

Readability



Agenda

01

연구 배경

주제를 선정하게 된 배경을 설명합니다.

02

연구 목표

얻고자 하는 결과를 위해 정한 목표를 설명합니다.

03

연구 내용

결과를 얻고자 노력한 과정을 설명합니다.

04

연구 결과

얻은 결과를 해석적으로 설명합니다.

05

연구 시사점 및 질의 응답

연구가 제시하는 의의를 살펴보고, 간단한 질의응답을 받습니다.



연구 배경

Readability를 선택한 이유





Kaggle, 트럼프 대통령의 Tweet

위와 같이, 정치인, CEO, 연예인, 인플루언서 등
다양한 사람들이 SNS를 활용하여 여러 의견을 표하는 시대

IT 시대 여러 정보를 효과적으로 받거나 주기 위해선
가독성은 매우 중요

Readability



가독성 평가 수식

01

FK Level (Flesch Kincaid Grade Level) – 미국 국방부에서 사용

영어 지문의 가독성을 정의된 수식으로 계산하여 **미국의 학년**으로 나타낸 것
(FK Level을 계산한 결과로 값이 5가 나왔다면, 그 텍스트는 미국의 5학년 학생 수준이 읽을 수 있는 텍스트)

02

G.F.I.(Gunning Fog Index)

해당 텍스트를 읽기 위해 얼마나 교육을 받아야 하는지 나타낸 것
(학술문서의 수치는 18로, 18년간의 교육이 필요하다는 의미)

03

FRE Score (Flesch Reading Ease Score)

영어 지문의 가독성을 정의된 수식으로 계산하여, 해당 문장을 **어느 정도 수준의 사람**이 **읽을 수 있는 지** 나타낸 것
(FRE Score는 대학 졸업자 및 분야별 전문가 수준까지로 나타낼 수 있으며, 점수에 따른 수준)



가독성 평가 수식

FK Level (Flesch Kincaid Grade Level)


$$FK Level = 0.39 \left(\frac{Total\ words}{Total\ sentences} \right) + 11.8 \left(\frac{Total\ syllables}{Total\ words} \right) - 15.59$$

G.F.I.(Gunning Fog Index)

$$GFI = 0.4 \left[\left(\frac{Total\ words}{Total\ sentences} \right) + 100 \left(\frac{Complex\ words}{total\ words} \right) \right]$$

FRE Score (Flesch Reading Ease Score)

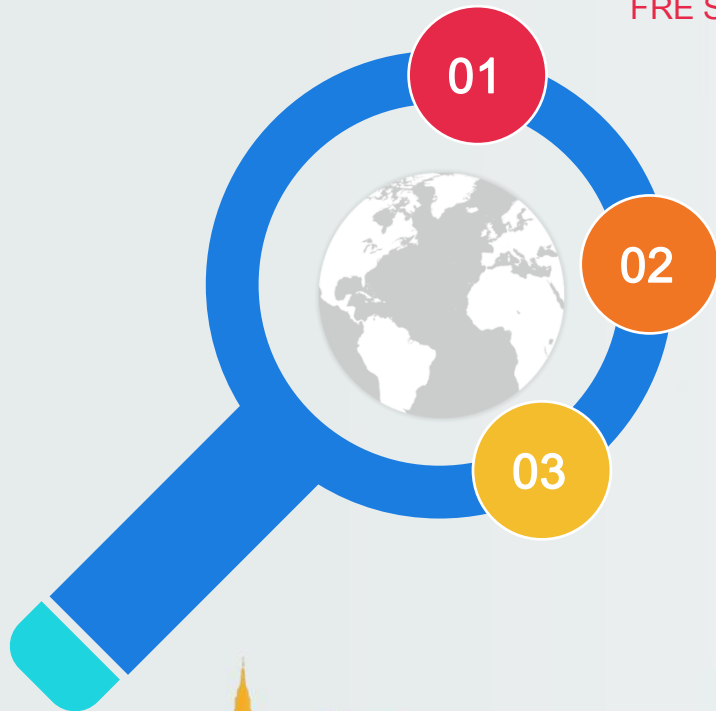
$$FRE Score = 206.835 * 1.015 \left(\frac{Total\ words}{Total\ sentences} \right) - 8.46 \left(\frac{Total\ syllables}{Total\ words} \right)$$



Score	Level
100 ~ 90	5th grade
90 ~ 80	6th grade
80 ~ 70	7th grade
70 ~ 60	8th grade, 9th grade
60 ~ 50	10th to 12th grade
50 ~ 30	College
30 ~ 10	College graduate
10 ~ 0	Professional

가독성 평가 수식

FRE Score (Flesch Reading Ease Score)



Score	Level
100 ~ 90	5th grade
90 ~ 80	6th grade
80 ~ 70	7th grade
70 ~ 60	8th grade, 9th grade
60 ~ 50	10th to 12th grade
50 ~ 30	College
30 ~ 10	College graduate
10 ~ 0	Professional



연구 목표 및 방법

Readability로 얻고자 하는 것



Our Goal with Readability



영미권 사람들의 문장력 수준 파악

- FK LEVEL
- G.F.I. (Gunning Fog Index)
- FRE Score
- 40명의 인원, (SNS, 연설문 | 분야별 10명)



송실대 학생들의 문장력 수준 파악 (Eng)

- 불특정다수 51명 (칼럼, 에세이)

(크롤링 : py)



문장력 평가 지표 간의 상관관계

- 영미권 사람들의 Data를 바탕으로
- 선형 회귀 함수로 3개 간의 관계를 나타낼 함수



학생들의 Data로 함수의 정확도 파악

- 사이트에 표기된 수치와, 함수 대입 값의 차이를 비교, 분석

(MATLAB)



연구 내용

연구의 과정 및 시행착오



Readability Calculator



This free online software tool calculates readability : Coleman Liau index, Flesch Kincaid Grade Level, ARI (Automated Readability Index), SMOG. The measure of readability used here is the indication of number of years of education that a person needs to be able to understand the text easily on the first reading. Comprehension tests and skills training. This tool is made primarily for English texts but might work also for some other languages. In general, these tests penalize writers for polysyllabic words and long, complex sentences. Your writing will score better when you: use simpler diction, write short sentences. It also displays complicated sentences (with many words and syllables) with suggestions for what you might do to improve its readability.

