NLP COURSE PROJECT

STUDY RECOMMENDER SYSTEM

USING GPT-3

Jan Deller,
Erwin Smith &
Jan Peter Prigge

THE PROJECT

WHAT?

BUILDING A STUDY RECOMMENDER SYSTEM USING SEMANTIC SEARCH OF GPT-3

THE PROJECT

WHY?

PROBLEM

- × Study program titles don‘t always deliver what they promise
- × Students barely read the module descriptions thoroughly
- × Expectations are not beeing met

→ **HIGH STUDENT DROPOUT RATES** **≈30%**

THE PROJECT

WHY?

SOLUTION

- ✓ Search through large study program and module descriptions automatically
- ✓ Match them with input of students like: Study field, future job goals, skills you want to learn, study and exam type
- ✓ Give recommendation about study program
  → **TO MAKE BETTER DECISIONS**

THE TEAM

WHO?

JAN

Lightning fast Coder

ERWIN

Python based data science dude

PETER

Joined for free Wifi

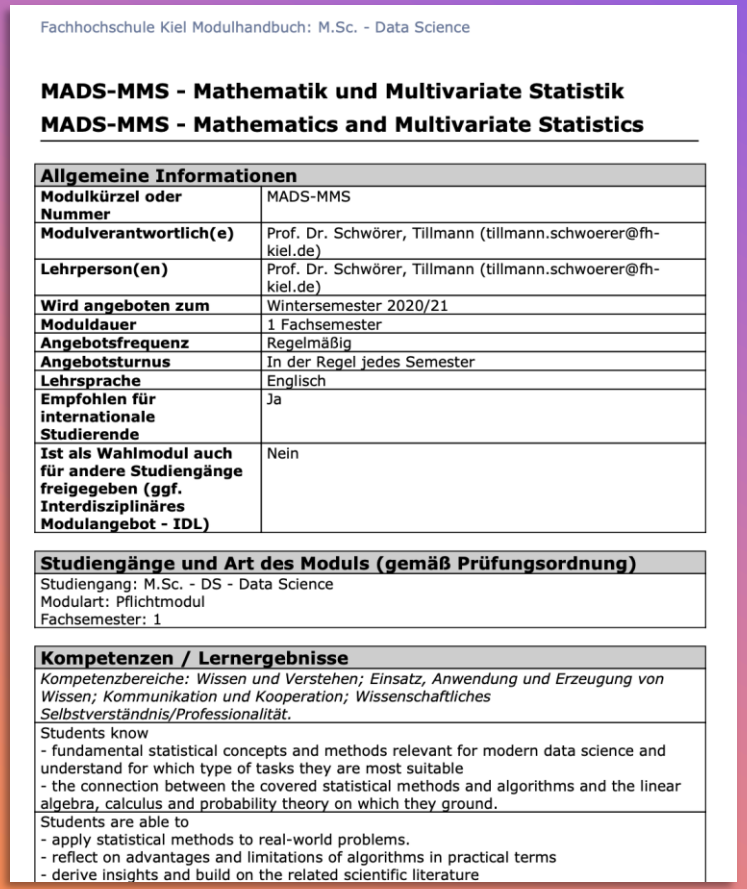
THE PROCESS

HOW?

1. Gather module data of study programmes
2. Gather data of students as examples
3. Determine input parameters
4. Match with semantic search by GPT-3
5. Check results & test with other model

1. MODULE DATA

THE PROCESS



Fachhochschule Kiel Modulhandbuch: M.Sc. - Data Science

MADS-MMS - Mathematik und Multivariate Statistik
MADS-MMS - Mathematics and Multivariate Statistics

Allgemeine Informationen	
Modulkürzel oder Nummer	MADS-MMS
Modulverantwortlich(e)	Prof. Dr. Schwörer, Tillmann (tillmann.schwoerer@fh-kiel.de)
Lehrperson(en)	Prof. Dr. Schwörer, Tillmann (tillmann.schwoerer@fh-kiel.de)
Wird angeboten zum	Wintersemester 2020/21
Moduldauer	1 Fachsemester
Angebotsfrequenz	Regelmäßig
Angebotsturnus	In der Regel jedes Semester
Lehrsprache	Englisch
Empfohlen für internationale Studierende	Ja
Ist als Wahlmodul auch für andere Studiengänge freigegeben (ggf. Interdisziplinäres Modulangebot - IDL)	Nein

