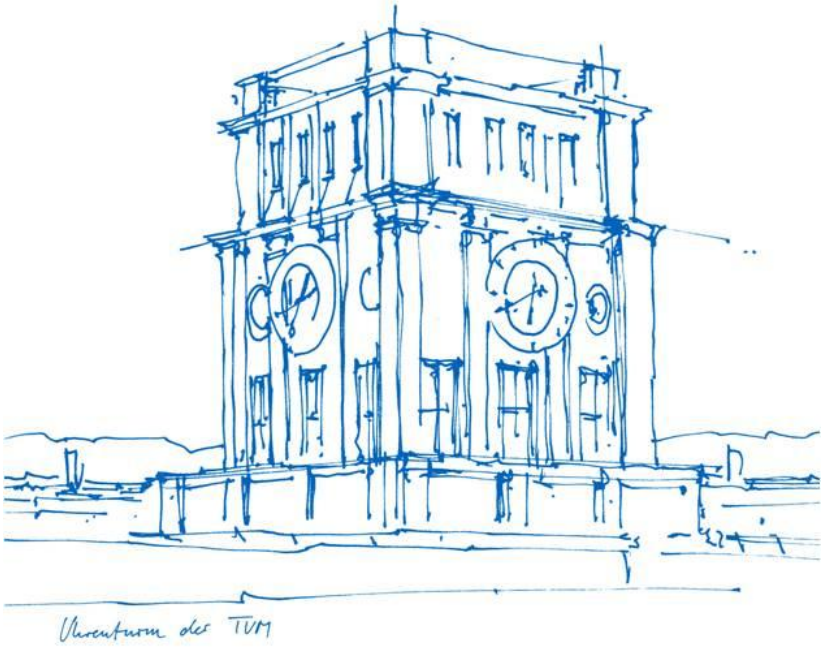
IDP - StudySmarter
Machine Learning Final Presentation

Technical University Munich

Supervisor Prof. Dr. Nicola Breugst

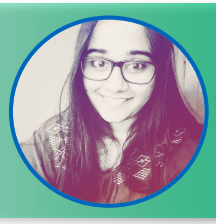
Munich, August 13, 2018



Meet the Team

Document Classification

NITIN



VINDHYA

Concept map

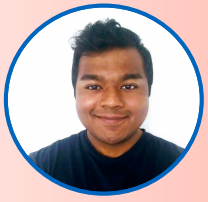
ROBERTO

ANUM

SAPTWARSHI

Recommender systems

ANSHUL



ARVINDH

Main Agenda


Concept Map


Document Classification


Recommender System


Organization & Management


Questions & Answers

Concept Map

1. Introduction

2. Development

3. Challenges & Learning

Introduction - Concept Map

- Concept Maps are nice tools for **knowledge visualization**.
- Represents the **relationship among key words**.
- Highly used among students of different study fields.

Introduction - Concept Map

Using IBM Watson

Pros

- Efficient
- Few parameters to use
- Good for fast development of product

Cons

- It's costly
- Rigid structure
- Built for general usage
- Not customizable
- Doubtful quality
- It's a black box

Introduction - Concept Map (Goal)

- Build our own customized pipelined.
  - Extracting more relevant lecture key words
  - Open the black box
- Use Machine Learning techniques/algorithms.

We want to help StudySmarter to make it better and free!

Introduction - Concept Map (Examples)

Lehrstuhl Elektrische Energiespeichertechnik

Lehrstuhlleitung:
Prof. Dr. Ing. Andreas Jossen
Tel: +49 (0) 89 / 289 – 26966
andreas.jossen@tum.de

Holger Hesse
Leitung Team stationäre Energiespeicher
Stellvertretende LS-Leitung
Tel: +49 (0) 89 / 289 – 26964
holger.hesse@tum.de

Franz Spingler
Li-Ion-Batterie Forschung
Tel: +49 (0) 89 / 289 – 26962
franz.spingler@tum.de

Holger Hesse, Franz Spingler | Ringvorlesung | 13.10.2017 2 / 50

Motivation

Start-ups und Innovationsfelder

Digitalisierung und Startups: Was die Bundesregierung nach der Wahl angehen muss

Im TV-Duell spielte die Digitalisierung keine Rolle. Man kann nur hoffen, dass diese Ignoranz kein Abbild der kommenden vier Jahre ist. Sonst verpassen wir die wichtigste Chance unserer Zeit.

Frankfurter Allgemeine

Bewegung im Energiespeicher

Adaptive Balancing Power will ein zentrales Problem der erneuerbaren Energien lösen. Einen Preis hat das in Darmstadt ansässige Start-up schon gewonnen.

Quelle:http://t3n.de/news/digitalisierung-startups-bundesregierung-wahl-859700/
http://www.faz.net/aktuell/rhein-main/die-erfindung-des-schwungmassenspeichers-15199298.html

Holger Hesse, Franz Spingler | Ringvorlesung | 13.10.2017 4 / 50

Klausur

- Es findet eine Finalklausur statt, keine Midterm.
  - Die Klausur dauert 90 Minuten.
  - Umfasst den gesamten hier vorgestellten Stoff und den gesamten Stoff der Übungen.
  - Es gibt zwei Termine für die Klausur, bei beiden gilt (falls erreicht) der Bonus, egal ob sie die erste Klausur schreiben oder nicht.
- Insbesondere in der **Zentralübung** gegen „Mitte des Semester" (wann immer das sein wird ... Während eines Vorlesungstermins ... Wird **nicht** angekündigt) wird anlassbezogen diskutiert, wie so eine Klausur in etwa aufgebaut ist und wie man sich am besten vorbereiten kann.
- Es wird bei diesem Termin auch eine kurze Probeklausur ausgeteilt, die aber nicht bewertet (nicht einmal eingesammelt) wird

Odd – even Partitioning

$$v_j = \sum_{k=0}^{n-1} c_k \cdot \exp\left(\frac{2\pi i j k}{n}\right) =$$

$$= \sum_{k=0}^{n/2-1} c_{2k} \cdot \exp\left(\frac{2\pi i j 2k}{n}\right) + \sum_{k=0}^{n/2-1} c_{2k+1} \cdot \exp\left(\frac{2\pi i j (2k+1)}{n}\right)$$

$$= \sum_{k=0}^{m-1} c_{2k} \cdot \exp\left(\frac{2\pi i j k}{m}\right) + \omega^j \cdot \sum_{k=0}^{m-1} c_{2k+1} \cdot \exp\left(\frac{2\pi i j k}{m}\right)$$

$$v_{m+j} =$$

$$= \sum_{k=0}^{m-1} c_{2k} \cdot \exp\left(\frac{2i\pi(j+m)k}{m}\right) + \omega^{j+m} \cdot \sum_{k=0}^{m-1} c_{2k+1} \cdot \exp\left(\frac{2i\pi(j+m)k}{m}\right)$$

$$= \sum_{k=0}^{m-1} c_{2k} \cdot \exp\left(\frac{2ijk\pi}{m}\right) - \omega^j \cdot \sum_{k=0}^{m-1} c_{2k+1} \cdot \exp\left(\frac{2ijk\pi}{m}\right)$$

Butterfly: a_j, b_j → ω^j → $a_j + \omega^j b_j$, $a_j - \omega^j b_j$

3

Development - Concept Map

Development of a **Concept-Map** which contains important keywords from a lecture

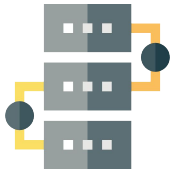
The Idea

As a student I can

- Add subjects and lecture slides to study,
- Each lecture slide is automatically fed to the concept-map pipeline
- A graph containing keywords which represent the key concepts of a lecture.

