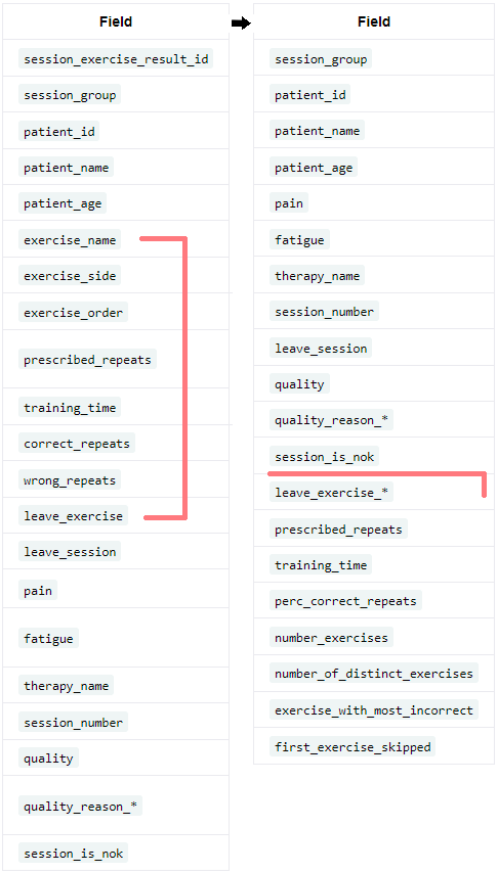
sword

Thursday, 25 April 2024

ML Engineer Test Notes

Este documento serve como apêndice à solução final, tentando providenciar contexto conciso e objetivo sobre o raciocínio, interpretação, exemplos e design choices em alguns dos pontos considerados importantes na minha resolução, deixando de fora detalhes relacionados com a minha exploração/estudo dos dados.

Q1



• Depois de me familiarizar com os dados, considerei central a distinção entre colunas dependentes de exercícios (marcadas à esquerda) e independentes (dependentes apenas a sessões e/ou pacientes).

O processo de transformação escolhido passa por:

1) criar e passar ao groupby().map() uma função personalizada capaz de lidar com possíveis transformações mais complexas que colunas novas necessitem (abaixo da linha);

2) concatenar o resultado anterior com a transformação agregada de colunas que se mantêm constantes (acima da linha).

• Contexto sobre a minha interpretação do significado de colunas específicas:

leave_exercise_**technical_issues** e leave_exercise_**difficulty** - Reparei que estes valores não existiam nos dados fornecidos (na coluna leave_exercise), ao contrário de outros. Inicialmente defini difficulty como exercícios saídos por “other” onde o rácio de correct e wrong repeats fosse baixo (tentar definir dificuldade sem entrar pelas colunas pain ou tired, que já têm o lugar delas) ou através da ease of use em casos de desistência. Para technical_issues inicialmente defini como exercícios saídos por “other” ligando a movement detection, tablet, tablet and motion. Acabei por assumir que estes valores poderiam existir noutros dados e simplificar. Gostaria de discutir este ponto.

perc_correct_repeats - Considerei o rácio correct/(correct+wrong) e não correct/prescribed_repeats por integridade semântica, por exemplo em casos de desistências onde não existe performance de n exercícios.

number_exercises(**performed** in the session) - Depois de estudar diferentes colunas (especificamente em situações complicadas através de sampling de leave_exercise/session rows) que poderiam ser úteis, decidi contar exercícios onde training_time>0. Importante também referir que keyword “performed” fez com que não contasse simplesmente o numero de rows da sessão.

number_distinct_exercises - Contando com a lógica acima para “performed” exercises, apenas adicionei a lógica de olhar para 2 exercícios “iguais” e contá-los uma única vez, com ajuda da coluna exercise_name (por exemplo, considerei 3 sets de push ups e 2 sets de abs, 2 exercícios distintos).

José Pedro Ferreira

exercise_with_most_incorrect - Tive o cuidado outra vez de agrupar os mesmos exercícios da sessão com a lógica acima antes de encontrar o exercício associado ao valor máximo de wrong_repeats.

first_exercise_skipped - Filtrando os exercícios da sessão com valores no campo leave_exercise, tive atenção às exercise_orders antes de localizar o nome do primeiro exercício.

Estas foram algumas das interpretações onde achei necessário dar algum contexto. Espero que este texto ajude a elucidar as minhas decisões. Num ambiente de equipa, tendo a fazer as perguntas necessárias.

