데이터 사이언스 인터뷰 질문 모음집

17 Feb 2018 in Data on Deep-Learning

- 데이터 사이언스 분야의 인터뷰 질문을 모아봤습니다. (데이터 분석가 / 데이터 사이언티스트 / 데이터 엔지니어)
  - **구직자**에겐 예상 질문을 통해 면접 합격을 할 수 있도록, **면접관**에겐 좋은 면접 질문을 할 수 있도록, 딥러닝 **공부**하는 분들에겐 용어를 알 수 있도록 도와드리기 위해 본 문서를 만들게 되었습니다. 다만 여기에 나온 모든 것들을 알 필요는 없습니다. 특정 분야에 이런 것들이 있나보다~ 정도의 키워드만 가져가도 충분히 좋을 것 같습니다!
  - (단, 테마가 어색하거나 질문이 이상한 것들이 존재할 수 있으니 꼭 선별해서 보시길 부탁드립니다!!!)
  - 정답은 따로 작성하지 않으며 질문 위주로 작성했습니다. (하용호님, 남세동님의 허락을 받아 내용을 포함했습니다)
- 질문의 테마를 넣기 애매할 경우엔 제가 임의로 넣었습니다. 또한 특정 분야는 제 얕은 지식으로 문제를 작성했으니, 문제가 부족하다고 느끼시면 언제든 댓글 날려주세요 :)
- Github에서 보실 분들은 링크를 클릭해주세요

- Data Science 공부에 대한 전반적인 내용이 궁금하신 분은 I-want-to-study-Data-Science 문서를 보시면 좋을 것 같아요 :)
- 개발 전반적인 면접 질문은 Interview_Question_for_Beginner를 추천합니다! 신입 데이터 엔지니어의 경우 해당 문서에 있는 CS 내용을 숙지하면 좋을 것 같습니다

Contents

공통 질문

- 왜 해당 직군으로 지원했나요?
- 왜 저희 회사에 지원하셨나요?
- 해당 직군의 매력이 무엇이라고 생각하나요?
- 해당 직군에서 본인의 장점은?

- 해당 직군을 하면서 이루고자 하는 목표는?
- 해당 직군을 하기 위해 어떤 노력을 했나요?
- 왜 저희가 지원자를 뽑아야 하나요?
- 지원자의 단점은 무엇인가요?

목차로 이동

프로젝트

- 데이터를 어떻게 구했나요?
- 해당 프로젝트에서 왜 이 알고리즘을 사용했나요?
- 그 알고리즘과 유사한 알고리즘이 존재하지 않나요?
- 해당 알고리즘의 단점은?
- 해당 프로젝트에서 지원자는 어떤 일을 했나요?
- 해당 프로젝트에서 지원자가 느낀 점은?
- 해당 프로젝트를 다시 진행한다고 하면 어떻게 할 것인가요?
- Kaggle에서 수상을 하면 데이터 분석을 잘 할까요?

목차로 이동

통계 및 수학

- 고유값(eigen value)와 고유벡터(eigen vector)에 대해 설명해주세요. 그리고 왜 중요할까요?
- 샘플링(Sampling)과 리샘플링(Resampling)에 대해 설명해주세요. 리샘플링은 무슨 장점이 있을까요?
- 확률 모형과 확률 변수는 무엇일까요?
- 누적 분포 함수와 확률 밀도 함수는 무엇일까요? 수식과 함께 표현해주세요
- 베르누이 분포 / 이항 분포 / 카테고리 분포 / 다항 분포 / 가우시안 정규 분포 / t 분포 / 카이제곱 분포 / F 분포 / 베타 분포 / 감마 분포 / 디리클레 분포에 대해 설명해주세요. 혹시 연관된 분포가 있다면 연관 관계를 설명해주세요
- 조건부 확률은 무엇일까요?
- 공분산과 상관계수는 무엇일까요? 수식과 함께 표현해주세요
- 신뢰 구간의 정의는 무엇인가요?
- p-value를 고객에게는 뭐라고 설명하는게 이해하기 편할까요?
- p-value는 요즘 시대에도 여전히 유효할까요? 언제 p-value가 실제를 호도하는 경향이 있을까요?

