

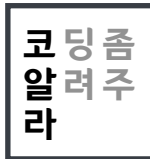
**COALA UNIV**

# 코알라 유니브 해커톤

준비 가이드

2019. 7.





# 코알라유니브 해커톤 준비 가이드

1. 해커톤 구성
2. 구현회 전까지 준비할 일
3. 구현회 당일 안내
4. 발표회 준비법
5. 팀별 피드백

## 해커톤 구성



코알라유니브의 DS 해커톤은 일반적인 해커톤과는 다르게 진행됩니다.

코알라 유니브에서는 장시간·철야로 중단없이 진행되는 기존 프로그래밍 해커톤과 달리 약 3주간 총 3번에 걸친 해커톤 프로젝트를 진행합니다. 더 좋은 아이디어를 구상하고 완성도 있는 프로젝트를 주도적으로 수행할 수 있도록 멘토 및 운영진이 함께 하고 있습니다.



**준비회 7/20 토 11:00 고려대과학도서관, 15:00 서강대J215**

프로젝트 아이디어를 구상하고 전체 실행계획에 대한 밑그림을 그리는 단계입니다.

**구현회 7/27 토 홍합밸리 11:00 ~ 17:00**

계획한 프로젝트를 실질적으로 수행하는 날입니다. 데이터를 수집한 후 모델을 검증하고 현실 세계에 적용하여 멋진 Insight와 효과적인 Solution을 제안해주세요.

**발표회 8/2 금 신촌-파랑고래 18:00 ~ 20:00**

수행한 프로젝트를 정리하여 5분 분량으로 발표합니다. Opensource의 생태계가 발전할 수 있도록 public open 될 예정입니다. \* 가장 우선적인 프로젝트 저작권은 프로젝트 참여자에게 있습니다.

## 구현회 전까지 준비할 일



구현회는 6시간! 그 시간만으로 부족하다면 미리 준비해주세요.

아래는 프로젝트 수행 프로세스입니다.

1. 아이디어 구상
2. 가설, 목적 수립
3. 데이터 수집
4. 머신러닝 등 수리/통계/과학적인 방법으로 데이터 분석
5. **Inishgt·Solution 제작**
6. 발표자료 제작
7. 발표

구현회 종료 후 1~5까지의 프로세스가 진행되어 있기를 권장합니다.

진하게 표시된 3~5 부분을 구현회 당일 수행할 수 있도록 미리 준비해주세요. 만약 프로젝트의 규모가 작아 당일 6시간동안 1~5번 프로세스까지 수행이 가능할 경우 준비하실 필요는 없습니다. 프로젝트의 규모가 커서 시간부족이 예상된다면 얼마든지 미리 준비해도 좋습니다.

\* 멤버들과 의논하여 프로젝트 예상시간을 추측해보고 규모를 미리 파악해주세요. 어렵다면 멘토진에게 도움을 요청하실 수 있습니다.

## 구현회 당일 안내



실질적인 프로젝트 작업을 수행해주세요!

구현회 당일, 아래 프로젝트 수행 프로세스 중 3~5번에 집중된 작업이 진행됩니다.

1. 아이디어 구상
2. 가설, 목적 수립
3. 데이터 수집
4. 머신러닝 등 수리/통계/과학적인 방법으로 데이터 분석
5. Inishgt·Solution 제작
6. 발표자료 제작
7. 발표

구현회가 끝나는 시점에 모델링과 검증, 활용방안 논의를 정리한 후 ‘발표자료 표지’ 첫 page 정도를 완성하며 끝내는게 가장 이상적입니다.

## FAQ

1. **발표자료도 구현회 날 만드나요?**  
만들어도 좋지만 그날 완성할 필요는 없습니다. 발표회 직전까지 만드시면 됩니다.
2. **구현회 날 다 끝내지 못했어요. 어떡하죠?**  
걱정할 필요 없이 발표회 전까지 준비하면 됩니다.
3. **아무래도 프로젝트가 실패할 것 같아요.**  
성공한 프로젝트만큼, 프로젝트 실패도 매우 값진 경험입니다. 우리의 시행착오를 이야기하는, 실험결과 가설과 다른 것으로 판명된 멋진 결론을 담은 실패기 발표를 준비해주세요.
4. **최종적으로는 서비스가 탄생해야 하나요? (분석 결과는 어떻게 보여주어야 하죠?)**  
서비스 형태로 기획하시는 팀이 많습니다. 서비스는 분석 결과를 매력적으로 보여주는 한가지 방법일 뿐입니다. 자유롭게 하시면 됩니다. 실제 서비스 구현없이 가상의 이미지나 프로토타이핑, 계획도 괜찮습니다.