(Features_expected.parquet dentro do projeto já estão corridas e transformadas com o meu transform())

Q2

- Reasoning/Considerations:

Na minha system message, escolhi focar-me na persona e objetivo do meu modelo, dando general guidelines baseadas no documento providenciado, e também inseri neste campo exemplos úteis.

Na minha user message, foquei-me em informação mais especifica como format guidelines que fora testadas com a minha própria API key, pois não queria estar a gastar dinheiro na fase de testes. (Nestes testes decidi que o modelo GPT3.5 era suficiente para esta task, sendo também um modelo com um preço muito mais barato por token). Também dentro da user message, após filtrar apenas colunas que achei importantes para o contexto da sessão (informação geral, informação de fatores técnicos e informação especifica), criei um pequeno texto descritivo com a informação total compilada, de modo a não apenas fornecer o dicionário da sessão em causa. Tentei usar alguns truques semânticos para lidar com certos valores que espero discutir.

Diria que alterar a temperatura é interessante desde que a mantenhamos baixa (a partir de 0.5 talvez perca qualidade). Acabei por escolher 0.2 depois de alguns testes. Importante referir que usei o argumento frequency_penalty de modo a penalizar word repetitions.

Alguns exemplos em casos onde houveram problemas técnicos na sessão:

```
Hey Kevin, great job on completing session number 9!
👏 It's awesome to see your commitment to therapy.
I noticed you had a couple of technical issues during specific exercises, but don't worry, we can work through those together.
Can you share more about what happened during those exercises? How are you feeling overall after the session?
```

```
Great job completing your session, Kathy Ali! 👏
It sounds like you had a bit of a challenging session today with some technical issues and higher pain and fatigue levels.
Remember, progress is not always linear. Let's work together to address these challenges.
How are you feeling overall after today's session?
```

```
Great job completing your session, James Jimenez Jr.! 👏
It's awesome to see you pushing through session number 45!
Your pain and fatigue levels were low at 2 each, which is fantastic.
Let's chat about the hiccups with movement detection. Can you share more about what happened there?
```

```
Hey Theresa, great job completing your session today! 👏
It sounds like you had a bit of trouble with the sitting neck side bending exercise, but that's totally normal for the first session.
I see you had some technical hiccups with movement detection on the tablet. Can you share more about what happened there?
How are you feeling after today's session?
```

(Vi vários outros emojis, apesar destes exemplos)

• Potential next steps:

Uma boa prática que na minha experiência ajuda bastante em prompt engineering é o uso de mais exemplos na minha system message. Quando comparados com o resto do contexto dado no system message e também general format guidelines inseridas no user message, exemplos criam como um molde que aumenta o rendimento do resto das técnicas usadas na prompt. O problema no limite é o tamanho da prompt em casos de vários exemplos, que leva a custos de uso maiores, respostas mais lentas e possivelmente a capacidade de generalização no caso de temperaturas baixas. Tudo isto obviamente requer trial and error, independentemente da nossa intuição.

Isto leva ao próximo possível next step, que seria dar fine-tune ao modelo que estejamos a usar com uma das mais recentes técnicas de baixo custo e alta qualidade, como LoRA ou QLoRA (versão de dupla quantização, ainda mais leve).

Ora, dar fine-tune a um model com estas técnicas em específico requer ter posse dos model weights, o que não é o caso para os modelos da OpenAI. Aqui chegamos ao último tópico que é usar outros modelos de qualidade localmente, como por exemplo o recente LLama 3.

Não é apenas um investimento que permite fazer fine-tuning da maneira proposta, visto que existem outras até mais simples através da OpenAI, mas sim um investimento em privacidade.

Penso que seria positivo manter os dados dos pacientes da Sword in-house (não assumindo que não o fazem, mas no contexto do challenge pode ter sido diferente) visto também que a informação tem o seu valor a longo prazo tirando a privacidade da equação.

Aumentar a qualidade das mensagens, baixar o custo e latência de requests (prompts aqui poderiam diminuir o tamanho) são exemplos de pros.

Uma ideia que tive também, que não passa pelos exemplos anteriores, é integrar no contexto dado para a criação das nossas mensagens informação de sessões anteriores do paciente, de modo a inserir também na mensagem algum contexto evolutivo, o que pode melhorar a interação visto que memória do paciente aumenta a sensação humana do nosso método, para além da mensagem simplesmente poder ficar mais rica.

Por último, (desta vez verdade), caso queiramos dar leverage da natureza contínua das mensagens de terapeutas sobre sessões no contexto de melhorar os exemplos da nossa prompt ou até focar apenas nas últimas mensagens especificamente de x paciente, podemos construir um sistema RAG onde o retrieval se foque em recent past sessions.