

Questão 1:

PROCESSAMENTO DE LINGUAGEM NATURAL é um conjunto de técnicas para a REPRESENTAÇÃO DA LINGUAGEM NATURAL. ELA TEM UMA IMPORTÂNCIA CENTRAL NO CONTEXTO DE ~~EM~~ MINERAÇÃO DE DADOS POIS GRANDE PARTE DO CONHECIMENTO ESTÁ DENTRO DE DOCUMENTOS ESTRUTURADOS, SEMI ESTRUTURADOS E NÃO ESTRUTURADOS. ADEMAIS, ELE POSSUI UM PAPEL FUNDAMENTAL DENTRO DE INTERFACE HUMANO ~~COM~~ COMPUTADOR, POIS TORNA A INTERAÇÃO COM A MÁQUINA INCLUSIVA, FACILITADO O ACESSO À INFORMAÇÃO UTILIZANDO A NOSSA PRÓPRIA LINGUAGEM.

DESSA MANEIRA, A REPRESENTAÇÃO DO TEXTO TORNA-SE UM PROBLEMA CENTRAL EM PROCESSAMENTO DE LINGUAGEM NATURAL. A FORMA DA REPRESENTAÇÃO DE DETERMINA COMO O TEXTO É REPRESENTADO, AFETANDO ~~DIRETAMENTE~~ DIRETAMENTE O DESEMPENHO, A INTERPRETABILIDADE E A SUA APLICAÇÃO. A REPRESENTAÇÃO PODERÁ SER SIMBÓLICA, ONDE O TEXTO ~~É~~ É REPRESENTADO ATRAVÉS DE SÍMBOLO COM REGRAS, OU REPRESENTAÇÃO VETORIAL QUE IRÁ REPRESENTÁ-LO EM UM ESPAÇO VETORIAL. NESSE CONTEXTO, EXISTEM ~~AS~~ AS CLASSES DE MODELOS BASEADOS EM CONTAGEM, PROBABILÍSTICOS E EMBEDDINGS.

~~Os modelos probabilísticos e embeddings são os mais utilizados para representar o texto e a sua aplicação.~~
~~Para as próximas vamos considerar que um documento é um conjunto de texto que ~~será~~ será modelado. No caso em busca e representação da informação, o objetivo é encontrar quais documentos que melhor ~~representa~~~~

Representam a partir de uma consulta textual. No caso da mineração de texto, gostaria que esse documentos sejam classificados e entre outros exemplos. Dessa maneira, a forma de representação mais ~~simples~~ simples é o Bag-of-words, onde cada documento é representado pela coleção de palavras que ele possui. Grande problema dessa representação, é que mesmo tirando algumas palavras sem significado aparente (exemplo stopwords), a palavra e o texto possui um significado que é alterado pelo contexto. Além disso, a frequência da palavra pode possuir uma importância dentro do texto. ~~Por~~ Isso sem contar os problemas de sinônimo ou stemming que são processos comuns ~~na~~ de pré-processamento desse tipo de representação. Para amenizar o problema da frequência, surgiu o TF-IDF que dá um peso inverso a ~~freq~~ frequência da palavra na ~~rep~~ representação. Os modelos de contagem foram precursores da representação textual. Contudo, a sua utilização atual acaba sendo bem restrita por causa da sua falta de representação da semântica do texto, recorrendo a estrutura léxica do texto. Contudo em contexto em que é necessário a presença de palavras, essa representação acaba sendo usada.

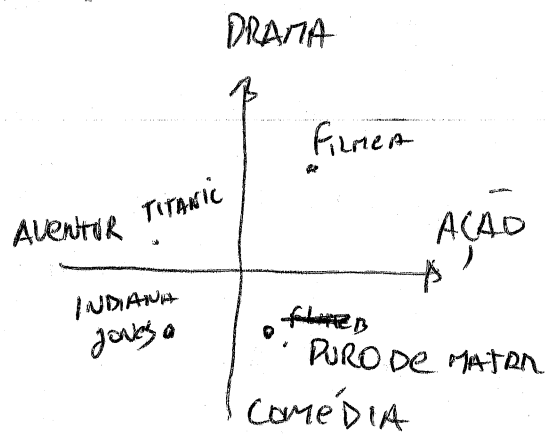
Por causa dos problemas da representação ~~na~~ baseada em contagem, surgem os modelos probabilísticos, principalmente representados pelo modelo de n-grams. Nessa