Basic text statistics are also displayed, including number of characters, words, sentences, and average number of characters per word, syllables per word, and words per sentence.

Enter text (copy and paste is fine) here:

or read it from a website (only plain text .TXT) :

Process text

평가 지표를 제공해주는 웹사이트
파이썬을 이용한 크롤링

https://www.online-utility.org/english/readability_test_and_improve.jsp

```

1 from selenium import webdriver
2 from selenium.webdriver.common.keys import Keys
3 import os
4 import csv
5 import time
6
7 def cal_text(path) :
8     driver = webdriver.Chrome('D:\Readability\chromedriver')
9     driver.implicitly_wait(3)
10
11     fp = open(path, mode = 'r', encoding = 'utf-8')
12     text = fp.read()
13     fp.close()
14
15     input_text = []
16     result = [0] * 3
17
18     input_text.append(text)
19
20     for input in input_text :
21         driver.get('https://www.online-utility.org/english/readability_test_and_improve.jsp')
22         time.sleep(3)
23         driver.find_element_by_class_name('bigtextarea').send_keys(input)
24         time.sleep(3)
25         driver.find_element_by_xpath('/html/body/form/input[2]').click()
26         time.sleep(3)
27
28         xpath_result = ['/html/body/table[3]/tbody/tr[2]/td[2]', # Gunning Fog Index
29                         '/html/body/table[4]/tbody/tr[3]/td[2]', # Fk Level
30                         '/html/body/table[5]/tbody/tr[2]/td[2]'] # Fre
31
32         temp_result = list()
33
34         for i in xpath_result :
35             temp_result.append(driver.find_element_by_xpath(i).get_attribute('textContent'))
36             time.sleep(3)
37
38         for j in range(0,3) :
39             result[j] += int(float(temp_result[j]) / len(input_text))

```

```

40
41         driver.close()
42         return result
43
44     def save_csv(part, name, data) :
45         csv_name = 'D:\Readability\Result//' + part + '.csv'
46
47         if os.path.isfile(csv_name) :
48             csv_fp = open(csv_name, 'a', newline='')
49             wr = csv.writer(csv_fp)
50             wr.writerow([name,data[0],data[1],data[2]])
51             csv_fp.close()
52
53         else :
54             csv_fp = open(csv_name, 'w', newline='')
55             wr = csv.writer(csv_fp)
56             wr.writerow(['', 'GFI', 'Fk Level', 'Fre'])
57             wr.writerow([name,data[0],data[1],data[2]])
58             csv_fp.close()
59
60     start_path = 'D:\Readability\Data'
61     dir_list = os.listdir(start_path)
62     file_list = []
63
64     for dir in dir_list :
65         file_list = os.listdir(start_path + '/' + dir)
66         for file in file_list :
67             data = cal_text(start_path + '/' + dir + '/' + file)
68             name = file.replace(".txt", "")
69             save_csv(dir, name, data)

```

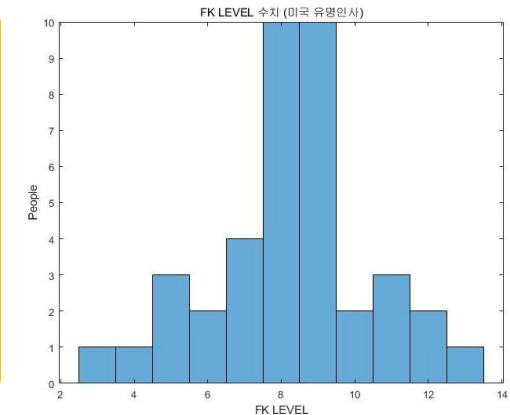
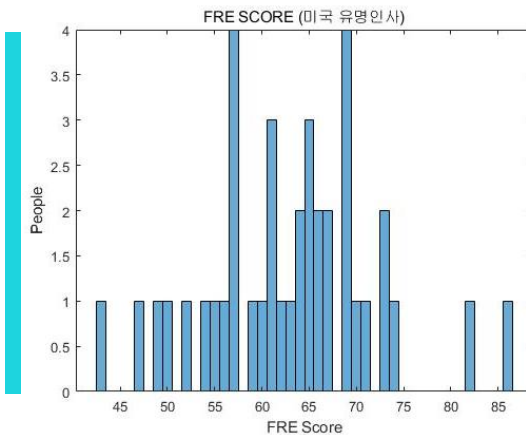
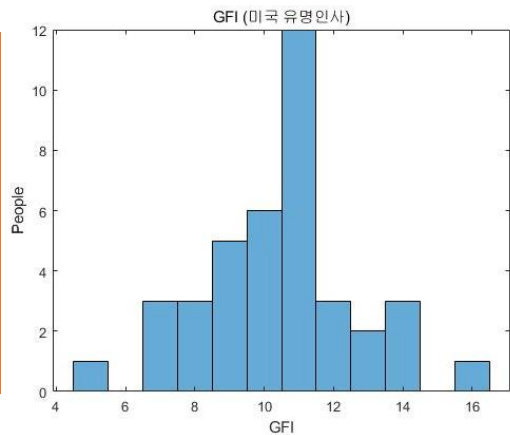
크롤링 코드