Development - Concept Map

Lecture Slide

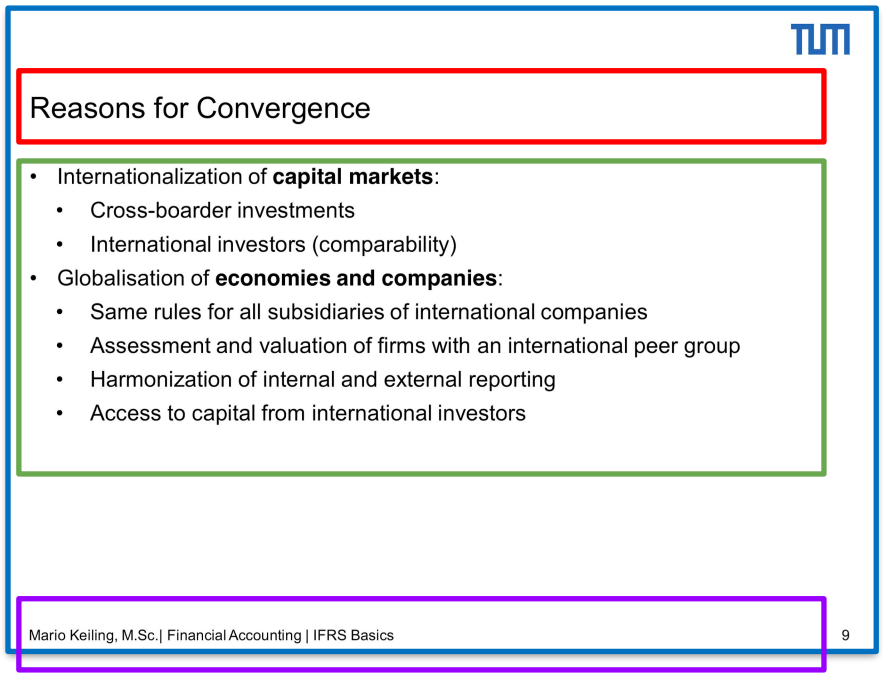


Concept-map pipeline

Concept-map

Development - Concept Map

PDF Structure (Page Content)



Header/ Title

Body

Footer

Lecture Slide

Corpus (Titles)

Corpus (Body)

Introduction

Development

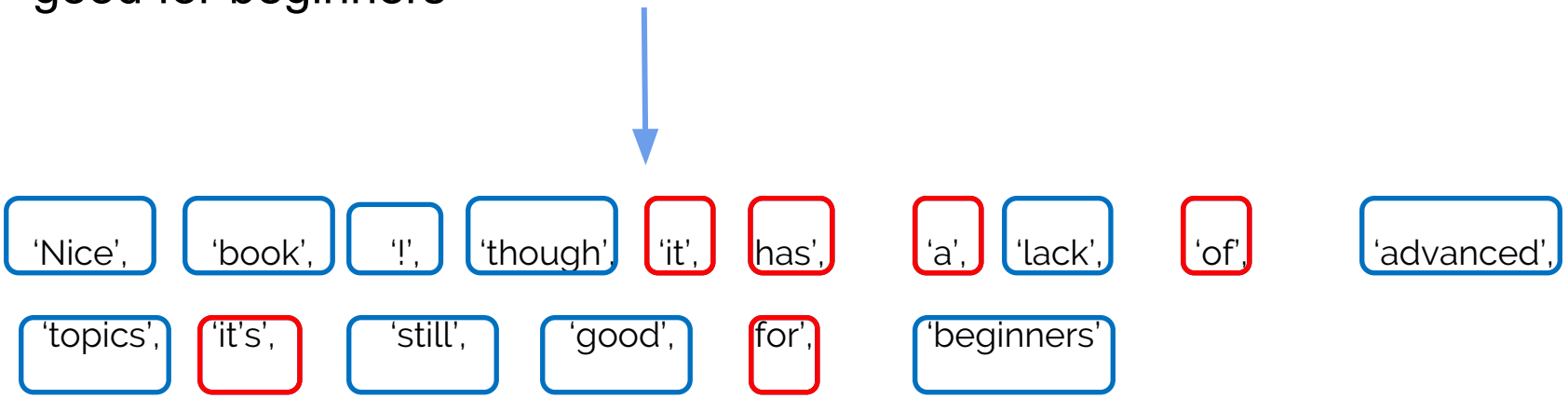
C & L

Development - Concept Map

Stop Words

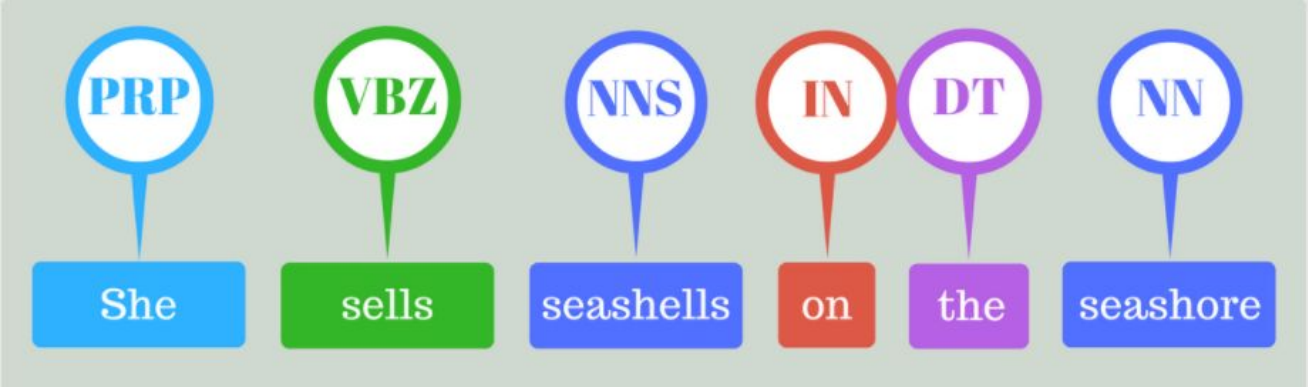
Word Tokenize & Stop words removal

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



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 ✓

Development - Concept Map

Keywords from Titles

Keywords from body of text

Challenges - Concept Map

GENERAL

- How to work **part time** in a **fast growing project.**
- How to **manage work** and **resources under** time constraints.
- **Knowledge transfer** between team members.
- Code management using Git and Bitbucket.

TECHNICAL

- **Metrics** to measure accuracy of our results.
- Different PDFs have different structure.
- Finding **open source** libraries.

Learnings – Concept Map (NLP)

