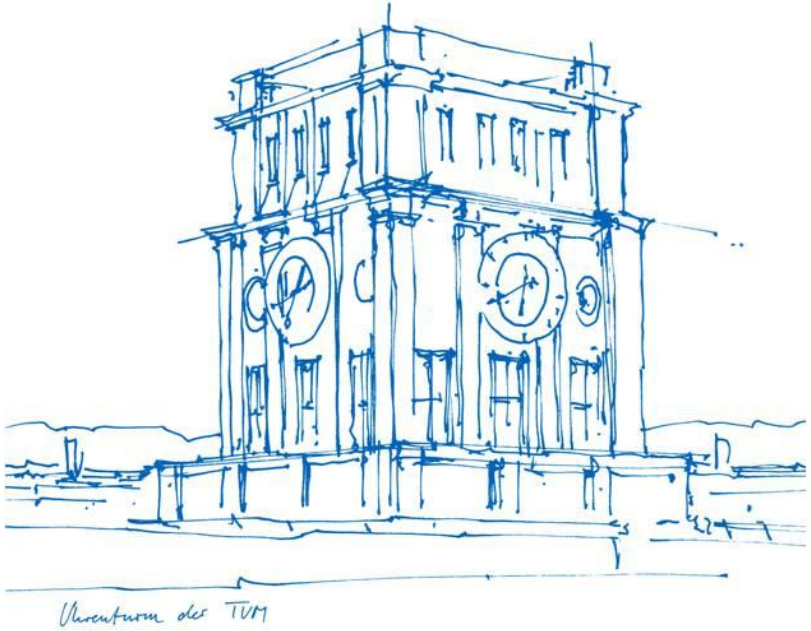
IDP StudySmarter – Document Classification Documentation

Technical University Munich

Supervisor Prof. Dr. Nicola Breugst

Munich, August 13, 2018



Document Classification Team

Nitin Vashisth

Vindhya Singh

Document Classification

1. Introduction

2. Development

3. Challenges & Learning

Tasks Performed

Research about machine learning and natural language processing, algorithms used, techniques implemented.

Coding the solution → Training the model → Evaluating the model

Algorithms Used : Stochastic Gradient Descent Classifier, Neural Network and Multinomial Binomial Classifier

Marketing and Innovation Management **Exam**

Documentation and **Presentation**

Objective

Automatic Classification of documents into respective subjects and extract exam dates using Machine Learning and NLP

Introduction

- Automatic document classification to respective subjects offers users to
    1. Manage the uploaded documents
    2. Eases the need to sort out the documents topicwise
    3. Helps classify similar topics making learning easier
    4. Suggested exam dates can help schedule studies

- Helps the developers to improvise on future aspects of designing a learning platform