- A/B Test 등 현상 분석 및 실험 설계 상 통계적으로 유의미함의 여부를 결정하기 위한 방법에는 어떤 것이 있을까요?
- R square의 의미는 무엇인가요?
- 평균(mean)과 중앙값(median)중에 어떤 케이스에서 뭐를 써야할까요?
- 중심극한정리는 왜 유용한걸까요?
- 엔트로피(entropy)에 대해 설명해주세요. 가능하면 Information Gain도요.
- 요즘같은 빅데이터(?)시대에는 정규성 테스트가 의미 없다는 주장이 있습니다. 맞을까요?
- 어떨 때 모수적 방법론을 쓸 수 있고, 어떨 때 비모수적 방법론을 쓸 수 있나요?
- “likelihood”와 “probability”의 차이는 무엇일까요?
- 통계에서 사용되는 bootstrap의 의미는 무엇인가요.
- 모수가 매우 적은 (수십개 이하) 케이스의 경우 어떤 방식으로 예측 모델을 수립할 수 있을까요?
- 베이지안과 프리퀀티스트간의 입장차이를 설명해주실 수 있나요?
- 검정력(statistical power)은 무엇일까요?
- missing value가 있을 경우 채워야 할까요? 그 이유는 무엇인가요?
- 아웃라이어의 판단하는 기준은 무엇인가요?
- 콜센터 통화 지속 시간에 대한 데이터가 존재합니다. 이 데이터를 코드화하고 분석하는 방법에 대한 계획을 세워주세요. 이 기간의 분포가 어떻게 보일지에 대한 시나리오를 설명해주세요
- 출장을 위해 비행기를 타려고 합니다. 당신은 우산을 가져가야 하는지 알고 싶어 출장지에 사는 친구 3명에게 무작위로 전화를 하고 비가 오는 경우를 독립적으로 질문해주세요. 각 친구는 2/3로 진실을 말하고 1/3으로 거짓을 말합니다. 3명의 친구가 모두 “그렇습니다. 비가 내리고 있습니다”라고 말했습니다. 실제로 비가 내릴 확률은 얼마입니까?
- 필요한 표본의 크기를 어떻게 계산합니까?
- Bias를 통제하는 방법은 무엇입니까?
- 로그 함수는 어떤 경우 유용합니까? 사례를 들어 설명해주세요

목차로 이동

분석 일반

- 좋은 feature란 무엇인가요. 이 feature의 성능을 판단하기 위한 방법에는 어떤 것이 있나요?
- “상관관계는 인과관계를 의미하지 않는다”라는 말이 있습니다. 설명해주실 수 있나요?
- A/B 테스트의 장점과 단점, 그리고 단점의 경우 이를 해결하기 위한 방안에는 어떤 것이 있나요?
- 각 고객의 웹 행동에 대하여 실시간으로 상호작용이 가능하다고 할 때에, 이에 적용 가능한 고객 행동 및 모델에 관한 이론을 알아봅시다.

- 고객이 원하는 예측모형을 두가지 종류로 만들었다. 하나는 예측력이 뛰어나지만 왜 그렇게 예측했는지를 설명하기 어려운 random forest 모형이고, 또다른 하나는 예측력은 다소 떨어지나 명확하게 왜 그런지를 설명할 수 있는 sequential bayesian 모형입니다.고객에게 어떤 모형을 추천하겠습니까?
- 고객이 내일 어떤 상품을 구매할지 예측하는 모형을 만들어야 한다면 어떤 기법(예: SVM, Random Forest, logistic regression 등)을 사용할 것인지 정하고 이를 통계와 기계학습 지식이 전무한 실무자에게 설명해봅시다.
- 나만의 feature selection 방식을 설명해봅시다.
- 데이터 간의 유사도를 계산할 때, feature의 수가 많다면(예: 100개 이상), 이러한 high-dimensional clustering을 어떻게 풀어야할까요?