## 최종 산출물

구현회 당일, 멘토와 도우미멘토가 적극적으로 도와줍니다.

저희의 해커톤은 시험도 경합도 아닙니다. 참여자들이 한학기 동안 배운 내용을 잘 마무리할 수 있는 마지막 수업인 동시에 커리어를 위한 경험입니다. 최고의 경험이 될 수 있도록 멘토들이 적극적으로 도와줄 예정이니 걱정하지마세요 😊

최종 산출물은 다음과 같습니다.

1. 프로젝트 기획 카드  
프로젝트 개요서
2. 구현 코드  
주피터 노트북 파일, pycharm 파일 등 분석에 사용된 프로그래밍 코드 일체
3. 프로젝트 발표 자료  
5min preseantion
4. 프로젝트 보고서(option): 구현코드를 마크다운으로 정리할 것  
발표 이후 공유/보고/포트폴리오 용도로 사용

이 가이드를 확인하는 시점에 모든 팀은 기획카드 제작이 완료되었을 것 입니다.

**2번** 구현 코드와 **3번** 발표자료 제작은 필수이며, 더 유용하게 사용되고 널리 공유되길 원한다면 **4번** 보고서까지 완성하길 권장합니다. 프로젝트 보고서는 포트폴리오로 활용되어 커리어에 큰 도움이 될 수 있습니다.

\* 구현회에서 시간이 남는다면 모든 산출물을 제작해보세요.



## 팀별 피드백



### 피드백 순서

1. 팀 이름이 뭐가 중요해 주제가 중요하지 **서강대**
2. 버블브레이커 **고려대**
3. 아기코알라 **연세대**
4. 대흥동 이제마 **서강대**
5. 오후반 2 **서강대**
6. 순수혈통유희CO **연세대**
7. 방가방가 **서강대**
8. 정빈과 아이들 **연세대**
9. 코랑이 **고려대**
10. 통소여의 모험 **고려대**

### \* 난이도 설명

하: 코알라 유니브에서 배운 DS 지식만으로 제작할 수 있다.  
중: 코알라 유니브에서 배운 DS 지식과 자신의 도메인 지식을 연관지어야 한다.  
상: 코알라 유니브에서 배운 DS 지식 외에 것을 추가 학습해야 한다.

# 1 팀이름이뭐가중요해주제가중요하지




프로젝트 예상 난이도	하~중상																					
머신러닝 방향성	<div>분류 알고리즘 위주로 설계해보세요. 일반적으로 아래와 같은 방식이 사용됩니다.</div> <table><tr><th>특징</th><th>추천 레시피(Y)</th><th>냉장고에 김치가 있나?(X1)</th><th>양파(X2)</th><th>돼지고기 (X...)</th><th>믹서시가 있나?(Xn)</th><th>화덕이 있는가? (Xn+1)</th></tr><tr><td></td><td>김치찌개</td><td>1</td><td>1</td><td>1</td><td>0</td><td>0</td></tr><tr><td></td><td>김치피자</td><td>1</td><td>0</td><td>0</td><td>0</td><td>1</td></tr></table>	특징	추천 레시피(Y)	냉장고에 김치가 있나?(X1)	양파(X2)	돼지고기 (X...)	믹서시가 있나?(Xn)	화덕이 있는가? (Xn+1)		김치찌개	1	1	1	0	0		김치피자	1	0	0	0	1
특징	추천 레시피(Y)	냉장고에 김치가 있나?(X1)	양파(X2)	돼지고기 (X...)	믹서시가 있나?(Xn)	화덕이 있는가? (Xn+1)																
	김치찌개	1	1	1	0	0																
	김치피자	1	0	0	0	1																
제공될 도우미 소스파일	해당없음																					
관련 팀 프로젝트	팀이름이뭐가중요해주제가중요하지 / 오후반2																					
멘토 코멘트	<div>1. 위의 데이터 스케치를 이해하고 분류알고리즘을 테스트해보세요.</div> <div>2. 데이터를 잘 모으면 머신러닝 모델을 만드는 일은 어렵지 않습니다. 이후 어떠한 방식으로 추천할지, UI/UX 적인 면도 고려해보면 좋겠습니다.</div> <div>3. 저도 꼭 사용해보고 싶어요! 기대할게요:) by 자취생멘토</div>																					