Questão 1 (cont):

Representação é considerado que a palavra possui um contexto de utilização, ou seja, as palavras antecessoras, (ou até sucessoras) complementam o significado e a presença dela. O objetivo do n-grams é modelar o texto em graus de ocorrência e o tamanho do grão é o tamanho da vizinhança diferentemente da representação de contagens que considera a palavra individual, nessa forma vou considerar o conjunto de palavra, transformando ~~o problema de~~ o problema de ocorrência em um problema de probabilidade condicional. Ainda é necessário, dependendo do contexto, os mesmos tratamentos de pré-processamento para ~~em~~ eliminação de stop-words, sinônimos e radicais. Contudo, esse modelo começa trazer um significado contextual a representação. Esses modelos foram muito utilizados na criação de buscadores ou até problemas de mineração de texto como sumarização, análise de tópicos, o seu grande problema é não considerar a semântica da palavra, ou seja, ele considera as palavras como ocorrência morfológicas. Contudo, eles possuem mais do que um símbolo, um exemplo clássico dos livros são as palavras

REI, RAINHA, HOMEM e MULHER. ~~SA~~ Eles possuem significados relacionados em que esse modelo não representa.

Dessa maneira, na última década, surgiram os embeddings. Para tal entendimento, é necessário entender o conceito de variáveis latentes. Variáveis latentes são variáveis que não são observadas, mas ~~em um~~ é possível um modelo inferi-las. Um exemplo clássico de variável latente é na área de sistemas de recomendação onde podemos imaginar variáveis para cada gênero de um filme e ~~para~~ representar os filmes nesse espaço dimensional. Como mostra o exemplo abaixo.



Se ~~em~~ conseguirmos extrair features a partir de um modelo e representar esses dados em um espaço n dimensional, ~~em seguida aplicar modelos~~ inferido.

Essa é a ideia dos embeddings, representar as palavras nesse espaço inferido n dimensional, onde ~~cada palavra~~ esse espaço possua um significado semântico das palavras.

Dessa maneira, no exemplo dado como rei, homem

Questão 1 (cont):

MULHER e RAINHA ESTARIAM ESPACIALMENTE PRÓXIMOS.

Uma característica importante que as palavras nessa representação serão considerados vetores e as operações vetoriais possuem significado. Dessa maneira, através de aritmética vetorial é possível fazer deslocamento onde se preserva a semântica, algo que não era possível com as representações os modelos bag-of-words, TFIDF ou até com n-grams. Esse tipo de representação alterou bastante a fez e aumentou o desempenho de várias tarefas, e revolucionou a área.

Hoje em dia, os embeddings são muito utilizados. A sua maior dificuldade é obter os valores, mas existem diversas técnicas disponíveis para utilizarem. Diferentemente das outras técnicas é necessário treinar um modelo com um corpus relevante para obtermos os embeddings. Inclusive, para áreas que possuem termos específicos de domínio como medicina e jurídico. A utilização desse tipo de modelo pode se tornar mais difícil, pois será necessário obter ou embeddings treinados para esse corpus ou precisar criá-los. É claro que é possível utilizar transfer learning com um corpus menor, minimizando esse problema.

PARA CONCLUIR, A REPRESENTAÇÃO TEXTUAL É DE SUMA IMPORTÂNCIA PARA DIVERSAS TAREFAS E SUA ESCOLHA DEPENDERÁ MUITO DA TAREFA EM SI, INCLUSIVE ATÉ O PRÉ-PROCESSAMENTO POSSUI CONSEQUÊNCIA, COMO EXEMPLO É COMUM DEIXAR STOP-WORDS EM PROBLEMA DE ANÁLISE DE SENTIMENTOS. POR FIM, OS EMBEDDINGS SÃO MUITO UTILIZADOS E FORAM BASE PARA A ARQUITETURA TRANSFORMERS QUE SÃO UTILIZADAS EM MODELOS DE LINGUAGEM GRANDE QUE VÊM A CADA DIA QUEBRANDO RECORDES DE BENCHMARKS EM QUASE TODAS TAREFAS TEXTUAIS.