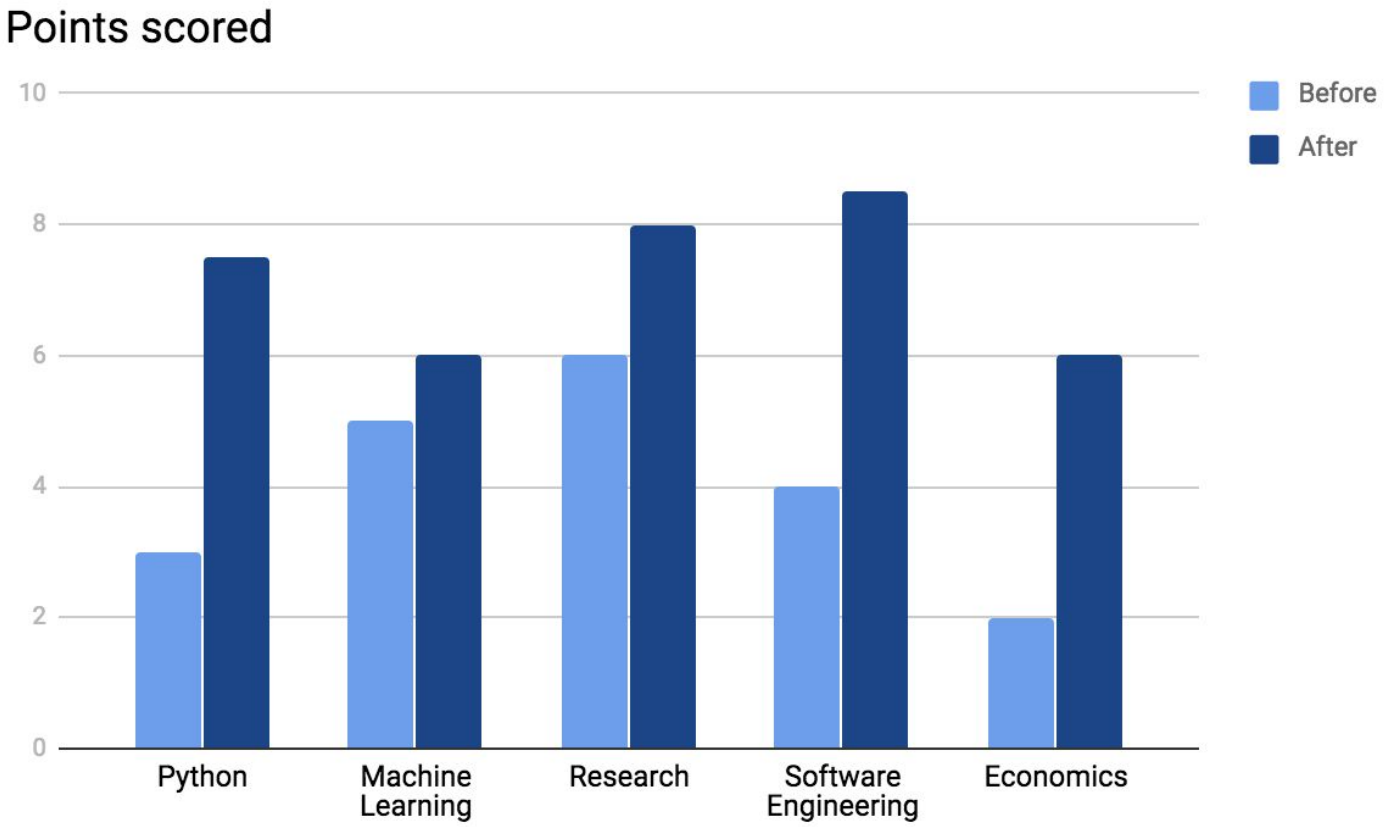
- Worked with state of the art **algorithms** and **techniques** in the field of **Natural Language Processing**.
- Literature Review.
- Familiarity with the **working pipeline** in NLP.
- Exposure to a **large dataset** of pdfs.

Learnings – Concept Map (Interdisciplinary Aspects)

- Familiarity with the working **hierarchy of startups**.
- Group discussions using **Slack** and **trello**
- Code management and integration using **Bitbucket**..
- Frequent meetings.