Studiengänge und Art des Moduls (gemäß Prüfungsordnung)
Studiengang: M.Sc. - DS - Data Science Modulart: Pflichtmodul Fachsemester: 1

Kompetenzen / Lernergebnisse
Kompetenzbereiche: Wissen und Verstehen; Einsatz, Anwendung und Erzeugung von Wissen; Kommunikation und Kooperation; Wissenschaftliches Selbstverständnis/Professionalität.
Students know - fundamental statistical concepts and methods relevant for modern data science and understand for which type of tasks they are most suitable - the connection between the covered statistical methods and algorithms and the linear algebra, calculus and probability theory on which they ground.
Students are able to - apply statistical methods to real-world problems. - reflect on advantages and limitations of algorithms in practical terms - derive insights and build on the related scientific literature

```
MADS-MMS - Mathematics and Multivariate Statistics
Kompetenzen / Lernergebnisse
Kompetenzbereiche: Wissen und Verstehen; Einsatz, Anwendung und Erzeugung von
Wissen; Kommunikation und Kooperation; Wissenschaftliches
Selbstverständnis/Professionalität.
Students know
- fundamental statistical concepts and methods relevant for modern data science and
understand for which type of tasks they are most suitable
- the connection between the covered statistical methods and algorithms and the linear
algebra, calculus and probability theory on which they ground.
Students are able to
- apply statistical methods to real-world problems.
- reflect on advantages and limitations of algorithms in practical terms
- derive insights and build on the related scientific literature
Students are able to
- correctly interpret and communicate the approach and results both in technical and
functional terms
- work successfully in teams, leveraging the individual skills of all team members
Angaben zum Inhalt
Lehrinhalte Statistics:
- Clustering
- Dimensionality reduction
- Linear regression
- Logistic regression
Literatur
Math:
- Basic linear algebra and calculus
- Similarity and distance measures
- Matrix decomposition techniques
- Gradient descent
- Leskovec, Rajaraman and Ullman: Mining of Massive Datasets. Cambridge
Univeristy Press, second edition. Available online: http://www.mmds.org.
- James, Witten, Hastie, and Tibshirani: An Introduction to Statistical
Learning with Applications in R. New York first edition. Available online:
https://web.stanford.edu/~hastie/Papers/ESLII.pdf.
- Hothorn and Everitt: A Handbook of Statistical Analyses Using R.
Routledge, third edition.
- Boyd and Vandenberghe: Introduction to Applied Linear Algebra.
```

2. INPUT DATA

Peron 1: JOE

- PREVIOUS INPUT:
- Bachelor of Science in Technology Management
- Node.js, JavaScript, HTML/CSS, SQL
- -
- English, German, French
- FUTURE STUDIES INPUT:
- Computer Science
- C, Python
- Full Stack developer
- Presentation
- Group projects

Person 2: MURAT

- PREVIOUS INPUT:
- Bachelor of Science in Electrical Engineering
- R&D Engineer
- Angular, Bootstrap, SQL, HTML, CSS, Javascript
- -
- English
- FUTURE STUDIES INPUT:
- Electrical Engineering
- C++, Python
- IoT Engineer
- written
- Group projects

Person 3: ISABEL

- PREVIOUS INPUT:
- Bachelor of Science Physics
- Software Engineer
- Angular, HTML, CSS, Javascript, SQL
- -
- English
- FUTURE STUDIES INPUT:
- Computer Application
- Project management, SCRUM
- Technical Lead
- presentation
- Group projects