Questão 2:

NAS ÚLTIMAS DÉCADAS, COM A EVOLUÇÃO DAS TECNOLOGIAS DE INFORMAÇÃO, ESTAMOS GERANDO UMA GRANDE QUANTIDADE DE DADOS. AINDA MAIS COM A WEB 2.0, USUÁRIOS DA INTERNET SE TORNARAM O ATOR PRINCIPAL NA GERAÇÃO DE DADOS PRINCIPALMENTE EM SISTEMAS COMO AS REDES SOCIAIS. ESSA GRANDEZA DESSA ESCALA FEZ A MUDANÇA DE DIVERSOS PARADIGMAS COMPUTACIONAIS PARA O PROCESSAMENTO DESSA ESCALA COM O OBJETIVO PRINCIPAL EM UTILIZAR ESSES DADOS PARA TOMADA DE DECISÃO AUTOMATIZADA.

~~ESSA TOMADA DE DECISÃO~~

COM ESSA GRANDE QUANTIDADE DE DADOS FOI POSSÍVEL CRIAR MODELOS QUE O SEU PODER DE DECISÃO É CAPAZ DE SER UM DIFERENCIAL DE NEGÓCIO ENTRE AS EMPRESAS E INFLUENCIAR A SOCIEDADE. ~~COMO~~ ~~EXEMPLO~~ SISTEMAS DE RECOMENDAÇÃO É UM EXEMPLO DESSE IMPACTO, ~~QUE~~ ~~PASSAM~~ ~~PROBLEMAS~~ ~~QUE~~ ~~TEM~~ ~~COMO~~ ~~OBJETIVO~~ MINIMIZAR O PROBLEMA DA ESCOLHA EM UMA GRANDE QUANTIDADE DE OPÇÕES. ESSAS ELAS PASSAM A TOMAR DECISÕES PARA O USUÁRIO DECIDINDO O QUE ELAS VÃO CONSUMIR OU QUAIS ITENS SERÃO CONSUMIDOS. NESSE CONTEXTO, SURTEM AS LEIS DE USO DE RESPONSABILIDADE DE DADOS E/OU TOMADA DE DECISÃO COMO NA EUROPA O GDPR ~~OU~~ E NO BRASIL A LGPD.

DESSA FORMA, CONSIDERANDO O DESENVOLVIMENTO DE UM SISTEMA DE RECOMENDAÇÃO EM UM AMBIENTE EM QUE FOI DETECTADO FAVORITISMO DE ALGUNS GRUPOS SOCIAIS, REPRODUTIBILIDADE DOS RESULTADOS E EXPLICABILIDADE LIMITADA, PODE-SE DISCUTIR NESSE CENÁRIO ABORDAGENS PARA MITIGAR ESSES PROBLEMAS.

PRIMEIRO É CONSIDERAR QUE NESSE SISTEMA SERÁ USADO TANTO POR EUROPEUS QUANTO BRASILEIROS. UM PONTO PRINCIPAL É DEFINIR COMO SERÁ A GOVERNANÇA ÉTICA PARA ESSE CENÁRIO. ~~NA~~ A IDEIA QUE O SISTEMA NÃO CUMpra AS NORMAS LEGISLATIVA, E SIM QUE O SISTEMA ELE SEJA CONSTRUÍDO EM JOINT DESSA GOVERNANÇA, NA LITERATURA, ESSE CONCEITO SE CHAMA ÉTICA BY DESIGN.

~~UMA~~ UM PONTO SENSÍVEL DENTRO DE UM SISTEMA DE RECOMEN-
DAÇÃO É O QUE É CHAMADO DE FAIRNESS. ESSE TÓPICO FOI MUITO ABORDADO PELA COMUNIDADE DA ÁREA NA FINAL DA ÚLTIMA DÉCADA. ESSE PROBLEMA PODE QUE O ALGORITMO DE RECOMENDAÇÃO FAVOREÇA UM GRUPO SOCIAL. NESSE CONTEXTO, ESSE FENÔMENO PODE VARIAR A CARACTERIZAÇÃO DEPENDENDO DO DOMÍNIO DO SISTEMA. NO CASO DE MÚSICA, O SISTEMA PODE FAVORECER UM GRUPO ESPECÍFICO DE BANDAS, NÃO DANDO CHANCE PARA BANDAS NOVAS DE SE RECONHECIAM. ESSE PROBLEMA TAMBÉM É CONHECIDO COMO VÍCIOS DE POPULARIDADE. NO CASO DE UM SISTEMA PEOPLE-TO-PEOPLE, COMO O DE EMPREGO, O SISTEMA PODE FAVORECER UMA ÉTNIAS.