미국인 분야별 유명인사 문장력 지표

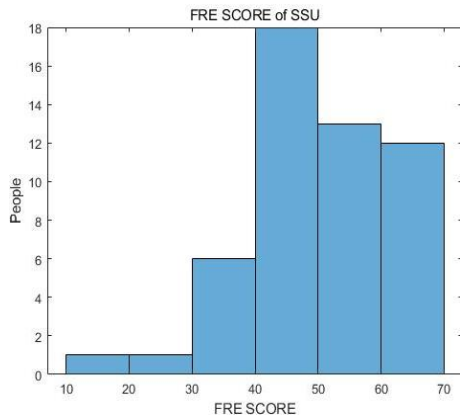
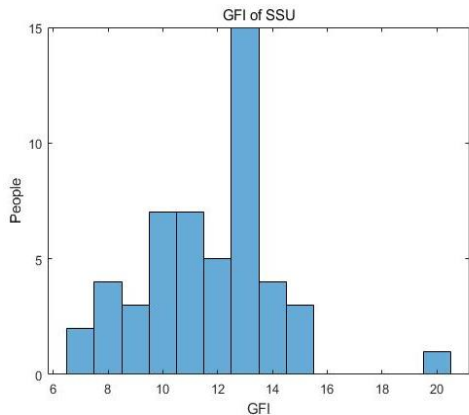
GFI 수치
동일 40명 대상
GFI 평균 : 10.4359

FK Level 수치
동일 40명 대상
FK LEVEL 평균 : 8.2564

FRE Score 수치
동일 40명 대상
FRE 평균 : 62.9487



분야 : CEO, 정치가, 연예인, 인플루언서 각 10명 (SNS, 연설문, 인터뷰, 에세이) | int 형 기준



송실대 학생 문장력 지표

GFI 수치

동일 51명 대상

GFI 평균 : 11.6863

FK Level 수치

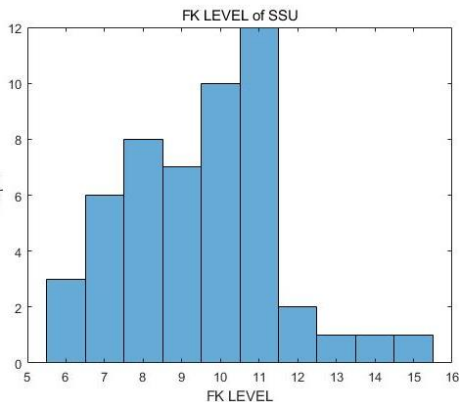
동일 51명 대상

FK LEVEL 평균 : 9.5098

FRE Score 수치

동일 51명 대상

FRE 평균 : 49.7843



미국인 vs 숭실대 학생 문장력 수준 비교

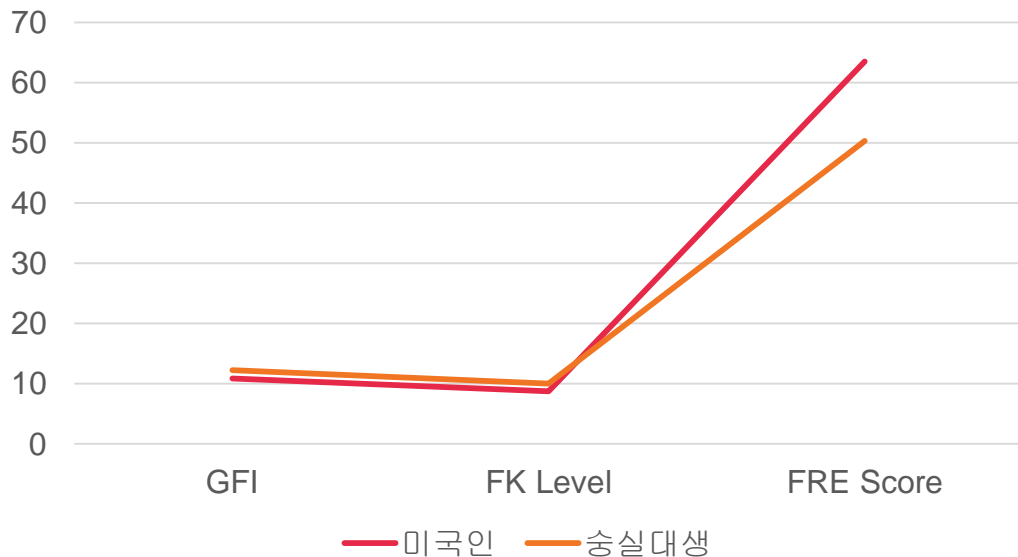
미국인	숭실대생
GFI 수치 GFI 평균 : 10.4359	GFI 수치 GFI 평균 : 11.6863
FK Level 수치 FK LEVEL 평균 : 8.2564	FK Level 수치 FK LEVEL 평균 : 9.5098
FRE Score 수치 FRE 평균 : 62.9487	FRE Score 수치 FRE 평균 : 49.7843

미국인 vs 숭실대 학생 문장력 수준 비교

미국인	숭실대생
GFI 수치 GFI 평균 : 10.8818	GFI 수치 GFI 평균 : 12.2325
FK Level 수치 FK LEVEL 평균 : 8.6985	FK Level 수치 FK LEVEL 평균 : 10.0502
FRE Score 수치 FRE 평균 : 63.4795	FRE Score 수치 FRE 평균 : 50.3200

미국인 vs 숭실대 학생 문장력 수준 비교

문장력 비교



모집단의 수가 제한적

데이터의 차이 (다방면 vs 칼럼위주)

크롤링 사이트의 인식률 차이

평가 지표 개수가 제한적

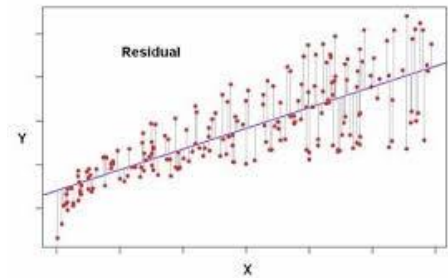
Else...