Person 4: HANNAN

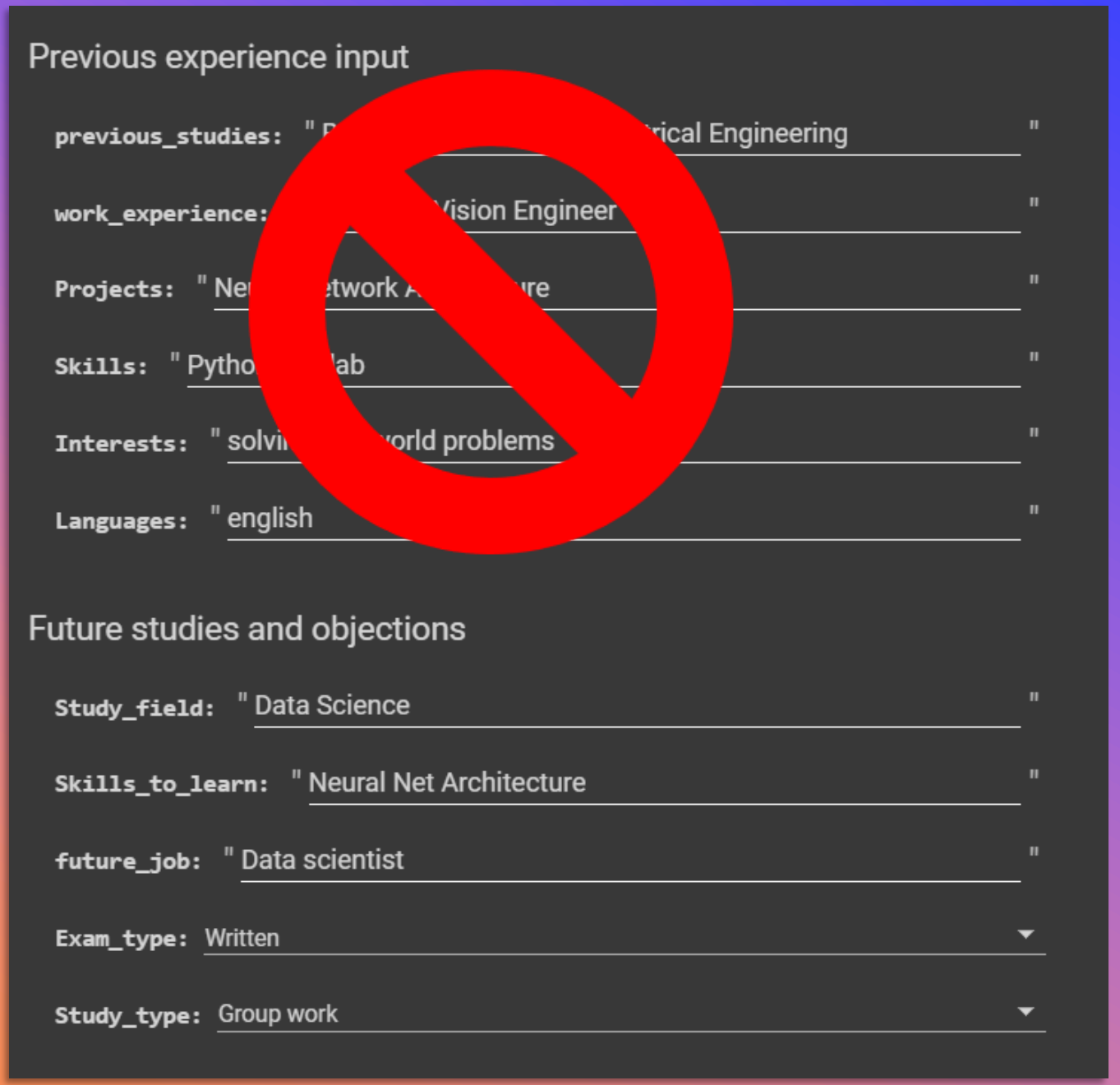
- PREVIOUS INPUT:
- Bachelor of Science in Electrical Engineering
- Computer Vision Engineer
- Python, Matlab, C++
- -
- English
- FUTURE STUDIES INPUT:
- Data Science
- Neural Net Architecture
- Data scientist
- written
- solo

Person 5: TOBY

- PREVIOUS INPUT:
- Bachelor of Arts in Communication
- UI Developer
- HTML5, CSS3, Java script, J Query, React
- -
- English
- FUTURE STUDIES INPUT:
- Software Development
- Agile development
- Senior UI/UX Engineer
- written
- group projects

3. PARAMETERS

THE PROCESS



5. TEST AGAINST OTHER MODELS


```
[ ]  result = get_word_matching_in_gpt3_fashion(
         documents=documents,
         query=inputParameters)
```

```
print_best_3(result,documents, names)
```

