1. Apresentação
   1. Meu nome é Gabriel e vou apresentar meu projeto final da formação cientista de dados
2. Visão geral
   1. Os objetivos do meu projeto são: . . .
   2. As etapas do projeto foram:
      1. Extração dos dados utilizando web scrapping
      2. Transformação dos dados
      3. Carregamento dos dados em arquivo CSV e RData
      4. Aplicação de algoritmos de cluster e classificação
      5. Desenvolvimento de dashboard
3. Solução técnica
   1. Para o desenvolvimento do projeto foram utilizadas as seguintes ferramentas:
      1. QGIS, PostgreSQL e PostGIS: foram utilizadas para montar a base de dados geográfica, para elaborar mapas
      2. R: foi utilizado para fazer o web scrapping e todo pre-processamento dos dados e também para aplicar algoritmo de cluster
      3. Python: foi utilizado para desenvolver modelos de classificação
      4. Para o desenvolvimento da plataforma web foram utilizadas as tecnologias HTML 5, CSS, JS, JQUERY, D3.JS, HIGCHARTS e BOOTSTRAP
   2. O processo do ETL foi executado da seguinte forma:
      1. EXTRAÇÃO: foi realizado um web scrapping em sites de vagas de emprego e de avaliações de empresas
      2. TRANSFORMAÇÃO: foram aplicadas várias limpezas e transformações nos dados, dentre elas:
         1. Remoção de caracteres especiais;
         2. Novas variáveis para armazenas skills, idioma, graduação referentes a vaga de emprego, estado, cidade das vagas de emprego e das avaliações sobre as empresas
         3. Conversão das datas de postagem das vagas de emprego e revisões
      3. LOAD: após os processos anteriores, estes dados foram armazenados em CSV para ser utilizado nos dashboards e RData um formato mais leve para disponibilizar para a comunidade;
   3. Descrição dos dados
      1. A base de dados de vagas de emprego é formada por 3.260 registros contendo 13 campos
      2. A base de dados de reviews é formada por 109.606 registros contendo 17 campos
      3. Para desenvolver os dashboard os campos utilizados foram:
         1. Para vagas de emprego: . . .
         2. Para reviews: ...
   4. DATA MINING
      1. Nesse projeto foram aplicados duas técnicas de minieração de dados:
         1. CLUSTER: foi utilizado para agrupar vagas de emprego que possuam requisitos semelhantes
         2. CLASSIFICAÇÃO: foi utilizada para classificara partir do comentário, se aquele funcionário/ex-funcionário recomenda ou não a empresa.
   5. Para fazer cluster de skill de vagas de emprego, dentre as 13 variaveis, foram utilizadas as variáveis: state, skills, edu...
      1. O parâmetro K, valor de cluster, foi de 4;
   6. Para desenvolver o modelo de classificação de reviews, dentre 17 variaveis, foi utilizada a variável Title como variável independente e a variável recommend como variável dependente
   7. Foram desenvolvidos diferentes modelos, como, NB, SVC, SVM e DL, o que obteve melhor resultado foi o modelo de ... com ....
   8. Agora irei apresentar o sistema desenvolvido que está disponível no GITHUB
4. Agora vou apresentar a Solução PROPOSTA técnica do projeto. . .
5. As ferramentas utilizadas nesse projeto foram
   1. QGIS, POSTGRESQL e POSTGIS para desenvolver os mapas que foram utilizados, em formato geojson.
   2. R foi utilizado em toda pipeline do projeto a coleta via web scraping, o pré-processsamento dos dados e data mining.
   3. O Python foi utilizado para desenvolver o modelo de classificação
   4. Já para o desenvolvimento da parte de dashboards web, foram utilizadas HTML, css, js, jquery, d3.js, bootstrap e highcharts
6. O processo de ETL foram executadas:
   1. Extração dos dados via web scrapping
   2. Transformaçacao dos dados: foram removidos caracteres especiais de alguns campos, foram criadas novas variáveis como skills, idioma, graduação estado, cidade dentre outras e conversão das datas
   3. Os dados foram salvos em formato CSV que foram utilizados nos dashboard web, com d3.js e em RData por ser um formato mais leve o que possibilita disponibilizar para download
7. A base de dados de vagas de emprego possui 3.260 registros com 13 variaveis, como, o ID, cargo, data de postagem da vaga, cidade dentre outras
8. Já a base de dados de avaliações, possui 109.606 registros, com 17 variaveis, dentre ela ID, nome da empresa, data, data da coleta, rating, recomendação, comentário positivos e negativos, dentre outras.
9. Para o desenvolvimento dos dashboards de vagas de emprego, foram utilizadas as variáveis data, cargo, idioma, graduação, skill, cidade, estado, clusters e país.
10. Para o desenvolvimento dos dashboards de avaliações, foram utilizadas as variáveis data, cargo, cidade, estado, comentários, recomendação, empresa e rating.
11. Foram aplicadas duas técnicas de data mining
    1. Cluster: com objetivo de agrupar vagas de emprego com requisistos semelhantes
    2. E, classificação: para classificar através do título da avaliação, se o avaliador recomenda ou não aquela empresa. Algumas avaliações não possuem a recomendação, então o modelo desenvolvido foi aplicado a esses dados.
12. Seleção de entidades...
    1. para aplicar a técnica de cluster, foram utilizadas as variáveis, estado, skills, graduação e idioma. Foi utilizado o algoritmo K-means com 4 cluster
    2. para o modelo de classificação, a variável dependendo é a recommend, que indica se recomenda ou não a empresa e tem 100% dos registros preenchidos. A variável independente utilziada foi a do titulo da avaliação que tem 99,53% dos registros preenchidos.
13. RESULTADOS
    1. Para desenvolver o modelo, a base de dados foi dividida em 75% como dados de treino e 25% como dados de teste
    2. Os algoritmos utilizados foram Naive Bayes Multinomial, SV, naive bayes binomial e deep learning. O modelo com melhor acuracia foi o Naive Bays multinomial
    3. Mesmo ainda não sendo uma acuracia ideal, ainda sendo necessário tratar melhor os dados e melhorar o modelo.
14. Apresentação do sistema
    1. Agora irei apresentar o resutlado após o sistema desenvolvido.