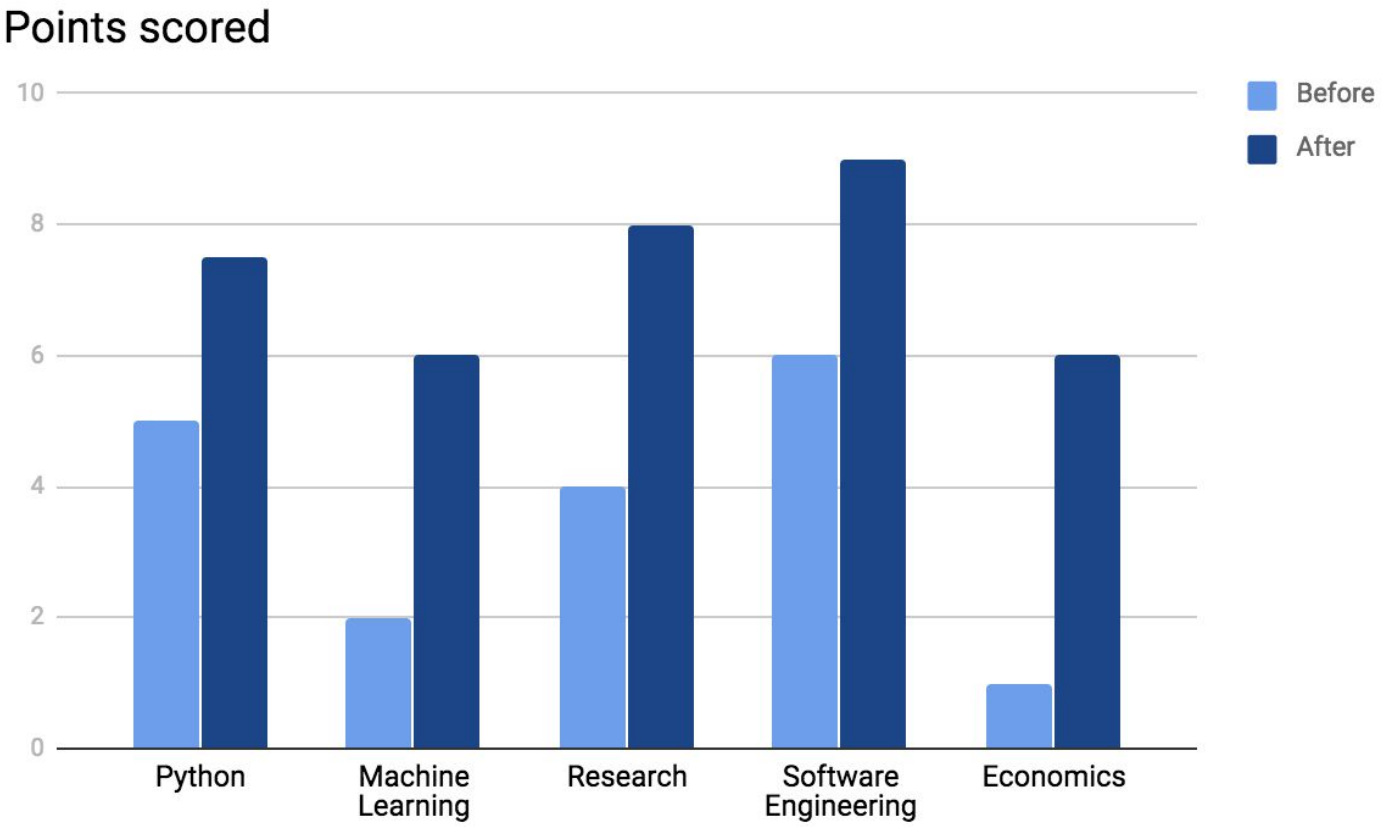
Learnings – Individual Learning Curve

Anum Afzal



Learnings – Individual Learning Curve

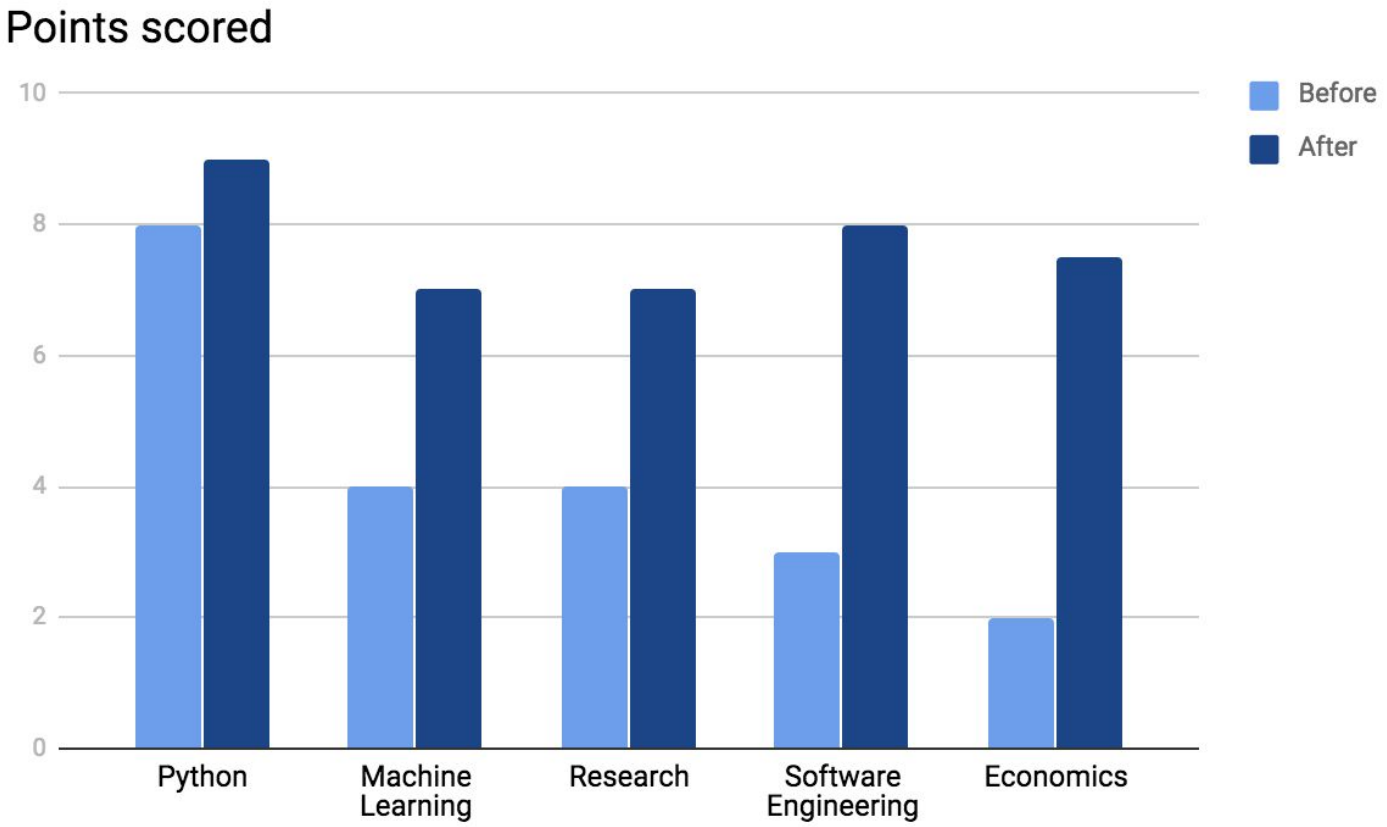
Saptwarshi Saha



Introduction | Development | C & L

Learnings – Individual Learning Curve

Roberto Pereira



Document Classification

1. Introduction

2. Development

3. Challenges & Learning

Objective

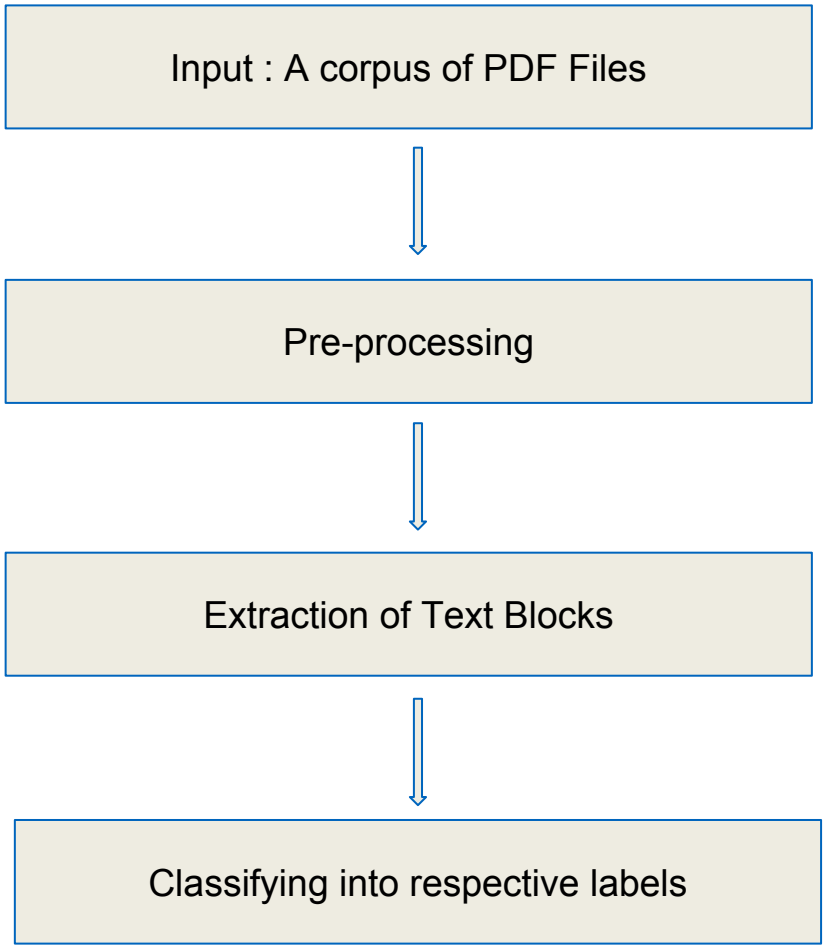
Automatic Classification of documents into respective subjects and extract exam dates using Machine Learning and NLP