목차로 이동

머신러닝

- Cross Validation은 무엇이고 어떻게 해야하나요?
- 회귀 / 분류시 알맞은 metric은 무엇일까요?
- 알고 있는 metric에 대해 설명해주세요(ex. RMSE, MAE, recall, precision …)
- 정규화를 왜 해야할까요? 정규화의 방법은 무엇이 있나요?
- Local Minima와 Global Minima에 대해 설명해주세요.
- 차원의 저주에 대해 설명해주세요
- dimension reduction기법으로 보통 어떤 것들이 있나요?
- PCA는 차원 축소 기법이면서, 데이터 압축 기법이기도 하고, 노이즈 제거기법이기도 합니다. 왜 그런지 설명해주실 수 있나요?
- LSA, LDA, SVD 등의 약자들이 어떤 뜻이고 서로 어떤 관계를 가지는지 설명할 수 있나요?
- Markov Chain을 고등학생에게 설명하려면 어떤 방식이 제일 좋을까요?
- 텍스트 더미에서 주제를 추출해야 합니다. 어떤 방식으로 접근해 나가시겠나요?
- SVM은 왜 반대로 차원을 확장시키는 방식으로 동작할까요? 거기서 어떤 장점이 발생했나요?
- 다른 좋은 머신 러닝 대비, 오래된 기법인 나이브 베이즈(naive bayes)의 장점을 옹호해보세요.
- Association Rule의 Support, Confidence, Lift에 대해 설명해주세요.
- 최적화 기법중 Newton's Method와 Gradient Descent 방법에 대해 알고 있나요?
- 머신러닝(machine)적 접근방법과 통계(statistics)적 접근방법의 둘간에 차이에 대한 견해가 있나요?
- 인공신경망(deep learning이전의 전통적인)이 가지는 일반적인 문제점은 무엇일까요?
- 지금 나오고 있는 deep learning 계열의 혁신의 근간은 무엇이라고 생각하시나요?
- ROC 커브에 대해 설명해주실 수 있으신가요?

- 여러분이 서버를 100대 가지고 있습니다. 이때 인공신경망보다 Random Forest를 써야하는 이유는 뭘까요?
- K-means의 대표적 의미론적 단점은 무엇인가요? (계산량 많다는것 말고)
- L1, L2 정규화에 대해 설명해주세요
- XGBoost을 아시나요? 왜 이 모델이 캐글에서 유명할까요?
- 앙상블 방법엔 어떤 것들이 있나요?
- SVM은 왜 좋을까요?
- feature vector란 무엇일까요?
- 좋은 모델의 정의는 무엇일까요?
- 50개의 작은 의사결정 나무는 큰 의사결정 나무보다 괜찮을까요? 왜 그렇게 생각하나요?
- 스팸 필터에 로지스틱 리그레션을 많이 사용하는 이유는 무엇일까요?
- OLS(ordinary least squre) regression의 공식은 무엇인가요?

목차로 이동

딥러닝

딥러닝 일반

- 딥러닝은 무엇인가요? 딥러닝과 머신러닝의 차이는?
- 왜 갑자기 딥러닝이 부흥했을까요?
- 마지막으로 읽은 논문은 무엇인가요? 설명해주세요
- Cost Function과 Activation Function은 무엇인가요?
- Tensorflow, Keras, PyTorch, Caffe, Mxnet 중 선호하는 프레임워크와 그 이유는 무엇인가요?
- Data Normalization은 무엇이고 왜 필요한가요?
- 알고있는 Activation Function에 대해 알려주세요. (Sigmoid, ReLU, LeakyReLU, Tanh 등)
- 오버피팅일 경우 어떻게 대처해야 할까요?
- 하이퍼 파라미터는 무엇인가요?
- Weight Initialization 방법에 대해 말해주세요. 그리고 무엇을 많이 사용하나요?
- 볼츠만 머신은 무엇인가요?
- 요즘 Sigmoid 보다 ReLU를 많이 쓰는데 그 이유는?
    - Non-Linearity라는 말의 의미와 그 필요성은?
    - ReLU로 어떻게 곡선 함수를 근사하나?
    - ReLU의 문제점은?
    - Bias는 왜 있는걸까?