Action Plan

Input : A corpus of PDF Files

↓

Pre-processing

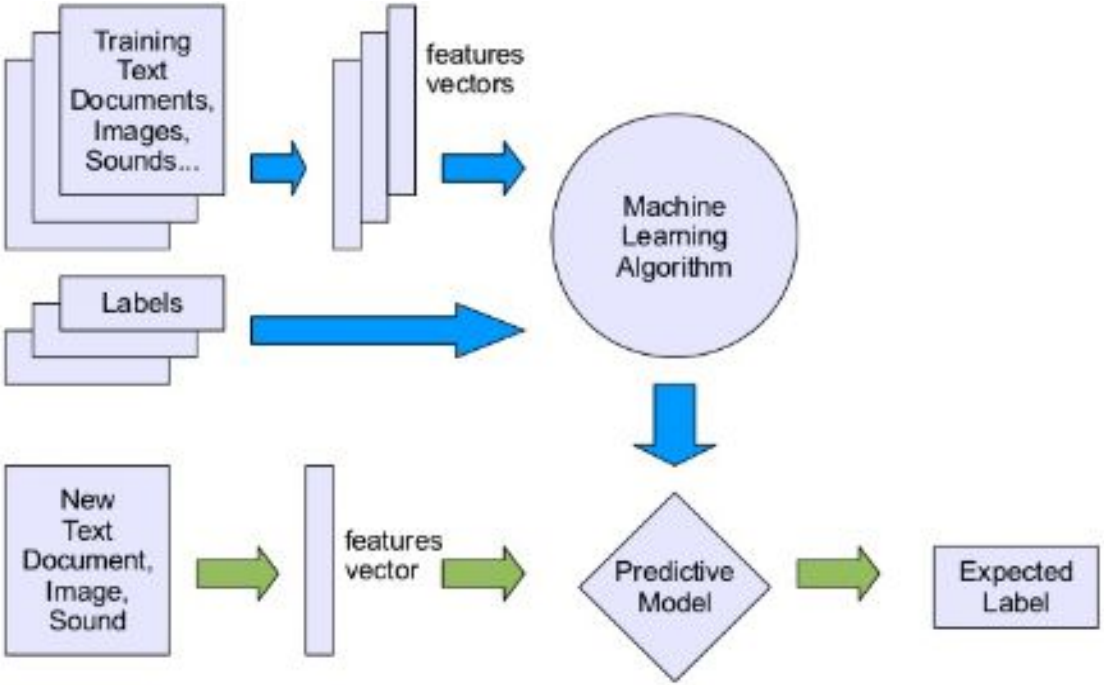
↓

Extraction of Text Blocks

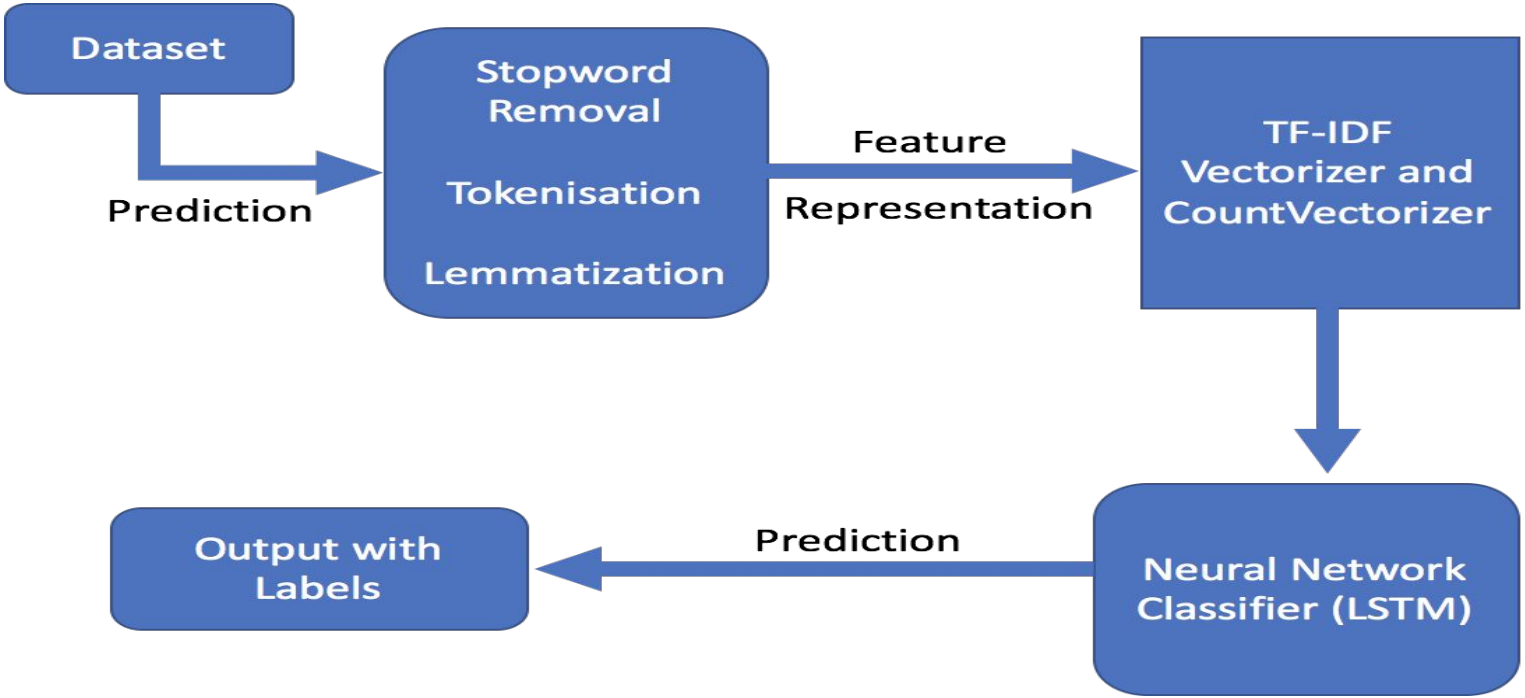
↓

Classifying into respective labels

Development Phase



General Flow of Development



Development

- Download of OCW (MIT) dataset with the list of 12,500 + documents
- Automating the conversion of PDF's to TXT's through pdfminer
- Preprocessing of text of each document and binding it with respective label
  A. Converted all text to lowercase
  B. Removed all text of length size upto 3
  C. Removed all junk characters and numbers
  D. Word lemmatization (This helps to cut short the word into the regular base words)
  E. Binding the processed text and the label together.
- Split the processed dataset into 80:20
- Create stochastic gradient classifier model, input the values of TfidfVectorizer, hinge loss, learning rate as 0.001
- Train the model and calculate the accuracy
- Test it on the sample pdf and predict the label of sample pdf by using the trained model.