Questão 2 (cont):

Os métodos mais comuns para mitigação são mecanismos para incorporar no pipeline da construção do algoritmo. Normalmente podendo ser no pré-processamento dos dados, equilibrando os dados para um treinamento equiparado. Outra possibilidade é no pós-processamento em que é possível mudar a classificação e o ranqueamento ajustando para alguma métrica de injustiça. Por fim, é possível modelar o problema no próprio algoritmo fazendo ele aprender a não cometer injustiça.

Outro problema comum, é a questão da reprodutibilidade. Na engenharia de software o artefato principal, em muitos casos, é o código em si, normalmente possuindo características determinísticas dado a ~~entrada~~ sua entrada. Com a vinda da utilização de modelos e ciência de dados não só como suporte mas também como parte no sistema, esse cenário mudou. Existe agora mais uma dimensão que são os dados, incluindo a engenharia de dados ~~na~~ ~~no~~ ~~processo~~. Dessa maneira, incluindo a engenharia de dados no processo de criação de software.

~~Uma~~ Agora reproduzir o software não é somente o código, agora ~~temos~~ temos modelos nesse processo e eles são sensíveis aos dados em si ~~em~~ e são necessários para

SUA CONSTRUÇÃO. SENDO QUE OS DADOS ~~POSSUAM~~ ~~FEITURA~~, SEGUINDO AS LEGISLAÇÕES SÃO DOS SEUS USUÁRIOS. ELAS TEM O DIREITO DE SABER ~~SE~~ SOBRE AS INFORMAÇÕES COLETADAS, EDIÇÃO E REMOÇÃO. ISSO TORNA O PIPELINE MAIS DIFÍCIL.

CONSIDERANDO O QUE FOI ABERDADO E CONCEITUADO SOBRE O PROBLEMA DE TOMADA DE DECISÃO AUTOMÁTICO E ÉTICA, É POSSÍVEL PROPOR ~~UMA~~ UMA ESTRATÉGIA DE GOVERNANÇA PARA ESSE TIPO DE CONTEXTO. PRIMEIRO É NECESSÁRIO PENSAR QUE A LEGI LAÇÃO NÃO É SO UMA NORMA E SIM FAZER PARTE DO PROCESSO.

PRIMEIRO, A GDPR TEM ALGUMAS NORMAS PARA O TRATAMENTO DE ~~EM~~ EUROPEUS FORA DO PAÍS. DESSA FORMA, OS DADOS SEGUIRIAM ESSAS NORMAS ATÉ PODENDO ESTÁ EM DATACENTERS LOCAIS.

SERIA LEVANTADO OS RISCOS DESSOS MODELOS E DECIDIDO A RESPONSABILIDADE DE CADA UM. ALÉM DISSO, SERIA NECESSÁRIO UM PLANO DE RISCOS PARA UZAMENTO DE DADOS. NO DESENVOLVIMENTO DO MODELO, É NECESSÁRIO DE USO DE CONTAINERS. ~~AND~~ ELAS VÃO SER RESPONSÁVEIS PARA UNIFORMIZAR O AMBIENTE DE TREINAMENTO E EXECUÇÃO ENTRE EQUIPES. ~~DE~~ ELAS SERÃO PARTE DO CÓDIGO E SERÃO VERSIONADOS. OS MODELOS POSSUEM UMA SENSIBILIDADE ALÉM DOS DADOS, MAS TAMBÉM DO AMBIENTE. ~~DE~~ DESSA FORMA, UM DOS PRIMEIROS PROBLEMAS BÁSICOS DE REPRODUTIVIDADE É RESOLVIDO.