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

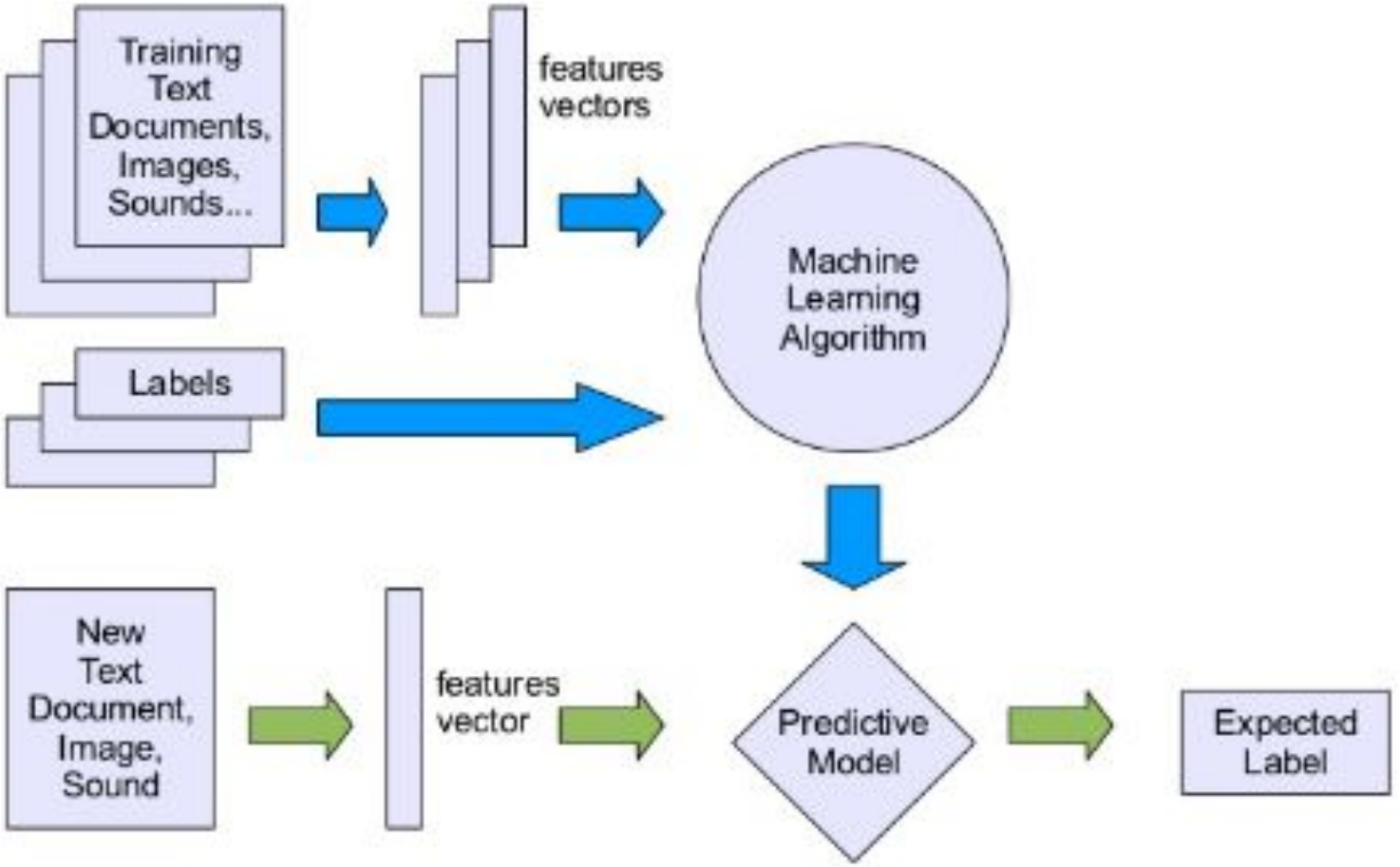
Action Plan


Input : A corpus of PDF Files
Pre-processing
Extraction of Text Blocks
Classifying into respective labels

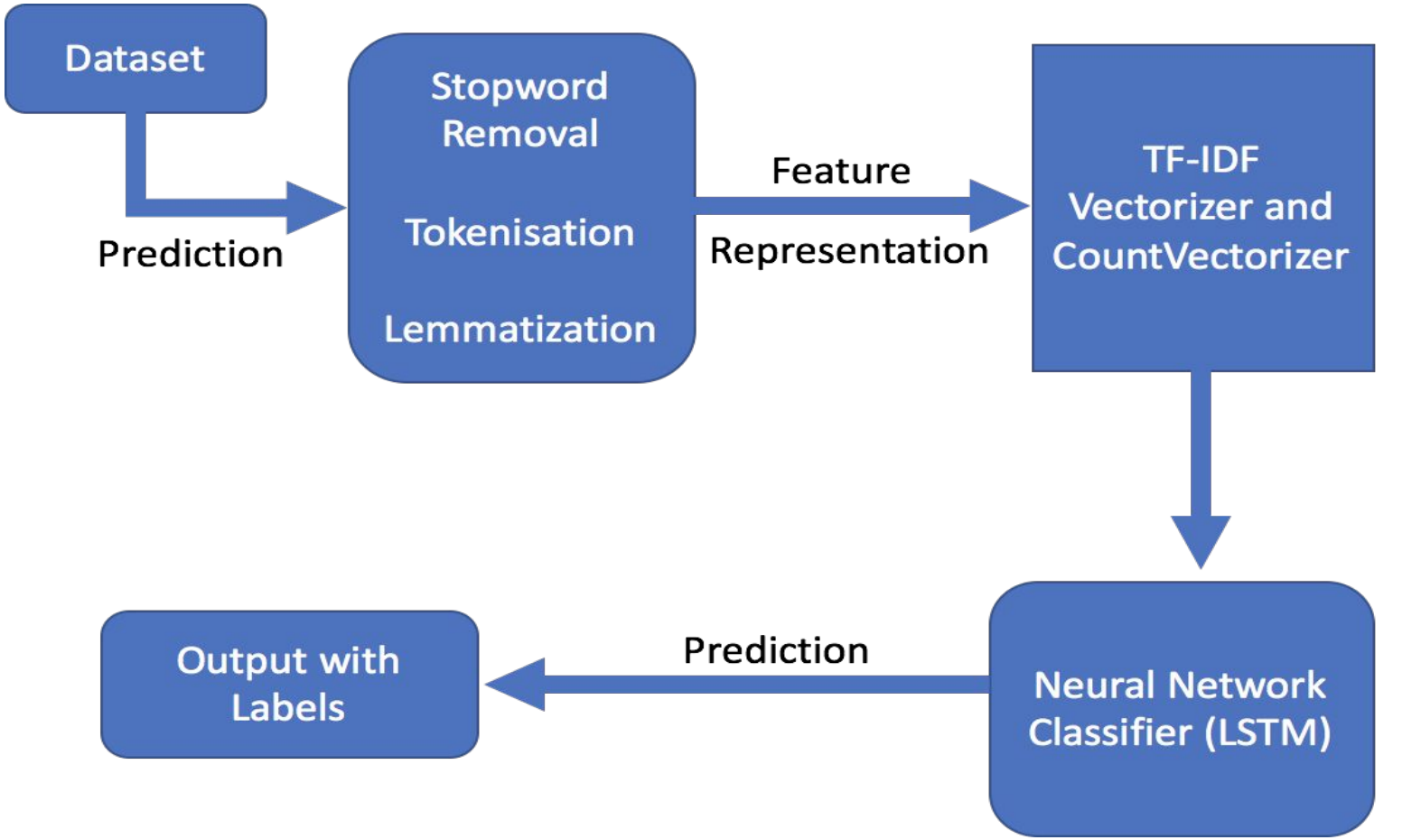
Tasks Performed

- **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
- Python Libraries Used for Date Extraction : datetime
- **Documentation** and **Presentation**

Development Phase



General Flow of Development



Development

- Downloaded OCW (MIT) dataset with the list of 12,500 + documents
- Automating the conversion of PDFs to TXTs 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
- Created stochastic gradient classifier model, input the values of Tfidf Vectorizer, hinge loss, learning rate as 0.001
- Trained the model and calculate the accuracy
- Tested it on the sample pdf and predict the label of sample pdf by using the trained model.

