo analisa padrões léxicos e suas influências na velocidade e acurácia de digitação. Esse estudo envolveu digitadores experientes em um ambiente de testes de digitação a partir de áudios. Foram medidas acurácia, latência de resposta, entre ouvir uma palavra e começar a digitá-la, bem como a velocidade de digitação. Algumas métricas analisadas foram a frequência de vezes que duas letras iguais aparecem próximas ou em sequência em uma palavra (frequência de bigramas) e a frequência de repetição de palavras em um texto. Foi concluído que repetições de palavras em um texto aumentam a velocidade de digitação por permitir um resgate de memória recente para o digitador. Também foi constatada uma correlação forte entre a velocidade de um digitador em digitar bigramas e sua performance como digitador.

III - Material e métodos

A base de dados principal contém um total de 600 quotes divididos nos 10 idiomas mais utilizados na plataforma (60 para cada). Além dos quotes, ela contém alguns dados estatísticos que foram obtidos durante a exploração, como a média de velocidade de digitação (WPM) e a acurácia apresentada neles, esses dados se referem aos 10 melhores “jogos” e também aos 10 mais recentes.

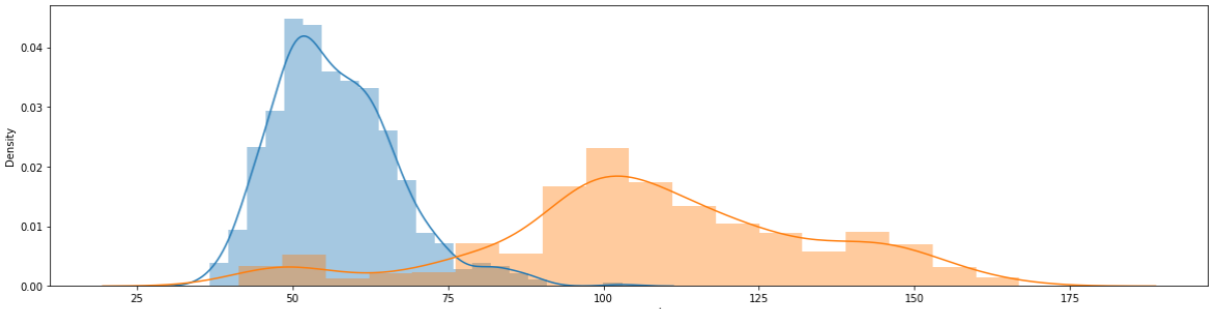

Figura 1: Distplots mostrando a densidade de WPM (palavras digitadas por minuto) para jogadores comuns (em azul) e para jogadores de alta performance (em laranja)

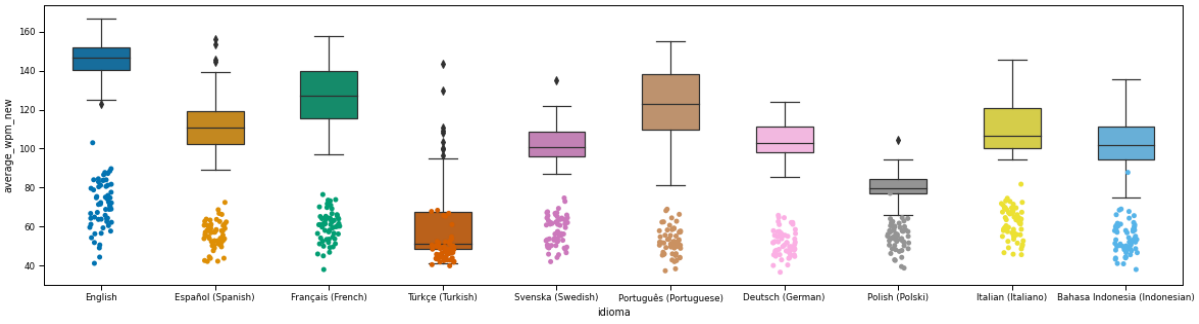

Figura 2: Boxplots mostrando os melhores jogos para cada idioma e stripplot para os mais recentes

Outros dados importantes são referentes aos quotes, são eles: Tamanho do quote, tamanho médio das palavras, quantidade de letras maiúsculas, caracteres não alfanuméricos e acentos. Todos esses dados foram obtidos após o scraping.

Como a maior parte dos dados são obtidos processando o quote e os dados obtidos via scraping estavam bem estruturados no site, não foi necessário tratar dados ausentes,

porém optamos por utilizar os dados obtidos pelo processamento como porcentagem ao invés do valor absoluto, com exceção do tamanho do quote e tamanho médio das palavras. Pelo fato de não termos uma grande quantidade de atributos, não foi utilizada nenhuma técnica de seleção, porém foram analisados as matrizes de correlação de Pearson e Phik (φk) para ter uma noção de como os atributos se relacionam.

Pela matriz de Pearson, foi possível notar que o tamanho médio da palavra e a quantidade de acentos eram os fatores que mais influenciavam a velocidade de digitação nos melhores jogos, seguido pelo tamanho do texto, esses três fatores também eram os mais influentes na acurácia (também se tratando dos melhores jogos), porém não tanto quanto na velocidade. Os jogos mais recentes possuíam um comportamento semelhante, porém de maneira bem mais fraca. A matriz Phik também foi utilizada, pois mostra a relação entre dados categóricos e dados quantitativos, permitindo a análise do idioma.