Questão 2:

Outro aspecto FUNDAMENTAL é a RASTREABILIDADE DAS DECISÕES. EM CONJUNTO DO CÓDIGO, SERÁ ~~MAIS~~ REPENSADO O DESENVOLVIMENTO ORIENTADO A ~~SER~~ SPECS (ESPECIFICAÇÃO). TODA DECISÃO E USO DE DADOS SERÁ DOCUMENTADA POR QUESTÃO DE RASTREABILIDADE e RESPONSABILIDADE. DESSA FORMA, TORNA-SE POSSÍVEL QUE SEJA APLICADO AUDITORIAS EXTERNAS PARA O LEVANTAMENTO DE CAUSAS DE PROBLEMAS. ESSE PROCESSO TORNA-SE UM DESAFIO EM AMBIENTES UTILANDO AUTO ML (AUTO APRENDIZADO DE MÁQUINA), MAS É POSSÍVEL USAR GUARD RAILS PARA MINIMIZAR.

UM GRANDE DESAFIO É O DADO SENSÍVEL. MUITO DELES PODENDO CARRREGAR VIÉSSES QUE PODEM FAVORECER ALGUM GRUPO, COMO FOI DISCUTIDO ANTERIORMENTE. POR ESSA RAZÃO, É NECESSÁRIO O LEVANTAMENTO DE RISCOS E A RASTREABILIDADE DAS DECISÕES. DESSA MANEIRA, DEMONSTRARÁ QUE A EQUIPE RECONHECEU COMO PROBLEMA E TENTOU FORMAS DE MITIGAR E MANTER COMO UM DOS OBJETIVOS.

NA QUESTÃO DOS USOS DE DADOS, É POSSÍVEL CRIAR ESTRATÉGIAS DE ESEMBLES, ONDE É POSSÍVEL CRIAR MODELOS REGIONAIS OBEDECENDO A LEGISLAÇÃO LOCAL e NEM QUE FAÇA ESSA AGREGAÇÃO. INCLUSIVE AJUDARIA ATÉ NA QUESTÃO DE VAZAMENTO, POIS OS DADOS SERÃO UTILIZADOS

SOMENTE REGIONAL COM A CRIPTOGRAFIA LOCAL. OUTRA ABERDAGEM COMPLEMENTAR É GERAR DADOS COM ~~AS~~ AS CARACTERÍSTICAS DA DISTRIBUIÇÃO ORIGINAL.

Por fim, um problema comum em sistemas de recomendação é a transparência. Também é um tópico dentro das legislações em que é necessário transparência na tomada de decisão. Em sistemas de recomendação isso é agravado, pois, dependendo da construção do sistema, é fundamental a confiança dele no usuário e se manter fiel ao sistema. A explicabilidade é um dos mecanismos que ajuda na transparência do sistema. É algo que os projetistas precisam ter em mente na construção do algoritmo.

Existem técnicas que podem ser utilizadas, mesmo utilizando modelos que não são transparentes. ~~da~~ ~~é possível~~ ~~por exemplo~~ construir um ~~modelo~~ normalmente em recomendação é utilizado modelos baseados em redes neurais, mas é possível em conjunto ter alternativas para amenizar o problema.

Por fim, ~~por~~ a etapa fundamental é a observabilidade. Para que o ciclo ~~se~~ funcione é necessário que exista toda uma estrutura para observar as tomadas de decisões sendo realizadas. Dessa maneira, é possível

ed

Questão 2 (cont):

SABER AS CONSEQUÊNCIAS DA TOMADA DE DECISÃO, SE O MODELO JÁ COMETENDO INJUSTIÇA, SE MECANISMO DE EXPLICAÇÃO ESTÃO SENDO USADOS E SE O MODELO AINDA ESTÁ TENDO BONS RESULTADOS

Além disso, a seleção de atributos se torna essencial no contexto atual de Big Data. É possível capturar muitos dados, como exemplo, uma instituição escolar pode utilizar dados como histórico escolar, indicadores socio-econômicos, registros de frequência e entre outros. Cada conjunto de dados possui diversos atributos. Um problema vindo desta característica é a maldição da dimensionalidade. Com o aumento do número de dimensões, o espaço com as observações torna-se menos denso. Diversas técnicas de mineração utilizam a distância entre as observações como meio de descoberta de padrões. Aumentando a esparsidade desse espaço, as similaridades tornam-se menos significativas.