## 2 버블브레이커

프로젝트 예상 난이도	하~상
머신러닝 방향성	<ol style="list-style-type: none"><li>1. 크롤링 후 얻게된 데이터를 정제하는 작업이 엄청난 이슈가 될 것 같습니다.</li><li>2. 카테고리 구성이 디테일하면 성능은 올라가지만 프로젝트 완성이 어려워질 수 있습니다. 적당한 정도에서 타협하세요. (20카테고리를 모두 사용하지 않고 3개로 줄인다든지)</li><li>3. 머신러닝 후 성능평가를 어떤식으로 할지 고안해보세요. 오차제곱평균, 합, 결정계수 등과 같은 CS적인 방법도 좋지만 실제 가격과 예상 가격간의 차이를 나타내는 기발한 방법을 고안해보세요.(정규분포와 같은 수리통계적 지식 사용도 좋음)</li></ol>
제공될 도우미 소스파일	해당없음
유사 프로젝트	버블브레이커 / 순수혈통유희co / 정빈과 아이들
멘토 코멘트	<ol style="list-style-type: none"><li>1. 데이터 가공 작업에서 난항이 예상됩니다. 파이팅! (힘들긴 하지만 할 수 있을거라 믿어요)</li><li>2. 분류가 적을 때는 영문feature를 숫자로 바꿔주는 순서가 상관없지만(male, female), 많을 때 (geforce100, 200, 300, ... 9900)는 값의 순서가 중요할 수 있습니다. (성능순으로 0,1,...989)</li><li>3. decision tree 외의 알고리즘을 추천합니다.</li><li>4. 특정 카테고리의 범위가 유난히 클 경우(0~20000) 스케일링 작업이 필요할 수 있습니다.(성능을 높이려면 필요, 없어도 어느정도 작동)</li><li>5. 거품이 크다는 것은 시일내에 가격이 내릴 가능성이 높다는 뜻으로 해석됩니다. 알고리즘 트레이딩을 하는 커머스 회사에서 관심 가지말 한 내용으로 높은 활용성이 기대됩니다. (용산 사장님만해도 바로 관심 가질 거 같아요) 그런 내용도 발표에 포함할 수 있겠네요.</li></ol>

### 3 아기코알라

프로젝트 예상 난이도	중~중상
머신러닝 방향성	Feature Engineering까지의 과정이 험난할 것 같습니다. 대표 색상을 찾아내고 해당 색상으로 이미지 데이터의 특징을 대신하는 작업을 pandas로 구현하는 과정에서 시행착오가 예상됩니다. 따라서 대표 색상 picker 작업을 어떤 수준까지 진행할지 정하는게 중요할 것 같습니다.
제공될 도우미 소스파일	 <code>image-color-picker.ipynb</code>  드라이브에 공유되었습니다.
유사 프로젝트	-
멘토 코멘트	<p>1. 유일하게 이미지 프로세싱을 다루는 팀입니다. 대표 색상 피킹 방법은 아래 세가지 중 하나를 권장합니다.              쉬운버전) 전체 픽셀의 평균색상값 - 공유된 파일              개선된버전) image를 blur처리 한 후 최빈 픽셀 색상값 (+ 픽셀 구역화도 가능)              ML버전) 머신러닝 kmeans 알고리즘 이용  <a href="https://zeevgilovitz.com/detecting-dominant-colours-in-python">https://zeevgilovitz.com/detecting-dominant-colours-in-python</a>  <a href="https://www.pyimagesearch.com/2014/05/26/opencv-python-k-means-color-clustering/">https://www.pyimagesearch.com/2014/05/26/opencv-python-k-means-color-clustering/</a></p> <p>2. 사실 이미지의 대표색상은 1개가 아닌 경우가 많습니다. 고민해보세요.              (대표색상은 빨간색일까요? 파란색일까요?)</p>  <p>3. 색상 선택결과 생각과 많이 다르다면? (2번의 문제일 가능성이 높습니다.)</p> <p>4. 일반적으로 ‘상을 받는 영화’와 ‘흥행하는 영화’는 다르다고 알려져 있습니다. 모델이 만들어진 후, 상을 받은 영화들의 포스터로 장르를 분석했을 때의 예측정확도와 대중영화의 예측 정확도에서 차이가 있을 수 있을까요? 저도 궁금하네요. (꼭 엄청난 활용분야를 찾을 필요 없이 재미있는 분석 결과만 제공해줘도 유용할 것 같습니다. 발전시키면, “(반전)성공하는 스릴러 영화의 포스터는 녹색이었다!” 라는 칼럼을 낼 수도 있지 않을까요? 업계 관계자에게는 유익하고 재미있는 칼럼이 될 것 같습니다.)</p>