활용1 각 평가 지표 간의 상관성 파악

선형회귀분석

변수 간의 인과관계를 설명하기 위함



최소 제곱법을 이용하여 기울기 a 와 y 절편 b 를 구함
그래프 간의 오차가 가장 최소가 되는 $f(x)$ 를 찾아줌

$$a = \frac{\sum_{i=1}^n (x - \text{mean}(x))(y - \text{mean}(y))}{\sum_{i=1}^n (x - \text{mean}(x))^2}$$

$$b = \text{mean}(y) - (\text{mean}(x) \cdot a)$$

최소 제곱법을 이용한 회귀 계수 추정

활용1 각 평가 지표 간의 상관성 파악

회귀 모형 평가 방식

RMSE와 결정계수

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

활용2 송실대 학생 DATA를 통한 모델의 정확도 파악

FK Level, FRE Score와 관련하여

실제 수치 대입 값(크롤링 값)과
미국인 DATA로 만든 함수에 대입한 예측 값
비교 / 분석

선형회귀함수

```
function [a, r2] = linregr(x,y)
```

```
n= length(x);
if length(y) ~= n, error('x and y must be same length'); end
x = x'; y = y';
sx = sum(x); sy = sum(y);
sx2 = sum(x.*x); sy2 = sum(y.*y); sxy = sum(x.*y);
a(1) = (n*sxy - sx*sy)/(n*sx2 - sx^2);
a(2) = sy/n - a(1)*sx/n;
r2 = ((n*sxy - sx*sy)/sqrt(n*sx2 - sx^2)/sqrt(n*sy2 - sy^2))^2;
xp = linspace(min(x), max(x), 2);
yp = a(1)*xp + a(2);
plot(x,y,'o',xp,yp);
grid on

end
```

함수 표기

```
---fuc.m---
```

```
function [y] = fuc(x,sx,sy) % 기존의 x축, 새로운 x축, 새로운 y축 입력 - 그래프 출력
```

```
y = 0.9524*x -1.6649 %이는 1,2,3번에서 직접 뽑은 기울기와 y절편으로 수정할 것.
plot(x,y);
grid on
hold on
scatter(sx,sy)
end
```

```
---파일에서 데이터 추출 코드---
```

```
CEO=CEO(:,:); % 표를 행렬로 변경
Entertainer=Entertainer(:,:);
Influence=Influence(:,:);
Politician=Politician(:,:);
SSU=SSU(:,:);
x=cast(vertcat(CEO(:,2),Entertainer(:,2),Influence(:,2),Politician(:,2)), 'double'); % 각 파일에서 데이터를 뽑아내 합침
y=cast(vertcat(CEO(:,3),Entertainer(:,3),Influence(:,3),Politician(:,3)), 'double');
z=cast(vertcat(CEO(:,4),Entertainer(:,4),Influence(:,4),Politician(:,4)), 'double');
sx=cast(SSU(:,2), 'double');
sy=cast(SSU(:,3), 'double');
sz=cast(SSU(:,4), 'double');
```

```
---각 수치의 평균---
```

```
mean(x)
mean(y)
mean(z)
mean(sx)
mean(sy)
mean(sz)
```

```
---각 미국의 데이터를 시각화---
```

```
linregr(x,y)
linregr(x,z)
linregr(y,z)
```

```
---출력대상의 데이터를 추출한 선형회귀 그래프와 함께 시각화---
```

```
fuc(x,sx,sy)
fuc(x,sx,sz)
fuc(y,sy,sz)
```

```
---평균제곱근 오차 구하기---
```

```
sqr(sqrt(sum(pow2(sy-(sx*??+??)))/size(sx,1))); % 해당하는 그래프의 기울기와 y절편 대입
sqr(sqrt(sum(pow2(sy-(sx*??+??)))/size(sx,1)));
```

```
---2차 선형회귀함수 구하기---
```

```
linregr(sx*??+??,sy) % 해당하는 그래프의 기울기와 y절편 대입
linregr(sx*??+??,sz)
```