- Gradient Descent에 대해서 쉽게 설명한다면?
  - 왜 꼭 Gradient를 써야 할까? 그 그래프에서 가로축과 세로축 각각은 무엇인가? 실제 상황에서는 그 그래프가 어떻게 그려질까?
  - GD 중에 때때로 Loss가 증가하는 이유는?
  - 중학생이 이해할 수 있게 더 쉽게 설명 한다면?
  - Back Propagation에 대해서 쉽게 설명 한다면?
- Local Minima 문제에도 불구하고 딥러닝이 잘 되는 이유는?
  - GD가 Local Minima 문제를 피하는 방법은?
  - 찾은 해가 Global Minimum인지 아닌지 알 수 있는 방법은?
- Training 세트와 Test 세트를 분리하는 이유는?
  - Validation 세트가 따로 있는 이유는?
  - Test 세트가 오염되었다는 말의 뜻은?
  - Regularization이란 무엇인가?
- Batch Normalization의 효과는?
  - Dropout의 효과는?
  - BN 적용해서 학습 이후 실제 사용시에 주의할 점은? 코드로는?
  - GAN에서 Generator 쪽에도 BN을 적용해도 될까?
- SGD, RMSprop, Adam에 대해서 아는대로 설명한다면?
  - SGD에서 Stochastic의 의미는?
  - 미니배치를 작게 할때의 장단점은?
  - 모멘텀의 수식을 적어 본다면?
- 간단한 MNIST 분류기를 MLP+CPU 버전으로 numpy로 만든다면 몇줄일까?
  - 어느 정도 돌아가는 녀석을 작성하기까지 몇시간 정도 걸릴까?
  - Back Propagation은 몇줄인가?
  - CNN으로 바꾼다면 얼마나 추가될까?
- 간단한 MNIST 분류기를 TF, Keras, PyTorch 등으로 작성하는데 몇시간이 필요한가?
  - CNN이 아닌 MLP로 해도 잘 될까?
  - 마지막 레이어 부분에 대해서 설명 한다면?
  - 학습은 BCE loss로 하되 상황을 MSE loss로 보고 싶다면?
  - 만약 한글 (인쇄물) OCR을 만든다면 데이터 수집은 어떻게 할 수 있을까?
- 딥러닝할 때 GPU를 쓰면 좋은 이유는?

    - 학습 중인데 GPU를 100% 사용하지 않고 있다. 이유는?
    - GPU를 두개 다 쓰고 싶다. 방법은?
    - 학습시 필요한 GPU 메모리는 어떻게 계산하는가?

- TF, Keras, PyTorch 등을 사용할 때 디버깅 노하우는?
- 뉴럴넷의 가장 큰 단점은 무엇인가? 이를 위해 나온 One-Shot Learning은 무엇인가?

목차로 이동

컴퓨터 비전

- OpenCV 라이브러리만을 사용해서 이미지 뷰어(Crop, 흑백화, Zoom 등의 기능 포함)를 만들어주세요
- 딥러닝 발달 이전에 사물을 Detect할 때 자주 사용하던 방법은 무엇인가요?
- Fatser R-CNN의 장점과 단점은 무엇인가요?
- dlib은 무엇인가요?
- YOLO의 장점과 단점은 무엇인가요?
- 제일 좋아하는 Object Detection 알고리즘에 대해 설명하고 그 알고리즘의 장단점에 대해 알려주세요
    - 그 이후에 나온 더 좋은 알고리즘은 무엇인가요?
- Average Pooling과 Max Pooling의 차이점은?
- Deep한 네트워크가 좋은 것일까요? 언제까지 좋을까요?
- Residual Network는 왜 잘될까요? Ensemble과 관련되어 있을까요?
- CAM(Class Activation Map)은 무엇인가요?
- Localization은 무엇일까요?
- 자율주행 자동차의 원리는 무엇일까요?
- Semantic Segmentation은 무엇인가요?
- Visual Q&A는 무엇인가요?
- Image Captioning은 무엇인가요?
- Fully Connected Layer의 기능은 무엇인가요?
- Neural Style은 어떻게 진행될까요?
- CNN에 대해서 아는대로 얘기하라
    - CNN이 MLP보다 좋은 이유는?
    - 어떤 CNN의 파라미터 개수를 계산해 본다면?
    - 주어진 CNN과 똑같은 MLP를 만들 수 있나?
    - 풀링시에 만약 Max를 사용한다면 그 이유는?

- 시퀀스 데이터에 CNN을 적용하는 것이 가능할까?