## 4 대흥동 이제마

프로젝트 예상 난이도	중
머신러닝 방향성	직접 데이터를 만들어야 하므로 데이터 준비작업 면에서의 노력이 필요해보입니다. 사실 크롤링 하는 다른 팀들보다 더 빨리 끝낼 수 있습니다.
제공될 도우미 소스파일	해당없음
유사 프로젝트	-
멘토 코멘트	<div>1. 설문조사에 참여하였습니다. 잘 준비되고 있는게 눈에 보입니다. 재미있는 프로젝트가 될 것 같아 기대하고 있습니다.</div> <div>2. 프로젝트 마무리 단계에서, ‘한의사인 나 보다 더 잘맞추네?’와 같은 인터뷰를 진행하거나, 다른 창의적인 방식으로 ‘주관적으로 진단내릴 수 밖에 없었던 기존의 방식’보다 나아진 모델의 실제 영향력을 확인할 수 있는 자료가 있으면 좋겠습니다.</div> <div>3. 빨리 제 체질 확인 좀..!</div>



## 5 오후반2

프로젝트 예상 난이도	하~중																								
머신러닝 방향성	<div>다음과 같은 데이터 구성이 한가지 방법이 될 수 있습니다.</div> <table><tr><th>특징</th><th>흥행여부</th><th>유해진</th><th>유아인</th><th>아이유</th><th>...</th></tr><tr><td></td><td>썩박</td><td>1</td><td>1</td><td>1</td><td>0</td></tr><tr><td></td><td>대박</td><td>1</td><td>0</td><td>0</td><td></td></tr><tr><td></td><td>손익분기</td><td>0</td><td>1</td><td>0</td><td></td></tr></table>	특징	흥행여부	유해진	유아인	아이유	...		썩박	1	1	1	0		대박	1	0	0			손익분기	0	1	0	
특징	흥행여부	유해진	유아인	아이유	...																				
	썩박	1	1	1	0																				
	대박	1	0	0																					
	손익분기	0	1	0																					
제공될 도우미 소스파일	해당없음																								
유사 프로젝트	팀이름이뭐가중요해주제가중요하지 / 오후반2																								
멘토 코멘트	<div>1. 관객수를 기준으로 할지 흥행여부를 기준으로 할지 고민해봐야 할 것 같습니다. 둘 모두 장단점이 있어보입니다.</div> <div>2. 학습 후 랜덤으로 조합을 구성해 흥행도를 측정하려는 것 같습니다. 100C10(nCr)은 17310309456440(17조)이므로 어떻게 조합을 구성할지 제대로 계획해야 시간내에 예측할 수 있겠습니다.</div> <div>3. 주연과 조연의 구분을 어떻게 가져갈건지 고민해보세요.</div> <div>4. 오후반2에서 생각하고 있는 재미있는 흥행 insight가 많을 것 같습니다. 영화 관계자들을 놀라게 하는(?) 프로젝트 기대할게요.</div>																								