매트랩 코드

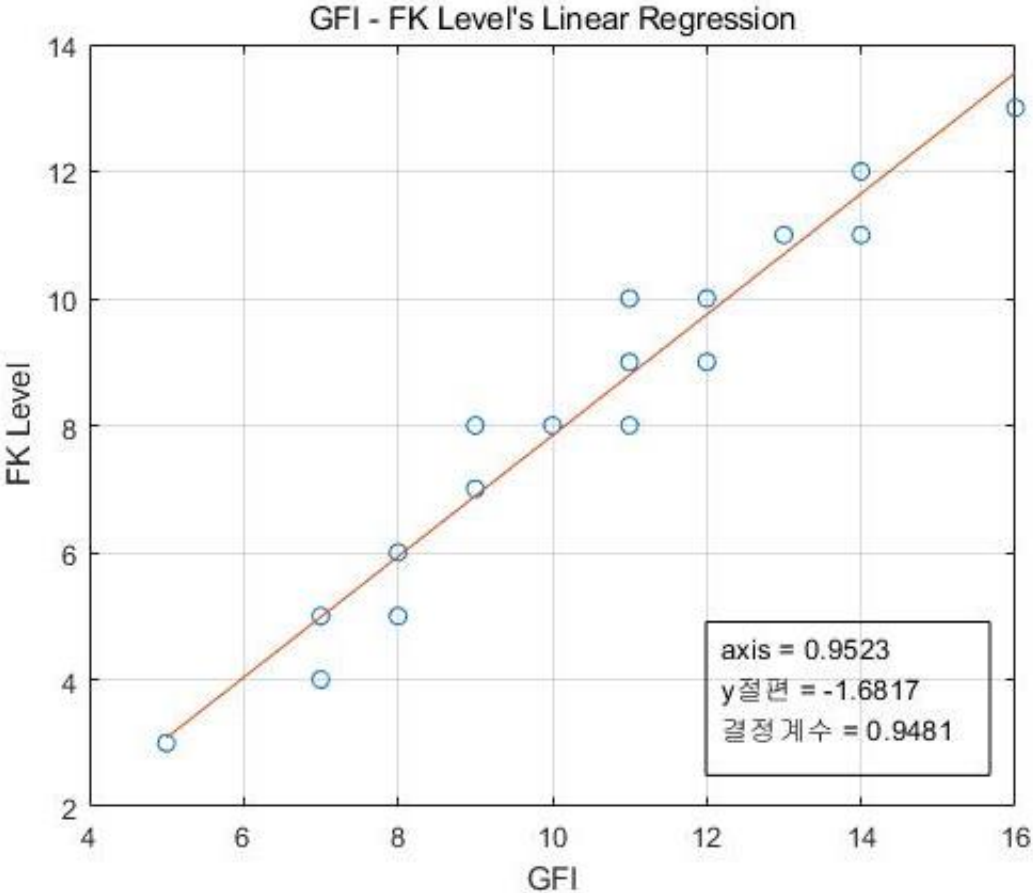


연구 결과

연구 결과를 설명



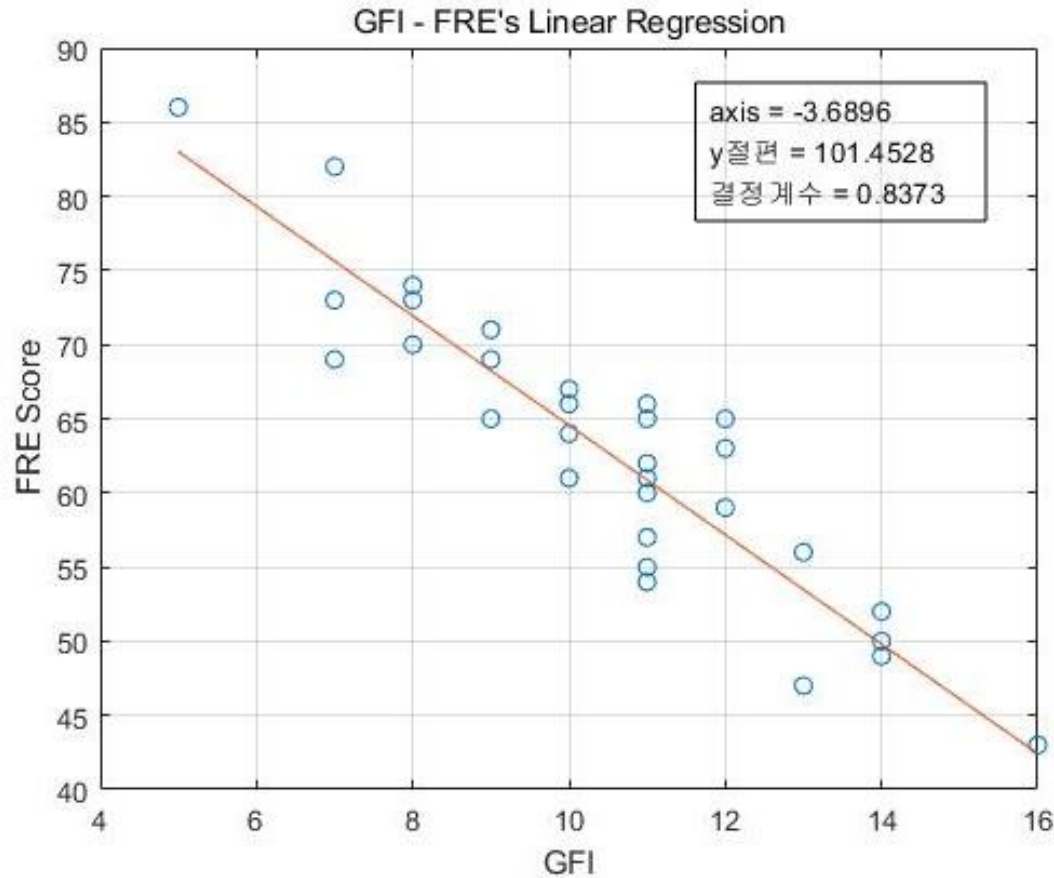
활용1 각 평가 지표 간의 상관성 파악



미국인 문장력 지표 GFI – FK 간의 선형회귀함수

GFI – FK 간의 (float) 회귀 계수와 결정계수	
X	GFI
y	FK Level
Axis	0.9524
Y절편	-1.6649
결정계수	0.9847
RMSE	1.0947