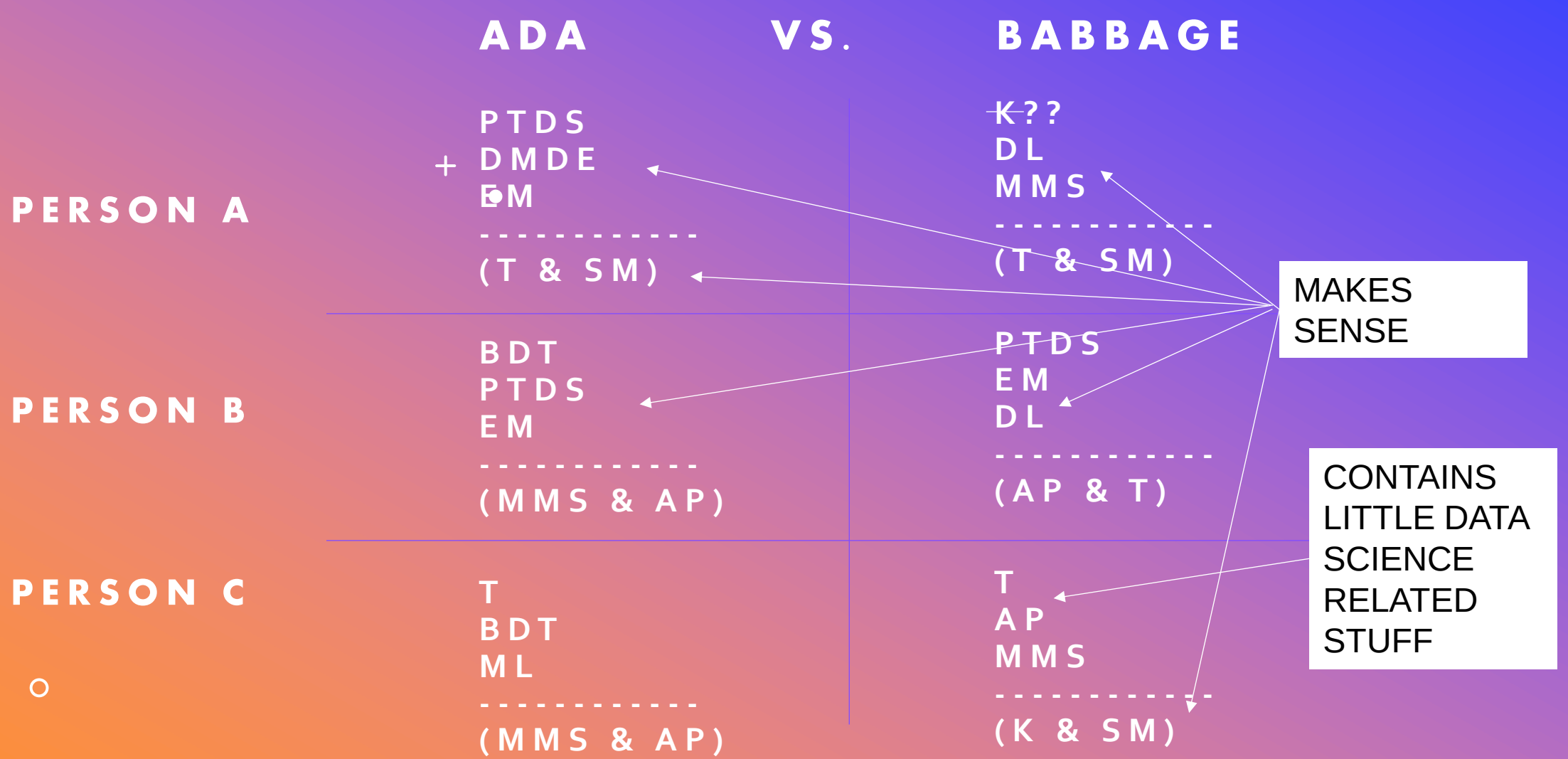
```
Based on our analysis of your inputted data, we have found the following best three matching study programmes
for you (with a corresponding level of suitability):

1st Best Ranked Course:  MADS-DMDE
Matching Score:          99%
Level of Suitabilty:     Outstandingly High

2nd Best Ranked Course:  MADS-AP
Matching Score:          73%
Level of Suitabilty:     Fairly Suitable

3rd Best Ranked Course:  MADS-BDT
Matching Score:          61%
Level of Suitabilty:     Least Suitable
```

LOOKING INTO THE NOTEBOOK

THE TEST

	+ PERSON A	PERSON B	PERSON C
INTEREST	Solving problems	Computer Vision	Mechanical industry
STUDY FIELD	Data Science	Artificial Intelligence	Engineering+Business
SKILLS2LEARN	Neural Net Architecture	Python	Project Mgmt, SAP
FUTURE JOB	Data Scientist	Technical Lead	Manager
EXAM TYPE	Written	Portfolio	Don‘t care
STUDY TYPE	Group Work	Group Work	Solo