6 순수혈통유희co

프로젝트 예상 난이도	하~중
머신러닝 방향성	데이터 조건이 좋고 목적이 뚜렷하여 우수한 성능의 머신러닝 모델이 기대됩니다.
제공될 도우미 소스파일	-
유사 프로젝트	버블브레이커 / 순수혈통유희co / 정빈과 아이들
멘토 코멘트	<div>1. 예상처럼 리뷰 수가 많을 수록, 도심에서 위치가 가까울 수록 가격이 높을까요? 가격이 싸기 때문에 리뷰수가 많은 경우도 적지 않을 것 같습니다. 사실 저도 궁금해요. 분석 후 결과 공유 기대할게요.</div> <div>2. airbnb의 호스트가 되려고 할 때 뿐만 아니라, 국내에서 특히 게스트하우스 상권과 지방에서 민박값을 쉽게 정하지 못하는 사람들에게 큰 도움이 될 것 같습니다. 사회적으로 큰 도움이 될 수 있을 것 같아요. 여러 가지 활용방안, 활용처를 제안해주시면 좋겠습니다.</div> <div>3. 꼭 잘 만드셔서 대박나는 노다지 좀 추천해주세요:)</div>

## 7 방가방가

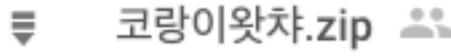

프로젝트 예상 난이도	하~상
머신러닝 방향성	크롤링과 분석법 자체는 우리가 많이 다루워왔던 그 방식 그대로입니다. 다만 가방의 상세조건을 수집하는 과정이 생각보다 까다로울 것 같습니다.
제공될 도우미 소스파일	 코랑이왓차.zip  드라이브에서 확인해주세요.
유사 프로젝트	방가방가 / 코랑이
멘토 코멘트	<ol style="list-style-type: none"><li>조건 입력 시 가방가격을 예측 해주는 방식은 크게 어려울 것 같지 않습니다. 구현회 당일 멘토들에게 질문하여 HTML페이지로 만들어도 재밌을 거 같아요.</li><li>해당 조건과 해당 가격대의 가방 모델을 추천하는 일을 정의해보세요. 해당 조건의 가방을 그대로 보여주는 것이라면 단순 listing, 해당 조건의 가방 중 이 사람에게 맞는 가방, 혹은 인기 가방을 보여주는 것이라면 추천에 해당합니다. 만약 추천을 생각하는거라면 위의 도우미 소스파일을 확인하고 추천시스템을 사용해 볼 수 있습니다. 다른 창의적인 방법도 가능합니다. 사실 실제 업계에서 매우 원시적이고 허접한 방법으로도 추천을 하고 있습니다. 본인이 만든 추천 알고리즘이 부족해보여도 충분히 가치가 있으니 자신감을 가지고 만들어보세요.</li><li>제 가방은 바자회 수제 제품이라 해당없겠네요.. ㅎㅎ</li></ol>





## 8 정빈과 아이들

프로젝트 예상 난이도	하~중
머신러닝 방향성	데이터와 머신러닝의 특성을 잘 이해한 프로젝트로 우수한 성능의 머신러닝모델이 기대됩니다. 다만 비상장 회사와 상장회사는 지표의 차이가 크고 공유하지 않는 항목이 있으므로 고려바랍니다.
제공될 도우미 소스파일	-
유사 프로젝트	버블브레이커 / 순수혈통유희co / 정빈과 아이들
멘토 코멘트	<ol style="list-style-type: none"><li>1. (프로젝트 기획카드에서) 상장 기업을 train한 후 비상장 기업을 test set으로 평가할 수 없습니다. 상장기업을 train 한 후 한번도 train 하지 않은 다른 상장기업으로 test하는 것이고, 비상장기업 평가는 모델 완성 후 실세계에 활용하는 범위입니다.</li><li>2. 이번 univ 해커톤 프로젝트 중에서 가장 격식있는(?) 프로젝트입니다. 가장 formal한 데이터를 사용하고 분석결과도 그렇습니다. 조금 딱딱해보여도 고급스러운 리포트를 만드는게 좋은 방법이 될 것 같아요.(증권가 분석처럼)</li><li>3. 정빈과 아이들 = 어벤져스</li></ol>