Result

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

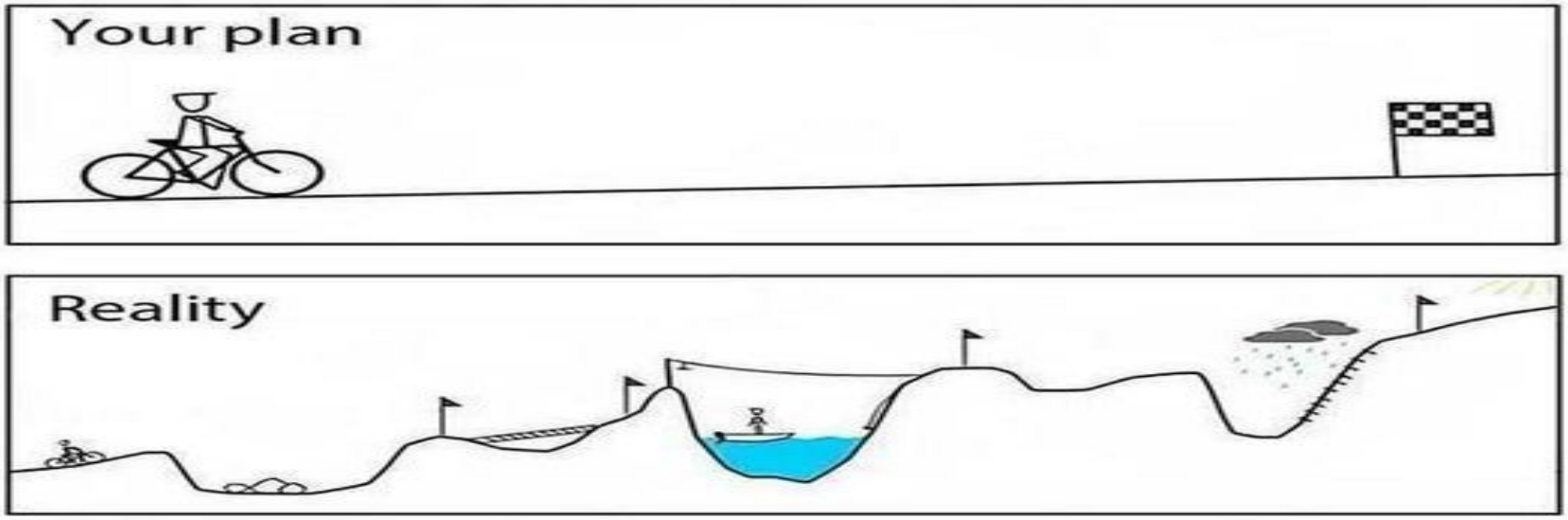
Result

```
python3.6 MIT_Department_Train.py
Done Extraction
Done splitting
Done training
Done prediction
0.891583166326653
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

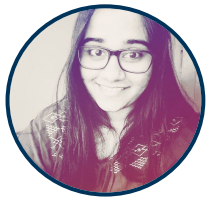
- How to work **part time** in a **fast growing project**
- How to **manage work** and **resources under** time constraints
- **Knowledge transfer** between team members



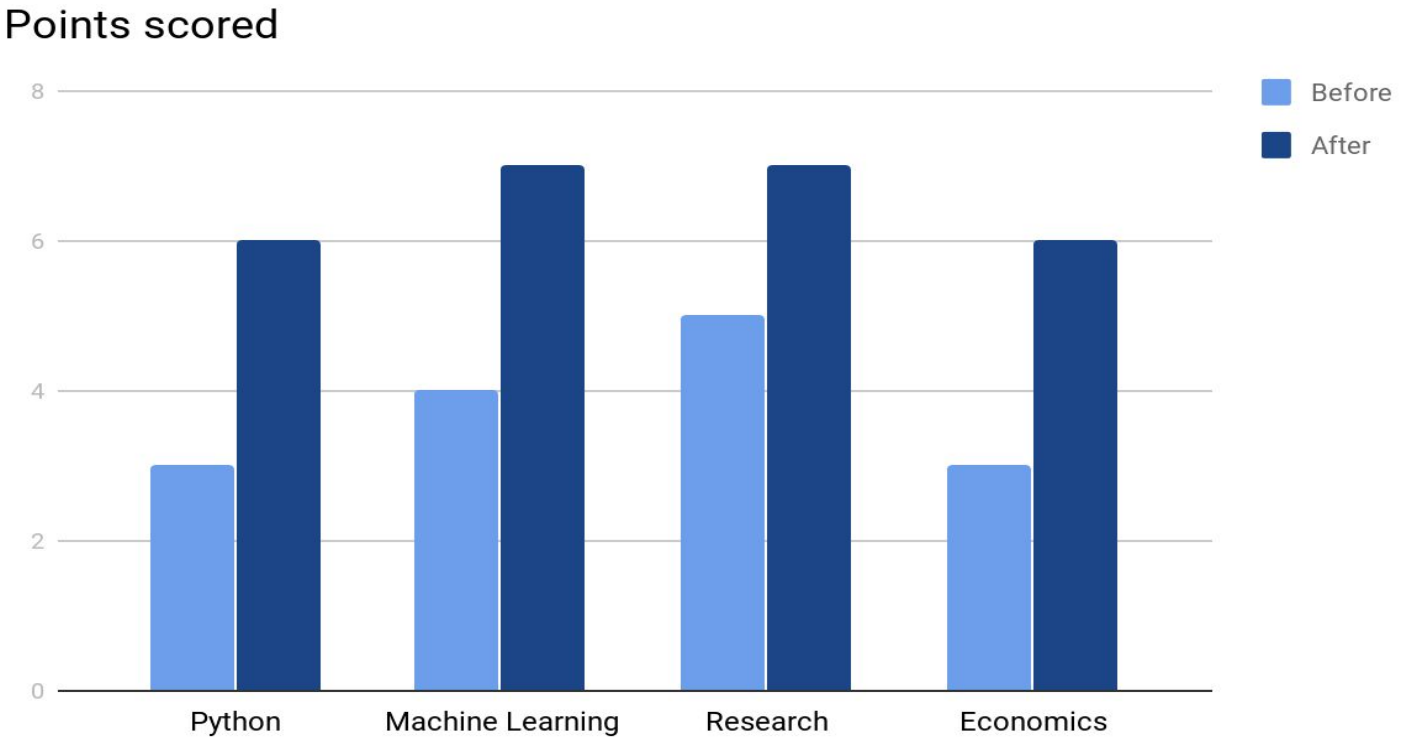
Technical Challenges

- 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



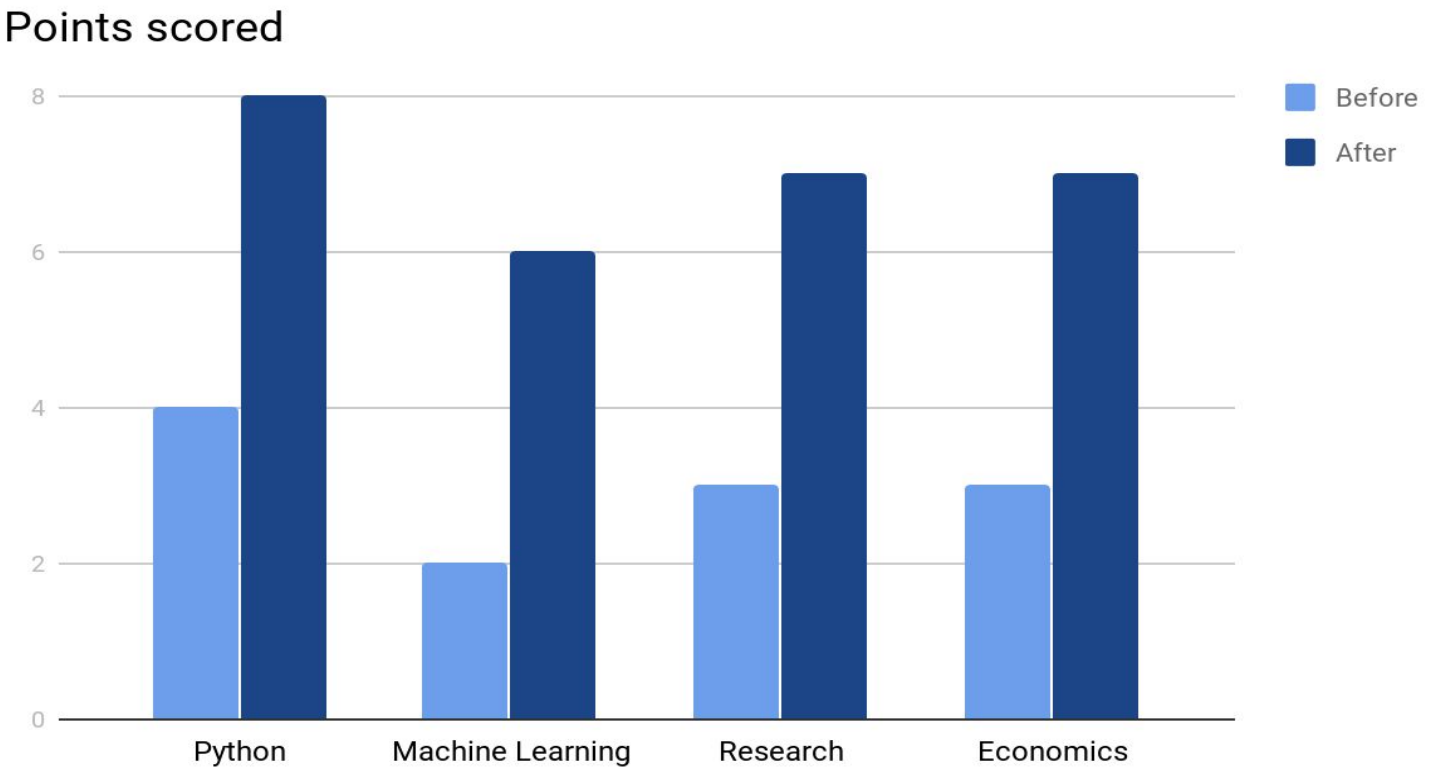
Vindhya Singh



Introduction | Development | C & L

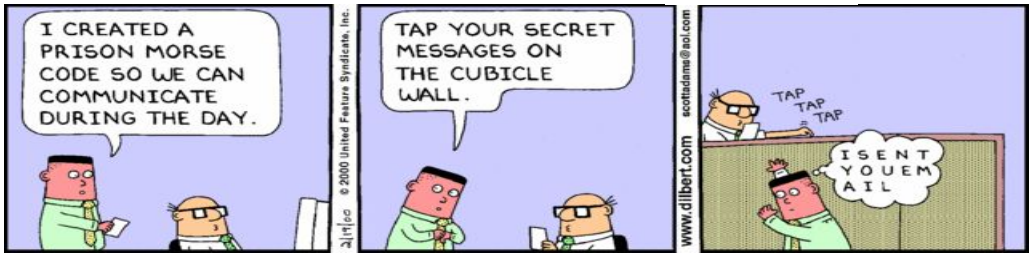
Learnings – Individual Learning Curve

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

Recommender System

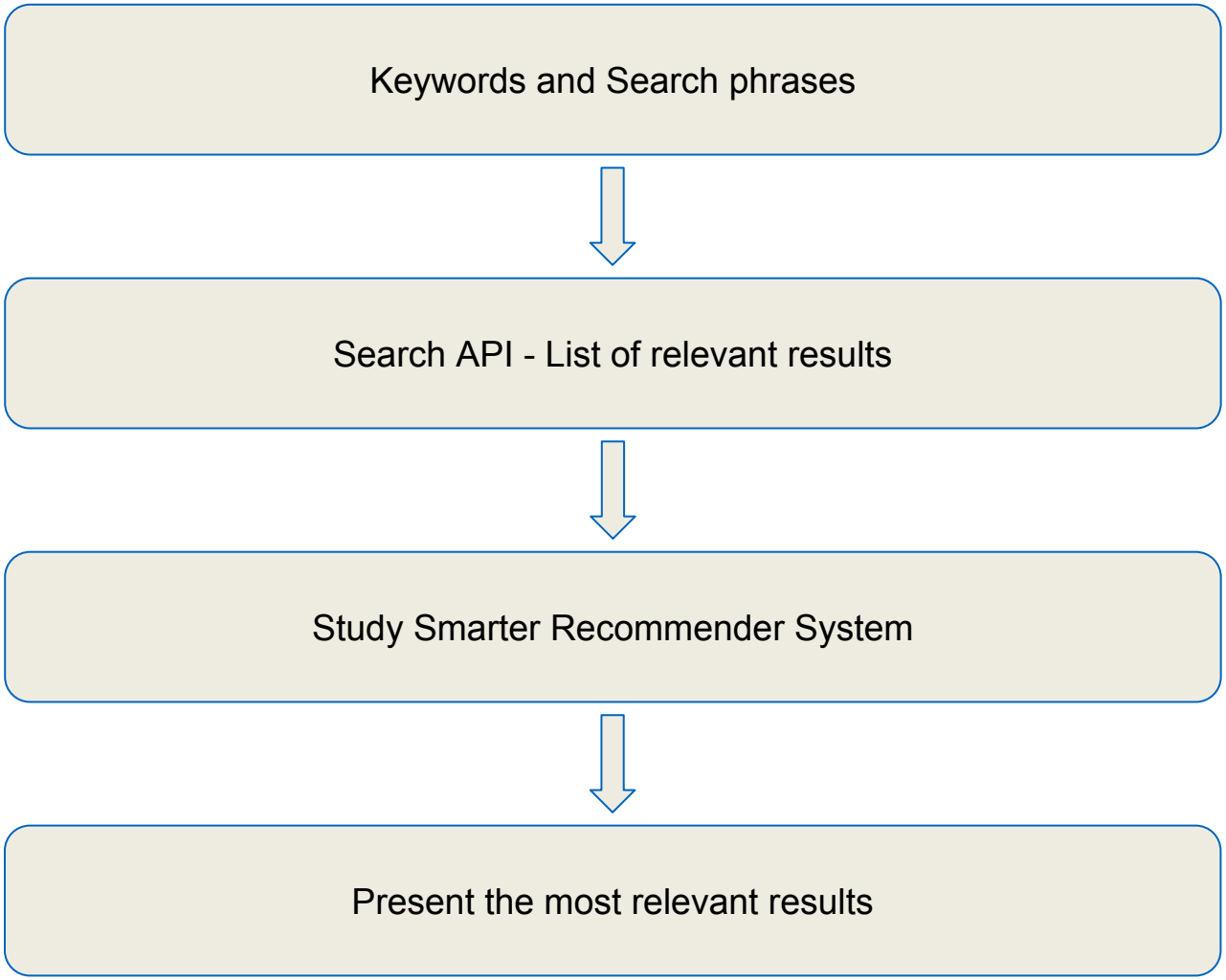
1. Introduction
2. Development
3. Challenges & Learning

Introduction

- While studying it is helpful to have some relevant information collected.
  - It saves time, by presenting you relevant information right in front of you
  - It helps you keep focused on the study material
  - It improves comprehension and learning

- What is relevant information ?
  - It can be be relevant definitions of important terms on the slide
  - It can be video lectures associated to the topic at hand.
  - It can be other similar courses

How do we do this ?



Youtube Search

- To get relevant videos from Youtube is an enormous task, as the search space is very large and always growing

- Disadvantages :
  - It can have a non zero cost
  - Or it can be restrictive with the amount of requests per day

- Advantages :
  - Returns the most accurate or relevant results for the search terms