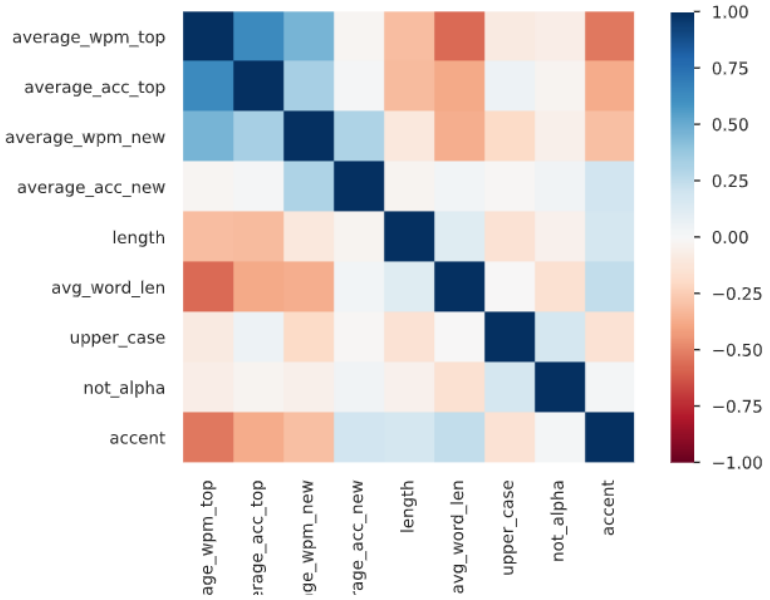

Figura 3: Matriz de correlação de Pearson

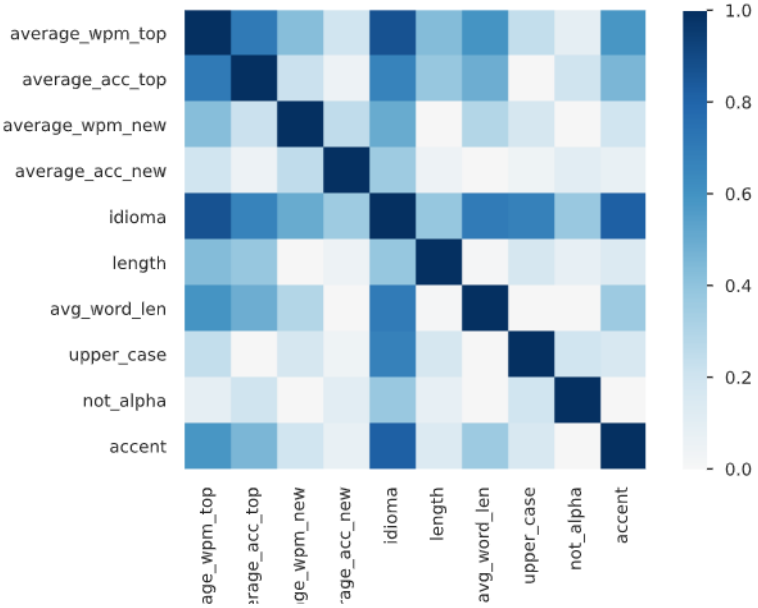

Figura 4: Matriz de correlação Phik

Os modelos de regressão utilizados foram: a regressão linear e polinomial, de grau 2 até grau 5, para todas as regressões, foram utilizados 2 conjuntos de atributos diferentes, sendo o primeiro composto por todos os atributos extraídos do quote (Tamanho, quantidade de acentos, etc) e o segundo composto somente por aqueles que demonstraram uma maior influência durante a análise das matrizes de correlação (Tamanho médio de palavras e quantidade de acentos).

IV - Experimentos

Como dito anteriormente, foram testados 10 modelos de regressão, divididos em 2 conjuntos de atributos. A tabela a seguir mostra os resultados utilizando o erro quadrado médio e também o R2 score.

Tipo de regressão	Atributos	Erro médio quadrado	R2 score
Linear	Todos (5)	308,54	55%
Polinomial (grau 2)	Todos (5)	284,73	58,5%
Polinomial (grau 3)	Todos (5)	263,66	61,5%
Polinomial (grau 4)	Todos (5)	258,55	62,3%
Polinomial (grau 5)	Todos (5)	249,77	63,6%
Linear	Mais significativos (2)	359,10	47,6%
Polinomial (grau 2)	Mais significativos (2)	345,36	49,6%
Polinomial (grau 3)	Mais significativos (2)	339,90	50,4%
Polinomial (grau 4)	Mais significativos (2)	335,26	51,1%
Polinomial (grau 5)	Mais significativos (2)	316,54	53,8%

Tabela 1: Resultados das regressões

Pela tabela, é possível perceber que uma regressão polinomial de grau 5 utilizando todos os atributos extraídos do texto teve o melhor desempenho enquanto as regressões que utilizaram apenas os atributos mais significativos não obtiveram um desempenho satisfatório.

V - Conclusões

Com base nos estudos feitos, foi observado como a distribuição de WPM de digitadores de alta performance possui uma variância muito maior do que a distribuição de digitadores comuns, o que demonstra o quão grande o aumento do WPM pode ser para quem pratica digitação corretamente e regularmente.

Foi observada a influência forte do idioma na velocidade de digitação. Sendo o inglês o idioma com melhores digitadores. Enquanto que idiomas menos utilizados e com grande quantidade de caracteres diacríticos como o turco demonstraram ter os piores desempenhos.

Foi concluído que a frequência de acentos, o tamanho médio das palavras de um texto e o tamanho do texto têm relação inversa com o desempenho de sua digitação.

VI - Referências

1. Artigo do site Medium, escrito pelo usuário Niraj Pandkar. https://medium.com/@nirupanda.296/what-affects-your-typing-speed-and-how-to-improve-it-b8f1d2b622aa.
2. Pinet, S., Ziegler, J.C. & Alario, FX. Typing is writing: Linguistic properties modulate typing execution. Psychon Bull Rev 23, 1898–1906 (2016). https://doi.org/10.3758/s13423-016-1044-3