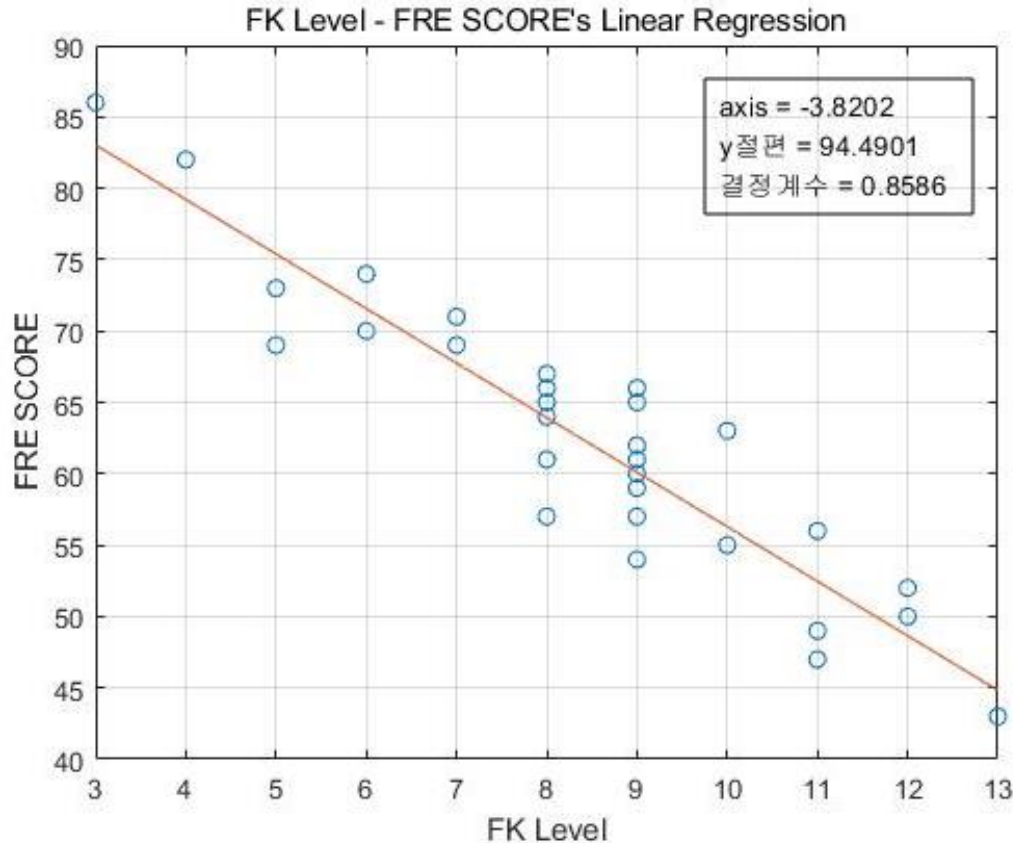
미국인 문장력 지표 GFI - FRE 간의 선형회귀함수



GFI - FRE 간의 (float)
회귀 계수와 결정계수

X	GFI
y	FRE Score
Axis	-3.6723
Y절편	103.4405
결정계수	0.8703
RMSE	0.9169

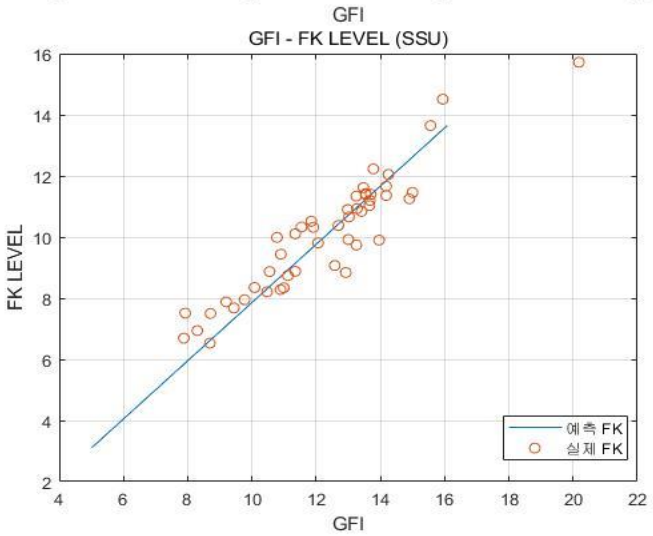
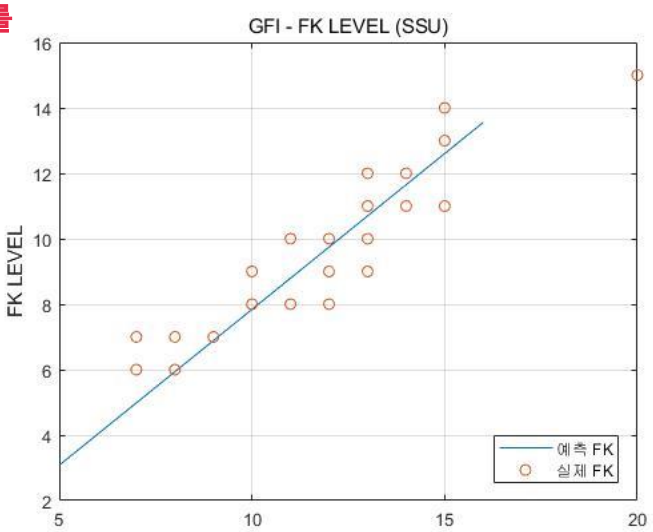
미국인 문장력 지표 FK – FRE 간의 선형회귀함수