  - Easy access to analytics useful for the recommendation system
  - Simplifies the problem of traversing a large search space
  - Easy to embed information about the videos in the StudySmarter interface
  - It is also possible to integrate a complete playlist related to the course

Custom Search

- To get relevant and useful textual information
  - Wikipedia snippets for information
  - Links about other courses on MOOCs similar to the current course
  - Blog posts or Pdfs explaining the topics

- Disadvantages :
  - Accuracy or usefulness depends on the keywords generated
  - Has a non-zero cost involved
- Advantages :
  - If the keywords are relevant, search results quality is high
  - Easy to modify and cherry pick favourable sources

Recommendation System

- To give good search results, the app has to collect information about the user as allowed by EU laws.
- The following data of the user needs to be monitored:
  - Relevancy rating
  - In app click counts
  - View count
  - Likes/Dislikes
  - In app ratings
  - In app view length
- Using this data, different algorithms using Item Similarity, Matrix Decomposition or Singular Value Decomposition algorithm can be employed to rate video/website links.

Additional Tasks

Data Collection

Dataset generation for Document Classification

- A dataset with multiple classes of labels for a large collection of documents
- These labels can reflect the granularity we want with the classification
  - Engineering -> Computer Science -> Artificial Intelligence -> Machine Learning
- Dataset should be unbiased with representation for a variety of fields of study
- Sources for the dataset should be organised and structured
- So the solution was to write a crawler to collect all the data, and then store it such that feeding it to classification algorithms is easy