Development - Concept Map

Stop Words

Word Tokenize & Stop words removal

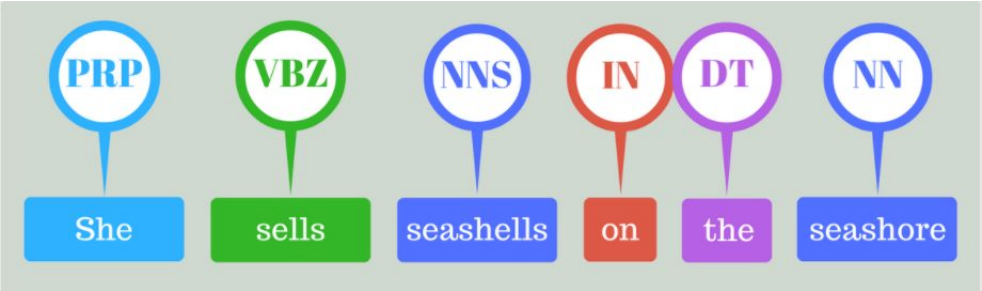
“Nice book! Though it has a lack of advanced topics it’s still good for beginners”

‘Nice’, ‘book’, ‘!’, ‘though’, ‘it’, ‘has’, ‘a’, ‘lack’, ‘of’, ‘advanced’, ‘topics’, ‘it’s’, ‘still’, ‘good’, ‘for’, ‘beginners’

Note : Slides taken from the concept map team as the algorithm used was the same

Development - Concept Map

Part of Speech Tagging



PRP → Personal Pronoun

VBZ → Verb 3rd person

NNS → Noun plural

IN → Preposition

DT → Pre Determiner

NN → Noun Singular

"accounting" → verb ✗

"international accounting standard" → adjective + noun + noun ✓

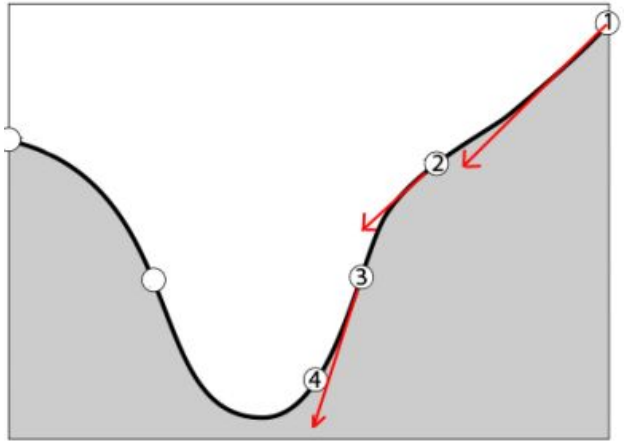
Note : Slides taken from the concept map team as the algorithm used was the same

Stochastic Gradient Descent Classifier

Gradient descent is a first-order iterative optimization **algorithm** for finding the minimum of a function. To find a local minimum of a function using **gradient descent**, one takes steps proportional to the negative of the **gradient** (or approximate **gradient**) of the function at the current point.

• Key idea
– Gradient points into steepest ascent direction
– Locally, the gradient is a good approximation of the objective function
• GD with Line Search
– Get descent direction, then unconstrained line search
– Turn a multidimensional problem into a one-dimensional problem that we already know how to solve

(Source : Machine Learning,Slide 06/22, WiSe 2017, Prof. Dr. Stephan Günnemann)