목차로 이동

자연어 처리

- One Hot 인코딩에 대해 설명해주세요
- POS 태깅은 무엇인가요? 가장 간단하게 POS tagger를 만드는 방법은 무엇일까요?
- 문장에서 “Apple”이란 단어가 과일인지 회사인지 식별하는 모델을 어떻게 훈련시킬 수 있을까요?
- 뉴스 기사에 인용된 텍스트의 모든 항목을 어떻게 찾을까요?
- 음성 인식 시스템에서 생성된 텍스트를 자동으로 수정하는 시스템을 어떻게 구축할까요?
- 잠재론적, 의미론적 색인은 무엇이고 어떻게 적용할 수 있을까요?
- 영어 텍스트를 다른 언어로 번역할 시스템을 어떻게 구축해야 할까요?
- 뉴스 기사를 주제별로 자동 분류하는 시스템을 어떻게 구축할까요?
- Stop Words는 무엇일까요? 이것을 왜 제거해야 하나요?
- 영화 리뷰가 긍정적인지 부정적인지 예측하기 위해 모델을 어떻게 설계하시겠나요?
- TF-IDF 점수는 무엇이며 어떤 경우 유용한가요?
- 한국어에서 많이 사용되는 사전은 무엇인가요?
- Regular grammar는 무엇인가요? regular expression과 무슨 차이가 있나요?
- RNN에 대해 설명해주세요
- LSTM은 왜 유용한가요?
- Translate 과정 Flow에 대해 설명해주세요
- n-gram은 무엇일까요?
- PageRank 알고리즘은 어떻게 작동하나요?
- depedency parsing란 무엇인가요?
- Word2Vec의 원리는?
  - 그 그림에서 왼쪽 파라메터들을 임베딩으로 쓰는 이유는?
  - 그 그림에서 오른쪽 파라메터들의 의미는 무엇일까?
  - 남자와 여자가 가까울까? 남자와 자동차가 가까울까?
  - 번역을 Unsupervised로 할 수 있을까?

목차로 이동

강화학습

- MDP는 무엇일까요?
- 가치함수는 무엇일까요? 수식으로도 표현해주세요
- 벨만 방정식은 무엇일까요? 수식으로도 표현해주세요
- 강화학습에서 다이나믹 프로그래밍은 어떤 의미를 가질까요? 한계점은 무엇이 있을까요?
- 몬테카를로 근사는 무엇일까요? 가치함수를 추정할 때 어떻게 사용할까요?
- Value-based Reinforcement Learning과 Policy based Reinforcement Learning는 무엇이고 어떤 관계를 가질까요?
- 강화학습이 어려운 이유는 무엇일까요? 그것을 어떤 방식으로 해결할 수 있을까요?
- 강화학습을 사용해 테트리스에서 고득점을 얻는 프로그램을 만드려고 합니다. 어떻게 만들어야 할까요?

목차로 이동

GAN

- GAN에 대해 아는대로 설명해주세요
- GAN의 단점은 무엇인가요?
- LSGAN에 대해 설명해주세요
- GAN이 왜 뜨고 있나요?
- Auto Encoder에 대해서 아는대로 얘기하라
  - MNIST AE를 TF나 Keras등으로 만든다면 몇줄일까?
  - MNIST에 대해서 임베딩 차원을 1로 해도 학습이 될까?
  - 임베딩 차원을 늘렸을 때의 장단점은?
  - AE 학습시 항상 Loss를 0으로 만들수 있을까?
  - VAE는 무엇인가?
- 간단한 MNIST DCGAN을 작성한다면 TF 등으로 몇줄 정도 될까?
  - GAN의 Loss를 적어보면?
  - D를 학습할때 G의 Weight을 고정해야 한다. 방법은?
  - 학습이 잘 안될때 시도해 볼 수 있는 방법들은?

목차로 이동

추천 시스템

- 추천 시스템에서 사용할 수 있는 거리는 무엇이 있을까요?

- User 베이스 추천 시스템과 Item 베이스 추천 시스템 중 단기간에 빠른 효율을 낼 수 있는 것은 무엇일까요?
- 성능 평가를 위해 어떤 지표를 사용할까요?
- Explicit Feedback과 Implicit Feedback은 무엇일까요? Impicit Feedback을 어떻게 Explicit하게 바꿀 수 있을까요?
- Matrix Factorization은 무엇인가요? 해당 알고리즘의 장점과 단점은?
- SQL으로 조회 기반 Best, 구매 기반 Best, 카테고리별 Best를 구하는 쿼리를 작성해주세요
- 추천 시스템에서 KNN 알고리즘을 활용할 수 있을까요?
- 유저가 10만명, 아이템이 100만개 있습니다. 이 경우 추천 시스템을 어떻게 구성하시겠습니까?
- 딥러닝을 활용한 추천 시스템의 사례를 알려주세요
- 두 추천엔진간의 성능 비교는 어떤 지표와 방법으로 할 수 있을까요? 검색엔진에서 쓰던 방법을 그대로 쓰면 될까요? 안될까요?
- Collaborative Filtering에 대해 설명한다면?
- Cold Start의 경우엔 어떻게 추천해줘야 할까요?
- 고객사들은 기존 추천서비스에 대한 의문이 있습니다. 주로 매출이 실제 오르는가 하는 것인데, 이를 검증하기 위한 방법에는 어떤 것이 있을까요? 위 관점에서 우리 서비스의 성능을 고객에게 명확하게 인지시키기 위한 방법을 생각해봅시다.