## 9 코랑이

프로젝트 예상 난이도	중~상
머신러닝 방향성	<ol style="list-style-type: none"> <li>1. 데이터 수집시 별점정보 부분이 까다로울 것 같습니다.(참고 코드 제공)</li> <li>2. kuklue 평가방식에 문제가 있을 수 있습니다. (난이도가 별1개이면 쉽다는 건지, 만족도가 낮다는 건지 헷갈린 유저들이 많이 있을것 같습니다.) 감안해주세요.</li> </ol>
제공될 도우미 소스파일	<ol style="list-style-type: none"> <li>1. 추천시스템   </li> <li>2. 별점 크롤러    (기획카드 엑셀파일 탭) </li> </ol>
유사 프로젝트	방가방가 / 코랑이
멘토 코멘트	<ol style="list-style-type: none"> <li>1. 총평을 예측하는 머신러닝의 난이도는 높지 않습니다. 그 이후 Insight을 찾거나 추천하는 과정은 난이도가 높을 수 있습니다.</li> <li>2. '만족스러운 교양과목을 예측할 수 있다.'는 사실 총평 4~5점인 강의를 선택하는게 최선이지 않을까 생각됩니다. 이 모델로 할 수 있는 더 의미있는 일이 많이 있을테니 창의적으로 고안해보세요.</li> <li>3. 가장 좋은 방법은 스스로에게 '나는 꿀 교양을 찾을 때 어떻게 하지?'를 묻고 그 과정을 글로 적은 후 기계화 할 수 있는지 고민해보는 것 입니다.</li> <li>4. 포함된 추천시스템 도우미 소스파일을 이용할 수 있겠지만 굳이 그렇게 하지 않고, 조금 어설픈 추천모델을 만들더라도 충분히 의미 있습니다. (사실 실제 유명서비스들의 일부 추천시스템은 수준이 매우 낮습니다.)</li> <li>5. HTML 파일로 디자인된 교양 수업 추천페이지도 있으면 재미있을것 같습니다.</li> <li>6. 하지만 이 모든걸 하려면? 프로젝트 규모가 너무 커지겠죠. 팀원들간 합의하여 프로젝트의 범위를 잘 정해보시기 바랍니다.</li> <li>7. 어쩌면 가장 유명한 고대인이 될 수도?</li> </ol>

## 10 통소여의 모험

프로젝트 예상 난이도	중~상
머신러닝 방향성	<ol style="list-style-type: none"> <li>1. 코알라유니브에서 다뤄보지 않은 텍스트 분석을 다루는 팀입니다. 제공된 코드 사용은 어렵지 않지만 익숙하지 않기에 시행착오는 분명 따라옵니다.</li> <li>2. stop word에 대해 고려해봅니다. (검색 + 도우미 소스파일 주석 참고)</li> <li>3. 남성장가와 여성작가의 차이가 없어졌다고 볼 수도 있지만, 모델의 예측력이 딸려 차이를 밝히지 못하는 것일 확률도 낮지 않습니다. 감안해주세요.</li> </ol>
제공될 도우미 소스파일	<div>  <span>텍스트분석하기.ipynb</span>  </div> <p>기획카드가 있는 구글 드라이브에 [참고자료] 폴더 내에 있음</p>
유사 프로젝트	-
멘토 코멘트	<ol style="list-style-type: none"> <li>1. 유일하게 텍스트 분석을 하는 팀입니다. 팀원들의 도전정신을 높이 평가합니다.</li> <li>2. train/valid에서 현대이전의 문학작품들을 활용한다면, 모델의 성능을 말하는 test도 마찬가지로 현대이전의 문학작품들을 활용해야합니다. 이때 정확도가 높아 모델을 신뢰할 수 있으면 활용범위를 늘려 현대의 작품에도 적용해볼 수 있습니다. 만약 정확도가 낮아 그 차이가 현저하면 성별별 언어사용이 시대에 따라 변화했다는 설명력을 가질 수 있습니다.</li> <li>3. 기획대로 현대 이전의 문학작품만을 대상으로 학습할 것이라 예상됩니다. 시대에 관계 없이 모든 작품을 학습한 결과도 만들어 실제로 모델간 차이가 있는지도 확인해볼 필요가 있습니다.</li> <li>4. 실제로 한글 작업을 진행할 필요는 없지만, 발표자료에서는 한글 케이스 적용시의 결과도 예상하거나 언급해주면 좋을 것 같습니다.</li> <li>5. 준비클래스 끝나고서도 남아서 고민하는 모습 감동받았어요;) 마지막도 잘 부탁드립니다.</li> </ol>