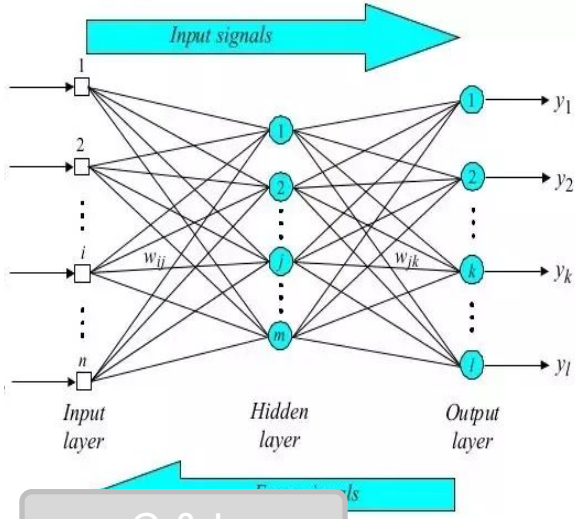


Stochastic gradient descent is an optimization method for unconstrained optimization problems. In contrast to (batch) gradient descent, SGD approximates the true gradient of by considering a single training example at a time. The class **SGDClassifier** implements a first-order SGD learning routine.

Neural Network

An **Artificial Neural Network (ANN)** : It is composed of a large number of highly interconnected processing elements (neurones) working in unison to solve specific problems. ANNs, like people, learn by example. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process.

Neural networks learn things in exactly the same way, typically by a feedback process called **backpropagation**. This involves comparing the output a network produces with the output it was meant to produce, and using the *difference* between them to modify the weights of the connections between the units in the network, working from the output units through the hidden units to the input units



Multinomial Binomial Classifier

MultinomialNB implements the naive Bayes algorithm for multinomially distributed data, and is one of the two classic naive Bayes variants used in text classification (where the data are typically represented as word vector counts, although tf-idf vectors are also known to work well in practice).

Result

S.No	Number of Labels Used	Prediction Accuracy
1	6	89.16%
2	60	73.86%
3	625	73.06 %

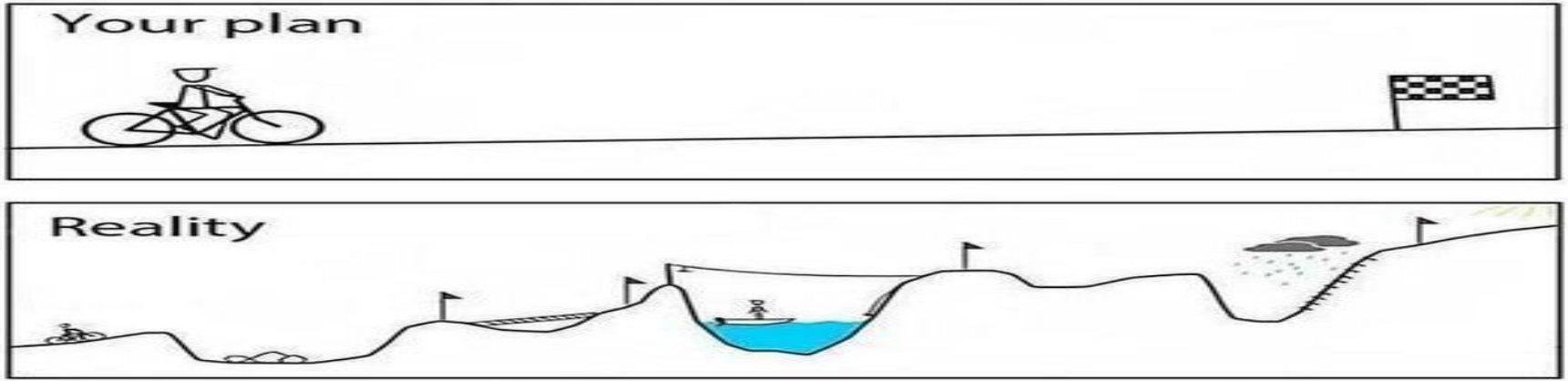
Result

```
python3.6 MIT_Department_Train.py
Done Extraction
Done splitting
Done training
Done prediction
0.8915831663326653
Done processing the unlabeled data
lec02.txt => Mathematics
CAA 108 Lecture.txt => Science
1408278839.txt => Business
2._principlesofdesign.txt => Business
pse2015_3_Design_Patterns_I.txt => Business
```

Result showcasing the documents being correctly classified to respective labels

General Challenges

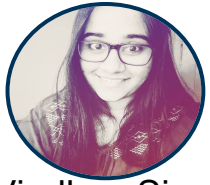
- How to work **part time** in a **fast growing project**
- How to **manage work** and **resources under** time constraints
- **Knowledge transfer** between team members