목차로 이동

데이터베이스

- PostgreSQL의 장점은 무엇일까요?
- 인덱스는 크게 Hash 인덱스와 B+Tree 인덱스가 있습니다. 이것은 무엇일까요?
- 인덱스 Scan 방식은 무엇이 있나요?
- 인덱스 설계시 NULL값은 고려되야 할까요?
- Nested Loop 조인은 무엇일까요?
- Windows 함수는 무엇이고 어떻게 작성할까요?
- KNN 알고리즘을 쿼리로 구현할 수 있을까요?
- MySQL에서 대량의 데이터(500만개 이상)를 Insert해야하는 경우엔 어떻게 해야할까요?
- RDB의 char와 varchar의 차이는 무엇일까요?
- 구글의 BigQuery, AWS의 Redshift는 기존 RDB와 무슨 차이가 있을까요? 왜 빠를까요?
- 쿼리의 성능을 확인하기 위해 어떤 쿼리문을 작성해야 할까요?
- MySQL이 요새 느리다는 신고가 들어왔습니다. 첫번째로 무엇을 확인하시고 조정하시겠나요?
- 동작하는 MySQL에 Alter table을 하면 안되는 이유를 설명해주세요. 그리고 대안을 설명해주세요

- 빡세게 동작하고 있는 MySQL을 백업뜨기 위해서는 어떤 방법이 필요할까요?

목차로 이동

데이터 시각화

- 네트워크 관계를 시각화해야 할 경우 어떻게 해야할까요?
- Tableau같은 BI Tool은 어느 경우 도입하면 좋을까요?
- “신규/재방문자별 지역별(혹은 일별) 방문자수와 구매전환율”이나 “고객등급별 최근방문일별 고객수와 평균 구매금액”와 같이 4가지 이상의 정보를 시각화하는 가장 좋은 방법을 추천해주세요
- 구매에 영향을 주는 요소의 발견을 위한 관점에서, 개인에 대한 쇼핑몰 웹 활동의 시계열 데이터를 효과적으로 시각화하기 위한 방법은 무엇일까요? 표현되어야 하는 정보(feature)는 어떤 것일까요? 실제시 어떤 것이 가장 고민될까요?
- 파이차트는 왜 구릴까요? 언제 구린가요? 안구릴때는 언제인가요?
- 히스토그램의 가장 큰 문제는 무엇인가요?
- 워드클라우드는 보기엔 예쁘지만 약점이 있습니다. 어떤 약점일까요?
- 어떤 1차원값이, 데이터가 몰려있어서 직선상에 표현했을 때 보기가 쉽지 않습니다. 어떻게 해야할까요?

목차로 이동

시스템 엔지니어링

- 지속적인 Cron 작업이 필요합니다. (dependency가 있는 작업들도 존재합니다) 어떻게 작업들을 관리할까요?
- 처음 서버를 샀습니다. 어떤 보안적 조치를 먼저 하시겠습니까?
- SSH로의 brute-force attack을 막기 위해서 어떤 조치를 취하고 싶으신가요?
- 프로세스의 CPU 상태를 보기 위해 top을 했습니다. user,system,iowait중에 뭐를 제일 신경쓰시나요? 이상적인 프로그램이라면 어떻게 저 값들이 나오고 있어야 할까요?
- iowait이 높게 나왔다면, 내가 해야하는 조치는 무엇인가요? (돈으로 해결하는 방법과 소프트웨어로 해결하는 방법을 대답해주세요)
- 동시에 10개의 컴퓨터에 라이브러리를 설치하는 일이 빈번히 발생합니다. 어떤 해결책이 있을까요?
- screen과 tmux중에 뭘 더 좋아하시나요?
- vim입니까. emacs입니까. 소속을 밝히세요.
- 가장 좋아하는 리눅스 배포판은 뭡니까. 왜죠?
- 관리하는 컴퓨터가 10대가 넘었습니다. 중요한 모니터링 지표는 뭐가 있을까요? 뭐로 하실건가요?

- GIT의 소스가 있고, 서비스 사용중인 웹서버가 10대 이상 넘게 있습니다. 어떻게 배포할건가요?