FK – FRE 간의 (float)
회귀 계수와 결정계수

X	GFI
y	FK Level
Axis	-3.8694
y절편	97.1371
결정계수	0.8899

활용2 숭실대 학생 DATA를
통한 모델의 정확도 파악



숭실대생 문장력 지표 예측 값과 실제 값 비교 GFI - FK

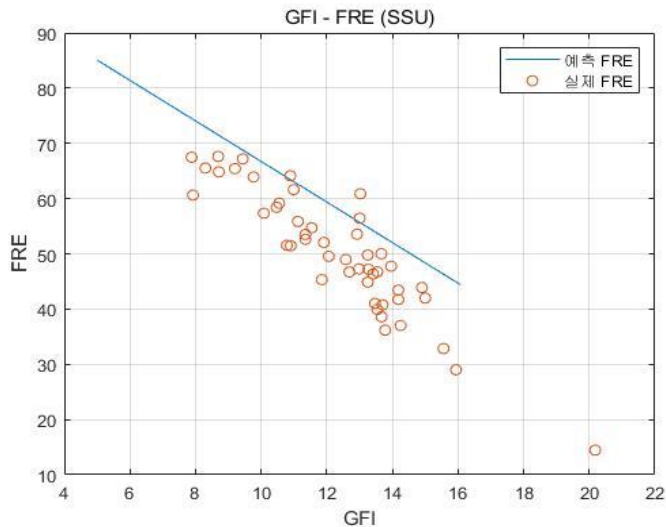
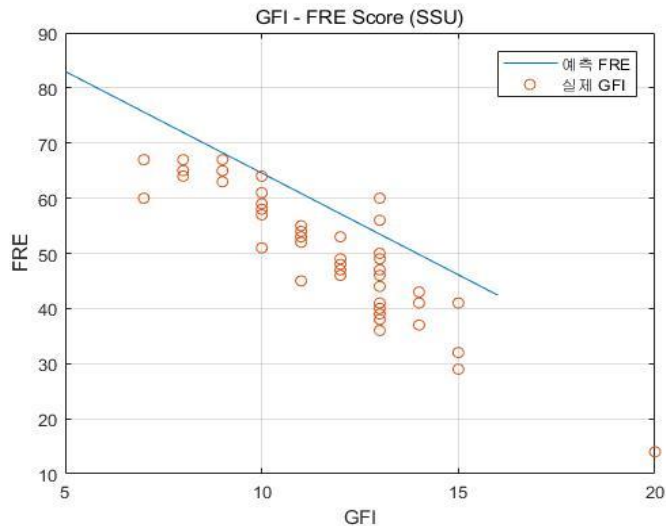
Int형 - 선형회귀 함수

$$y = 0.9523x - 1.6817$$

float형 - 선형회귀 함수

$$y = 0.9524x - 1.6649$$

송실대생 **문장력 지표** 예측 값과 실제 값 비교 GFI - FRE



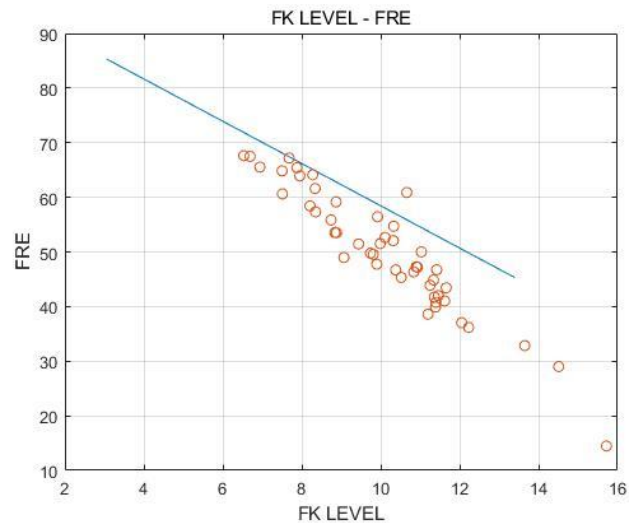
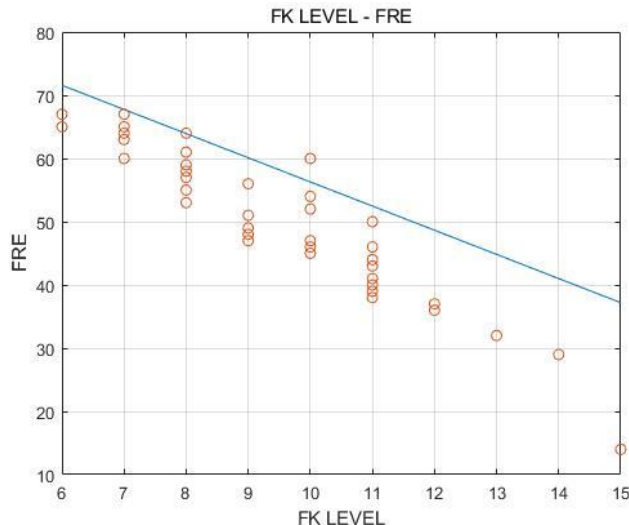
Int형 - 선형회귀 함수