THE RESULTS


ADA
VS.
BABBAGE
PERSON A
PTDS
DMDE
EM
(T & SM)
K??
DL
MMS
(T & SM)
PERSON B
BDT
PTDS
EM
(MMS & AP)
PTDS
EM
DL
(AP & T)
PERSON C
T
BDT
ML
(MMS & AP)
T
AP
MMS
(K & SM)
MAKES SENSE
CONTAINS LITTLE DATA SCIENCE RELATED STUFF

THE RESULTS

ADA VS. BABBAGE

- Very small scores for Person C → makes sense
- String-Matcher has almost same result for every person
- Both Engines show reasonable results
- „Babbage“ Engine seems to be performing better
  - → good, because cheapest engine

THE RESULTS

IT-SECURITY DATA SCIENCE INDUSTRIAL ENG.

A

```
result = openai.Engine("ada").search(
  documents=documents,
  query=inputParameters
)
program_scores = dc.calculate_scores_from_API_result(result)
print(program_scores)
```

```
[('IT Security', 17.595344827586203), ('Data Science', 26.470499999999998), ('Industrial Engineering', 8.020363636363635)]
```

String matcher

```
[('IT Security', 0.04507460892899427), ('Data Science', 0.08475135384837734), ('Industrial Engineering', 0.04665500749772917)]
```

```
result = openai.Engine("babbage").search(
  documents=documents,
  query=inputParameters
)
program_scores = dc.calculate_scores_from_API_result(result)
print(program_scores)
```

```
[('IT Security', 21.054379310344828), ('Data Science', 26.597583333333336), ('Industrial Engineering', 17.57963636363636)]
```

THE RESULTS

IT-SECURITY DATA SCIENCE INDUSTRIAL ENG.

B

```
result = openai.Engine("ada").search(
  documents=documents,
  query=inputParameters
)
program_scores = dc.calculate_scores_from_API_result(result)
print(program_scores)
```

```
[('IT Security', 42.23079310344828), ('Data Science', 46.46808333333333), ('Industrial Engineering', 34.03454545454545)]
```

String matcher

```
[('IT Security', 0.03480233659767262), ('Data Science', 0.04974712306521012), ('Industrial Engineering', 0.042531495888057665)]
```

```
result = openai.Engine("babbage").search(
  documents=documents,
  query=inputParameters
)
program_scores = dc.calculate_scores_from_API_result(result)
print(program_scores)
```

```
[('IT Security', 30.900827586206898), ('Data Science', 35.40408333333333), ('Industrial Engineering', 21.868272727272725)]
```

THE RESULTS

IT-SECURITY DATA SCIENCE INDUSTRIAL ENG.

C

```
result = openai.Engine("ada").search(
  documents=documents,
  query=inputParameters
)
program_scores = dc.calculate_scores_from_API_result(result)
print(program_scores)
```

```
[('IT Security', 33.93679310344829), ('Data Science', 30.75558333333333), ('Industrial Engineering', 39.35418181818182)]
```

String matcher

```
[('IT Security', 0.07004271796697953), ('Data Science', 0.08253510313833627), ('Industrial Engineering', 0.10848089101402013)]
```

```
result = openai.Engine("babbage").search(
  documents=documents,
  query=inputParameters
)
program_scores = dc.calculate_scores_from_API_result(result)
print(program_scores)
```

```
[('IT Security', 14.227448275862066), ('Data Science', 11.318916666666667), ('Industrial Engineering', 23.806999999999995)]
```

THE RESULTS

MULTIPLE PROGRAMS

- Overall reasonable results
- Better recommendations with GPT-3
- Babbage Engine also performs better here

THE LEARNINGS

- Individual setup / data preperation required for university specific module description structure
- The generated matching score might be intransparent
- Future related input matters
- Needs to be tested with German descriptions (not always available in english)

THE NEXT STEPS

- Gather more course description data and try more examples
- Create a proper concept for a use case at universities
- Build Frontend-Prototype
- Test with real users

THE NEXT STEPS

WILL IT MATCH?

Fill in a few details about your future studies and career:

In which study field would you like to study?

Standard form field

Which skills would you like to learn?

Standard form field

In which study field would you like to study?

Standard form field

Which exam type do you prefer?

Choose an option...

Which study type are you?

Choose an option...

START THE MATCH MAKER

NLP COURSE PROJECT

THANKS FOR LISTENING

Jan Deller,
Erwin Smith &
Jan Peter Prigge