목차로 이동

분산처리

- Apache Beam에 대해 아시나요? 기존 하둡과 어떤 차이가 있을까요?
- 좋게 만들어진 MapReduce는 어떤 프로그램일까요? 데이터의 Size 변화의 관점에서 설명할 수 있을까요?
- 여러 MR작업의 연쇄로 최종결과물이 나올때, 중간에 작업이 Fail날수 있습니다. 작업의 Fail은 어떻게 모니터링 하시겠습니까? 작업들간의 dependency는 어떻게 해결하시겠습니까?
- 분산환경의 JOIN은, 보통 디스크, CPU, 네트워크 중 어디에서 병목이 발생할까요? 이를 해결하기 위해 무엇을 해야 할까요?
- 암달의 법칙에 대해 말해봅시다. 그러므로 왜 shared-nothing 구조로 만들어야 하는지 설명해봅시다.
- shared-nothing 구조의 단점도 있습니다. 어떤 것이 해당할까요?
- Spark이 Hadoop보다 빠른 이유를 I/O 최적화 관점에서 생각해봅시다.
- 카산드라는 망한것 같습니다. 왜 망한것 같나요? 그래도 활용처가 있다면 어디인것 같나요.
- TB 단위 이상의 기존 데이터와 시간당 GB단위의 신생 로그가 들어오는 서비스에서 모든 가입자에게 개별적으로 계산된 실시간 서비스(웹)를 제공하기 위한 시스템 구조를 구상해봅시다.
- 대용량 자료를 빠르게 lookup해야 하는 일이 있습니다. (100GB 이상, 100ms언더로 특정자료 찾기). 어떤 백엔드를 사용하시겠나요? 느린 백엔드를 사용한다면 이를 보완할 방법은 뭐가 있을까요?
- 데이터를 여러 머신으로 부터 모으기 위해 여러 선택지가 있을 수 있습니다. (flume, fluentd등) 아예 소스로부터 kafka등의 메시징 시스템을 바로 쓸 수도 있습니다. 어떤 것을 선호하시나요? 왜죠?

목차로 이동

웹 아키텍쳐

- 트래픽이 몰리는 상황입니다. AWS의 ELB 세팅을 위해서 웹서버는 어떤 요건을 가져야 쉽게 autoscale가능할까요?
- 왜 Apache보다 Nginx가 성능이 좋을까요? node.js가 성능이 좋은 이유와 곁들여 설명할 수 있을까요?
- node.js는 일반적으로 빠르지만 어떤 경우에는 쓰면 안될까요?
- 하나의 IP에서 여러 도메인의 HTTPS 서버를 운영할 수 있을까요? 안된다면 왜인가요? 또 이걸 해결하는 방법이 있는데 그건 뭘까요?
- 개발이 한창 진행되는 와중에도 서비스는 계속 운영되어야 합니다. 이를 가능하게 하는 상용 deploy 환경은 어떻게 구현가능한가요? WEB/WAS/DB/Cluster 각각의 영역에서 중요한 변화가 수반되는 경우에도 동작 가능한, 가장 Cost가 적은 방식을 구상하고 시나리오를 만들어봅시다.

목차로 이동

서비스 구현

- 크롤러를 파이썬으로 구현할 때 BeautifulSoup과 Selenium의 장단점은 무엇일까요?
- 빈번한 접속으로 우리 IP가 차단되었을 때의 해결책은? (대화로 푼다. 이런거 말구요)
- 당장 10분안에 사이트의 A/B 테스트를 하고 싶다면 어떻게 해야 할까요? 타 서비스를 써도됩니다.
- 신규 방문자와 재 방문자를 구별하여 A/B 테스트를 하고 싶다면 어떻게 해야 할까요?
- R의 결과물을 python으로 만든 대시보드에 넣고 싶다면 어떤 방법들이 가능할까요?
- 쇼핑몰의 상품별 노출 횟수와 클릭수를 손쉽게 수집하려면 어떻게 해야 할까요?
- 여러 웹사이트를 돌아다니는 사용자를 하나로 엮어서 보고자 합니다. 우리가 각 사이트의 웹에 우리 코드를 삽입할 수 있다고 가정할 때, 이것이 가능한가요? 가능하다면, 그 방법에는 어떤 것이 있을까요?
- 고객사 혹은 외부 서버와의 데이터 전달이 필요한 경우가 있습니다. 데이터 전달 과정에서 보안을 위해 당연히(plain text)로 전송하는 것은 안됩니다. 어떤 방법이 있을까요?