$$y = -3.6896x + 101.4528$$

float형 - 선형회귀 함수

$$y = -3.6723x + 103.4405$$

송실대생 **문장력 지표** 예측 값과 실제 값 비교 FK - FRE



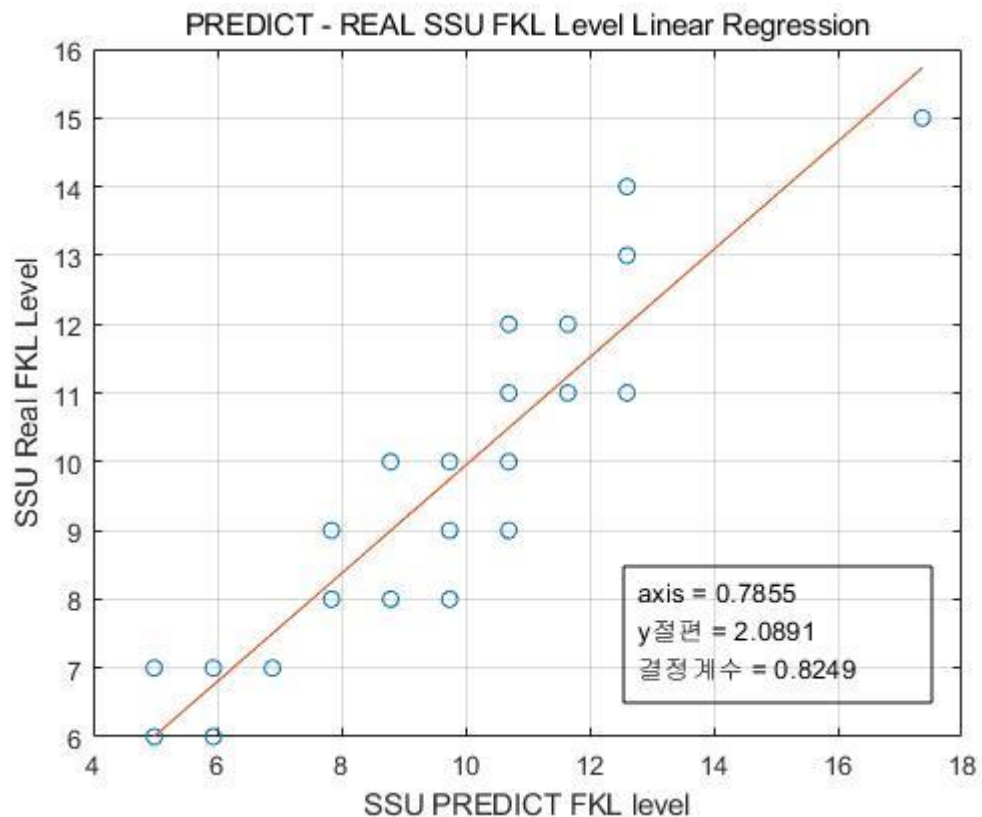
Int형 - 선형회귀 함수

$$y = -3.8202x + 94.4901$$

float형 - 선형회귀 함수

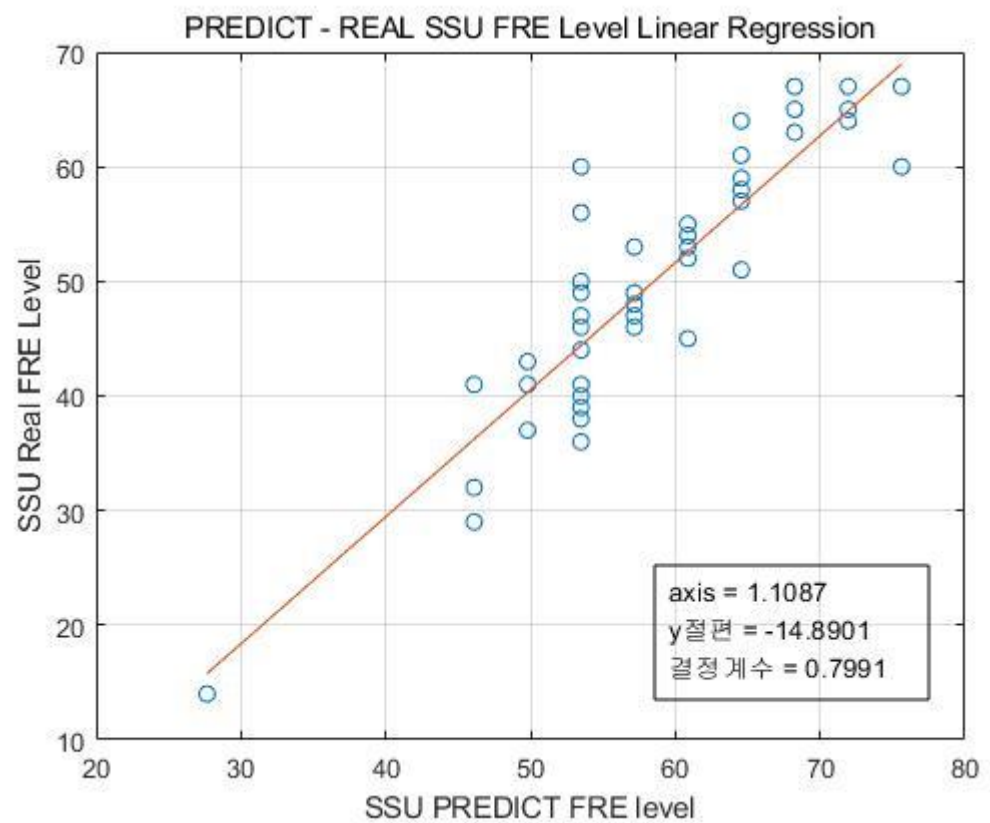
$$y = -3.8694x + 97.1371$$

숭실대생 문장력 지표 예측 값과 실제 값 비교 FK Level



FK – FRE 간의 (float)
회귀 계수와 결정계수

X	Predict FKL
y	Real FK Level
Axis	0.8063
Y절편	1.9987
결정계수	0.8757



숭실대생 문장력 지표 예측 값과 실제 값 비교 FRE Score

FK – FRE 간의 (float) 회귀 계수와 결정계수	
X	Predict FRE
y	Real FRE
Axis	1.1746
Y절편	-18.4187
결정계수	0.8311

연구 시사점

크롤링 사이트의 평가지표 관련
해당 사이트의 작동 방식/양상
집작

데이터 처리 中 보완 부분 발견
Data의 수 부족
다른 평가지표 활용 부재
다른 분석기법 활용 부재
...



새 문장력 평가 모델 구상
FK level + FRE + G.F.I.을 활용
한
새 평가 지표 구축

송실대 학생들의 영어 수준 집작
송실대 학생들의
영어 문장력 지표를 가늠
...> 오차 범위는 有 (Data의 차
이 등)

Else...



Readability

Thank you