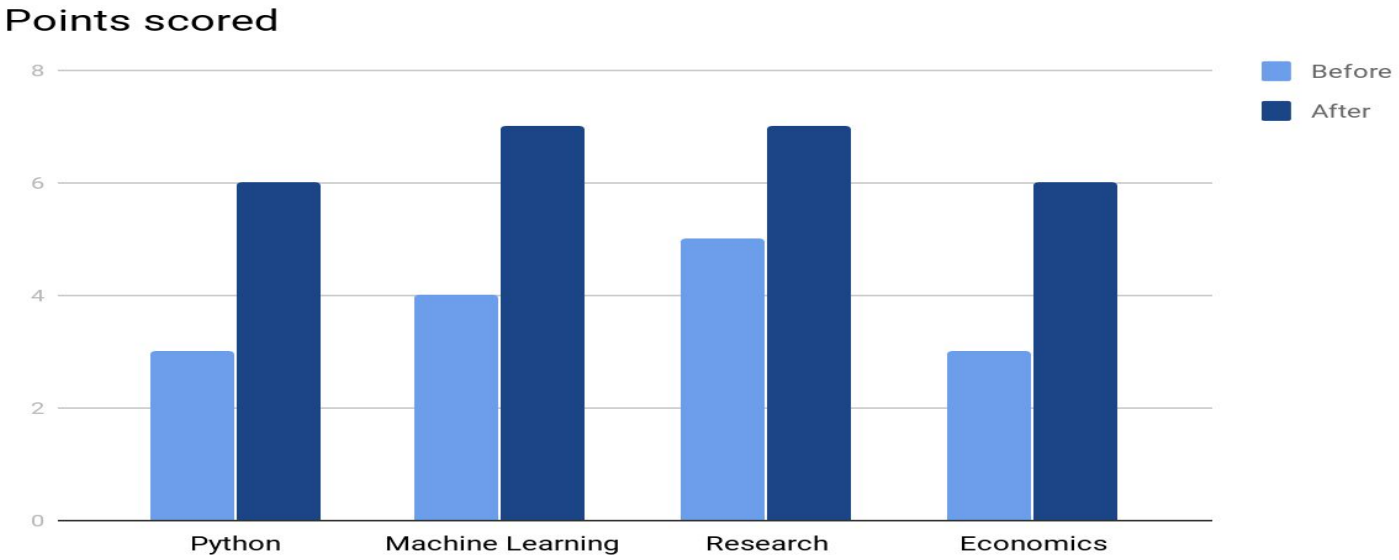
Technical Challenges

- Variety of documents - many languages - **stop word removal for all languages** was different
- **Stem words creation** - difficult due to inconsistencies in the original PDFs
- Some **PDFs had animations :** difficult to classify
- The uploads have Books. Problems- **size, number of tags**.
- The uploads for a subject have Exercises and their solutions. Most solutions to exercises contain **Greek symbols** that are considered garbage.
- Various **date formats**
- The problem with labels.
- Choosing which algorithm to use to train the model.