목차로 이동

대 고객 사이드

- 고객이 궁금하다고 말하는 요소가 내가 생각하기에는 중요하지 않고 다른 부분이 더 중요해 보입니다. 어떤 식으로 대화를 풀어나가야 할까요?
- 현업 카운터 파트와 자주 만나며 실패한 분석까지 같이 공유하는 경우와, 시간을 두고 멋진 결과만 공유하는 케이스에서 무엇을 선택하시겠습니까?
- 고객이 질문지 리스트를 10개를 주었습니다. 어떤 기준으로 우선순위를 정해야 할까요?
- 오프라인 데이터가 결합이 되어야 해서, 데이터의 피드백 주기가 매우 느리고 정합성도 의심되는 상황입니다. 우리가 할 수 있는 액션이나 방향 수정은 무엇일까요?
- 동시에 여러개의 A/B테스트를 돌리기엔 모수가 부족한 상황입니다. 어떻게 해야할까요?
- 고객사가 과도하게 정보성 대시보드만을 요청할 경우, 어떻게 대처해야 할까요?
- 고객사에게 위클리 리포트를 제공하고 있었는데, 금주에는 별다른 내용이 없었습니다. 어떻게 할까요?
- 카페24, 메이크샵 같은 서비스에서 데이터를 어떻게 가져오면 좋을까요?
- 기존에 같은 목적의 업무를 수행하던 조직이 있습니다. 어떻게 관계 형성을 해 나가야 할까요. 혹은 일이 되게 하기 위해서는 어떤 부분이 해소되어야 할까요.
- 인터뷰나 강의에 활용하기 위한 백데이터는 어느 수준까지 일반화 해서 사용해야 할까요?
- 고객사가 우리와 일하고 싶은데 현재는 capa가 되지 않습니다. 어떻게 대처해야 할까요?

목차로 이동

개인정보

- 어떤 정보들이 개인정보에 해당할까요? ID는 개인정보에 해당할까요? 이를 어기지 않는 합법적 방법으로 식별하고 싶으면 어떻게 해야할까요?
- 국내 개인 정보 보호 현황에 대한 견해는 어떠한지요? 만약 사업을 진행하는데 장애요소로 작용한다면, 이에 대한 해결 방안은 어떤 것이 있을까요?
- 제3자 쿠키는 왜 문제가 되나요?

목차로 이동

Reference

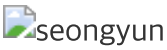
- 하용호님 자료
- 남세동님 자료
- Data Science Interview Questions & Detailed Answers
- Deep Learning Interview Questions and Answers
- Must know questions deeplearning : 객관식 딥러닝 문제
- My deep learning job interview experience sharing
- Natural Language Processing Engineer Interview Questions

카일스쿨 유튜브 채널을 만들었습니다. 데이터 사이언스, 성장, 리더십, BigQuery 등을 이야기할 예정이니, 관심 있으시면 구독 부탁드립니다 :)

이 글이 도움이 되셨다면 추천 클릭을 부탁드립니다 :)

Share this post

About



byeon

Machine Learning Engineer

Related Posts

Rules of Machine Learning: Best Practices for ML Engineering 정리 15 Dec 2019

CS224W - Machine Learning with Graphs 1강 정리 03 Dec 2019

지도 데이터 시각화 : Uber의 pydeck 사용하기 24 Nov 2019

Comments

어쩐지 오늘은

Machine Learning Engineer

- [ ] Diary
  - diary
- [ ] Data

- Machine-Learning
- Deep-Learning
- Mobility
- Engineering
- Optimization
- Paper
- Pytorch
- CS231n
- Reinforcement-Learning
- Time Series
- Analytics
- Kaggle
- Vision
- Interpretable ML
- Simulation
- Graph

☐

MLOps

- Basic
- Airflow
- MLflow
- Tensorflow Extended
- Serving
- Feature
- Experiment & Versioning

☐

Google Cloud Platform

- Basic
- BigQuery
- AI Platform
- Dataflow
- Cloud Functions

Development

- Linux
- Web
- Python
- SQL
- Scala
- OS
- Kubernetes
- DevOps
- Kotlin
- Julia
- Tools

ETC

- Lecture
- Book

About

Search

메모가 습관인 데이터쟁이입니다