Outro problema que vem do número de atributos é a redundância de informação agregada. Alguns modelos são mais sensíveis a esse problema do que os outros. Basicamente o modelo irá dar um peso maior para essa informação redundante e ~~isso~~ ~~é~~ ~~certamente~~ ~~prejudicial~~ ~~para~~ ~~o~~ ~~modelo~~ essa redundância não tem relação com a tarefa em si e somente com a disponibilidade da informação.

Por fim, outra consequência de uma grande quantidade de atributos é a interpretabilidade dos resultados. A descoberta de um conhecimento só é aplicável se ela possui uma relação e consequência no negócio. Caso contrário, ela é só uma informação. Dessa forma, uma etapa

Questão 3 (cont):

FUNDAMENTAL É A VISUALIZAÇÃO E INTERPRETAÇÃO DOS DADOS.

QUANDO EXISTEM UMA GRANDE QUANTIDADE DE DIMENSÕES, MAIS DIFÍCIL MOSTRAR, PRINCIPALMENTE PARA UM STACKHOLDER, A VISUALIZAÇÃO DESSES DADOS EM UM ESPAÇO N DIMENSIONAL. NESSE CONTEXTO, SURTIAM TÉCNICAS DE VISUALIZAÇÃO ONDE É POSSÍVEL MOSTRAR CLASSE DE DADOS EM DIMENSÕES MAIORES ATRAVÉS DE VISUALIZAÇÕES ESPECÍFICAS. ALÉM DISSO, É POSSÍVEL UTILIZAR TÉCNICAS DE REDUÇÃO DE ESPAÇO, COMO O PCA, PARA FAZER UMA TRANSFORMAÇÃO PARA UM ESPAÇO MENOR. PARA ~~SE~~ SER UTILIZADO ~~ALGUMA~~ TÉCNICAS ~~DE~~ DE VISUALIZAÇÃO EM BAIXA DIMENSÃO, A GRANDE DIFICULDADE DESSE TIPO DE TÉCNICA É QUE É NECESSÁRIO REDUZIR AS INFORMAÇÕES, E, DESSA MANEIRA, SENDO MUITO DEPENDENTE DA TÉCNICA UTILIZADA PARA REDUÇÃO. OUTRA DIFICULDADE TAMBÉM É A ~~RA~~ RELAÇÃO DO ATRIBUTO DO ESPAÇO REDUZIDO COM O DO ORIGINAL, TORNANDO A INTERPRETAÇÃO UM POUCO DIFICULTADA.

UMA ESTRATÉGIA COMUM É A EXLIMINAÇÃO DE ATRIBUTOS, SENDO POR EXEMPLO A CORRELAÇÃO OU A QUANTIDADE DE INFORMAÇÃO MÚTUA, E ~~ALGUMAS~~ EM ADIÇÃO É COMUM TAMBÉM A CRIAÇÃO DE NOVOS ATRIBUTOS DERIVADOS, PRINCIPALMENTE ALINHADO COM O NEGÓCIO. A GRANDE ~~DE~~ DIFICULDADE NA EXLIMINAÇÃO DE ATRIBUTOS É A REMOÇÃO DE INFORMAÇÃO

QUE SERIA ÚTIL PARA ~~EM MODELOS~~ A CRIAÇÃO DOS MODELOS.

COM TUDO O QUE FOI EXPOSTO, É POSSÍVEL QUE O PRE-PROCESSAMENTO POSSUI UM PAPEL FUNDAMENTAL DENTRO DO KDD. OS DADOS SÃO A BASE PARA QUE TÉCNICAS E MODELOS POSSAM ENCONTRAR PADRÕES E ESSAS INFORMAÇÕES POSSAM TER COMO CONSEQUÊNCIA A CRIAÇÃO DE CONHECIMENTO. INCLUSIVE, ATÉ A CAPTURA DOS DADOS PODEM IMPACTAR NESSE PROCESSO. UM PROBLEMA CLÁSSICO É O VIÉS DE AMOSTRAGEM, ~~PODEMOS~~ ~~REPERCUTIR~~ ~~EM~~ MESMO ASSIM, A QUANTIDADE DE OBSERVAÇÕES COM UMA GRANDE QUANTIDADE DE ATRIBUTOS PODEM TORNAR O PROCESSO DE DESCOBERTA UM DESAFIO.