Learnings – Individual Learning Curve



Vindhya Singh



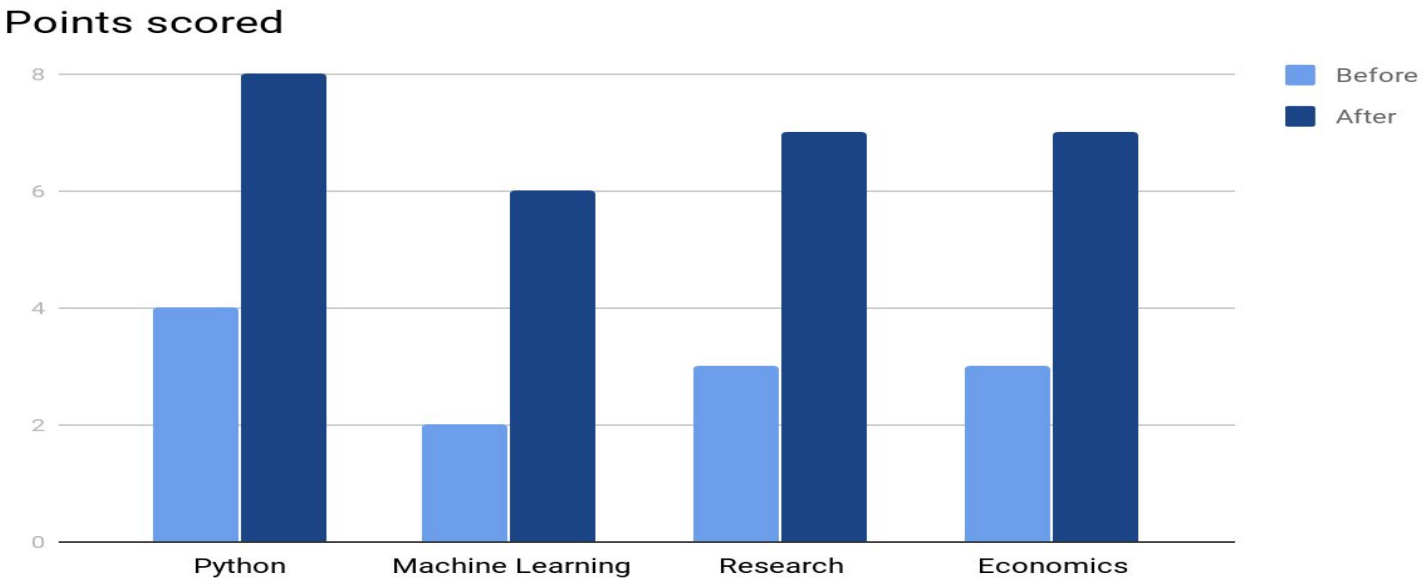
Introduction

Development

C & L

Learnings – Individual Learning Curve

Nitin Vashisth

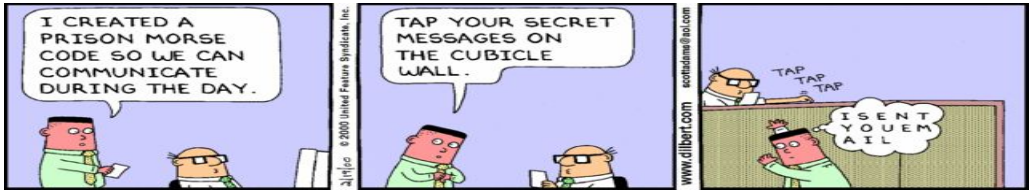


Introduction

Development

C & L

Learning



Technical Aspect

- Implementing Machine Learning project from scratch
- Working on multiple Python Libraries
- Understanding of Neural networks, their learning and functioning
- Implementing algorithms introduced in Research papers
- Understanding of why, when and where to use certain ML Algorithms
- Working with a document corpus

Interdisciplinary Aspect

- Marketing, working of startups : their challenges, vision, mission, strategies
- Using various platforms to organise and communicate : Slack, Trello, Bitbucket
- Technology and Innovation Management
- Organisation, Communication, Project Development in real time
- Working in Pipelines and code management between the team members

Organization & Management

Project Communication and Organization

Trello

- Individual and Team wise tasks assignment and tracking
- Weekly Task Monitoring

Atlassian Bitbucket

- Repositories for frontend and backend development
- Code reviews

slack

- Communication in channels
- Weekly standup with current status
- Collection of ideas

Planning - Trello

Boards | Trello | AA

studysmarter-ml | Personal | Private | AA SH AS AS 9 | Bitbucket | Show Menu

To Do

- Tune hyperparameter for all approaches — AA
- Find sub-candidates from candidates. — R
- Evaluation of Preprocessing + Textrank approach: How can we build upon it, what alternatives do we have — R SS
- Test current solution with key phrases — AS AS

+ Add another card

Doing

- relationship between keywords — SS
- Improve TextRank algorithm - Try out summary/conclusion to improve probability of keyphrase — R
- Calculate classification metrics — 2 — VS
- 1. Automatic KeyWord extraction (Trained classifier on Crowd500) — May 29 — SS
- Recommendation System

+ Add another card

In Review

- insert font weight into the pipeline — 1 — AA R
- Create a way to visualise "correct"/bad key-phrases based on the handcrafted — R
- Fix bug from mindmap2html — R
- Refine the extracted dates — VS

+ Add another card

Done

- Generate some hand-crafted keywords/keyphrases for the 2 pdfs we are using. — AA R SS
- Extract the date from the given PDFs for exams — VS
- make flowchart to better understand — R
- Each of us take one algorithm (GitHub posts) and try out. — AA R SS

+ Add another card

Dropped

- Find one or many datasets mapping text to concepts — 2 — AA R S
- 2. PageRank, Strength score, and (Degree + Closeness + Betweennes using tf-idf or tf for candidates extraction. — May 29 — 1 — S
- 5. kpex - This is the same as TF-IDF — May 16 — 1 — R
- 7. PKE - PositionRank — R

+ Add another card

IDP - StudySmarter

Thank you!

Technical University Munich

Supervisor Prof. Dr. Nicola Breugst

Munich, August 13, 2018