Sources for the Dataset

- MIT OCW
  - Very Structured
  - All courses are university courses
  - The courses are labeled with high granularity
  - Lacks in Medicine courses

- Johns Hopkins OCW
  - To cover the lack of medicine courses in MIT OCW
  - Courses aren’t labeled with same granularity as MIT OCW

Sources for the Dataset

- Other sources that were considered were :
  - Open Yale Courseware
  - Open Michigan Courseware
  - NPTEL
  - Udacity
  - Coursera and Edx
  - Khan Academy
  - Online Stanford

- After combining the two sources we have a dataset with 15,000+ documents with around 900 labels

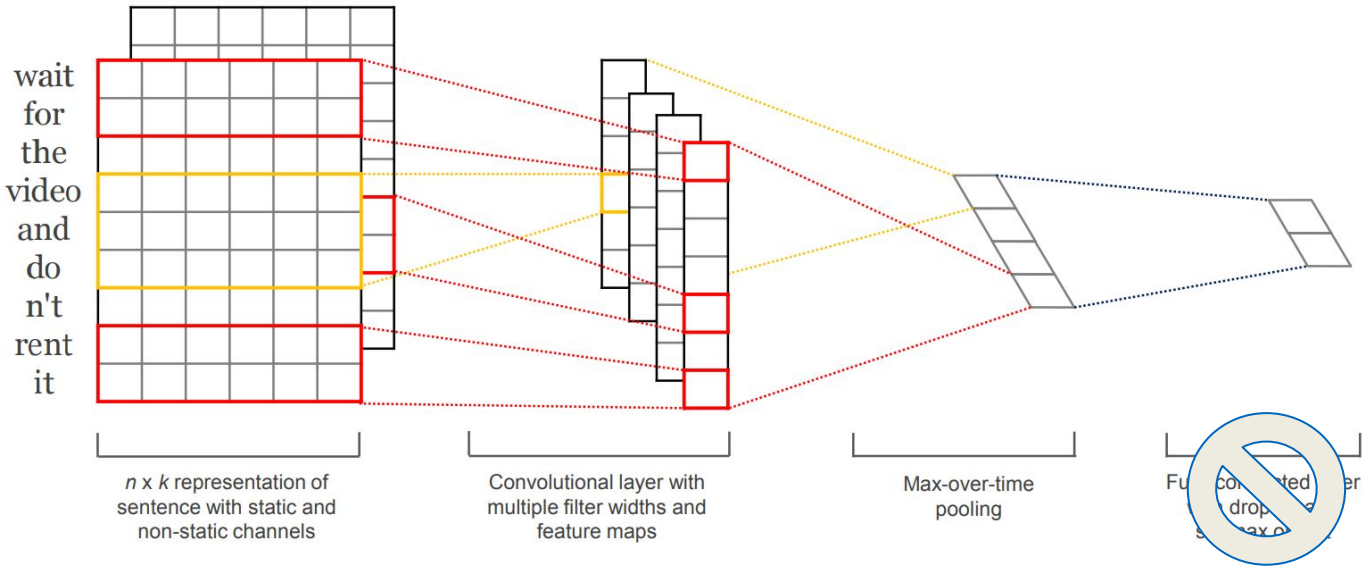
Advanced Development: Document Classification

Advanced Development: Document Classification

- The current development in NLP is shifting towards machine learning.
- Using deep learning in NLP has shown to give state of the art results.
- Industries have started to adopt deep learning for NLP such as:
  - CERN: Classification of high energy physics abstracts
  - IBM: Watson NLP Platform
  - Microsoft: Azure NLP Platform
  - Google: Google Cloud NLP Platform
- Implementation required researching new models completely different from existing standards.

Advanced Development

- Convolutional Neural Networks for Sentence Classification [Yoon Kim et al. 2014]
  - Use case: Classify sentence sentiment into Good or Bad



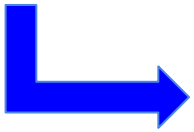
Advanced Development

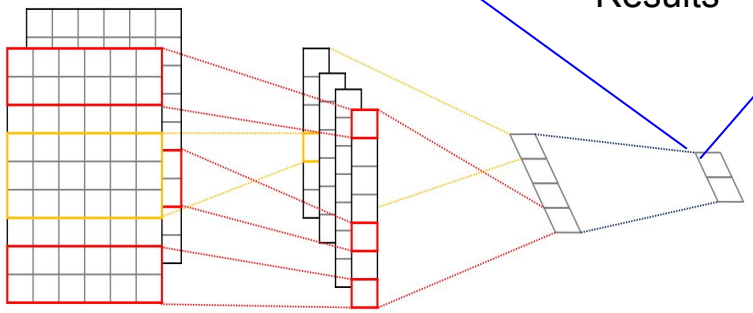
Flow of Data:

Dataset

Apple	0001
Orange	0002
Red	0003

Word to Vector





Class	Score [0, 1]
Physics	0.89
Chemistry	0.90
Economics	0.20

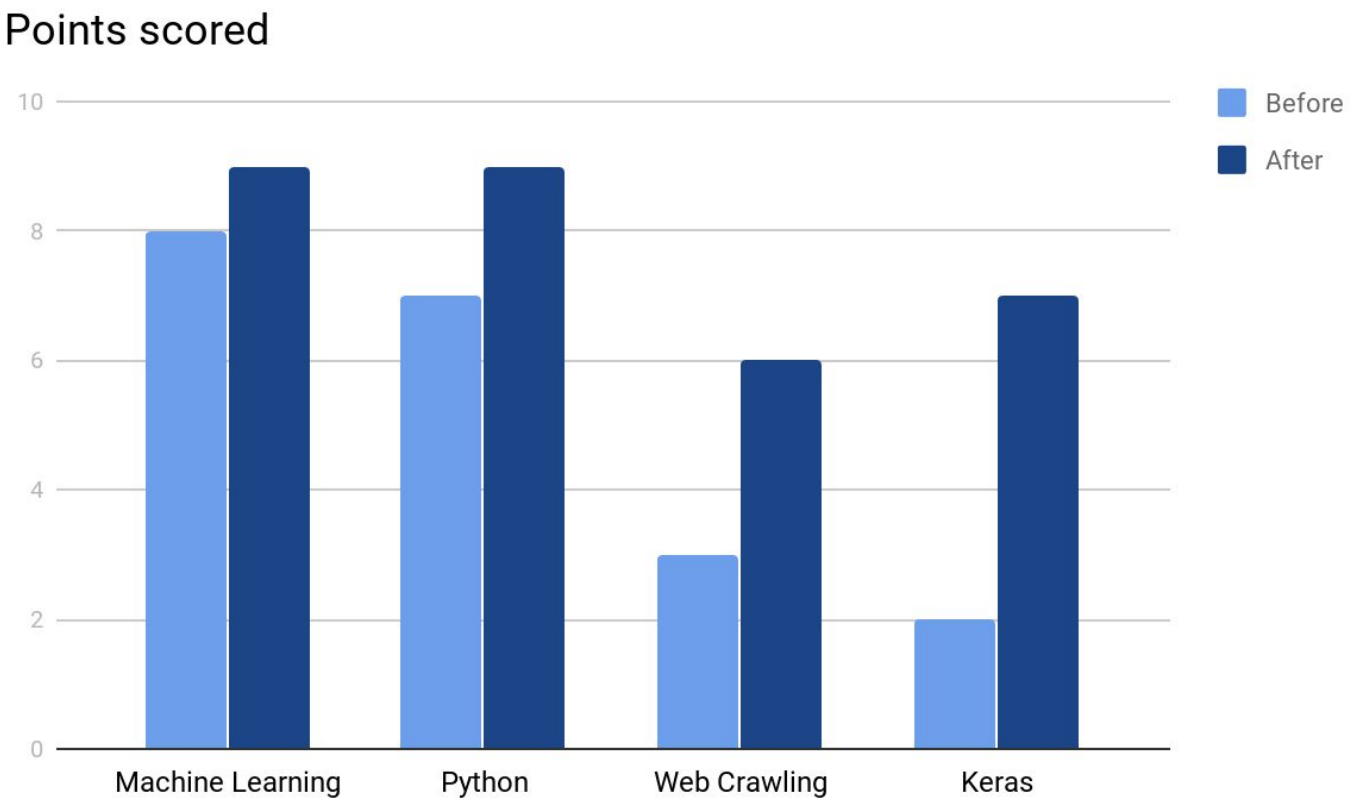
Results

Result - Advanced Development

Number of Labels Used	Prediction Accuracy
625	80.956%

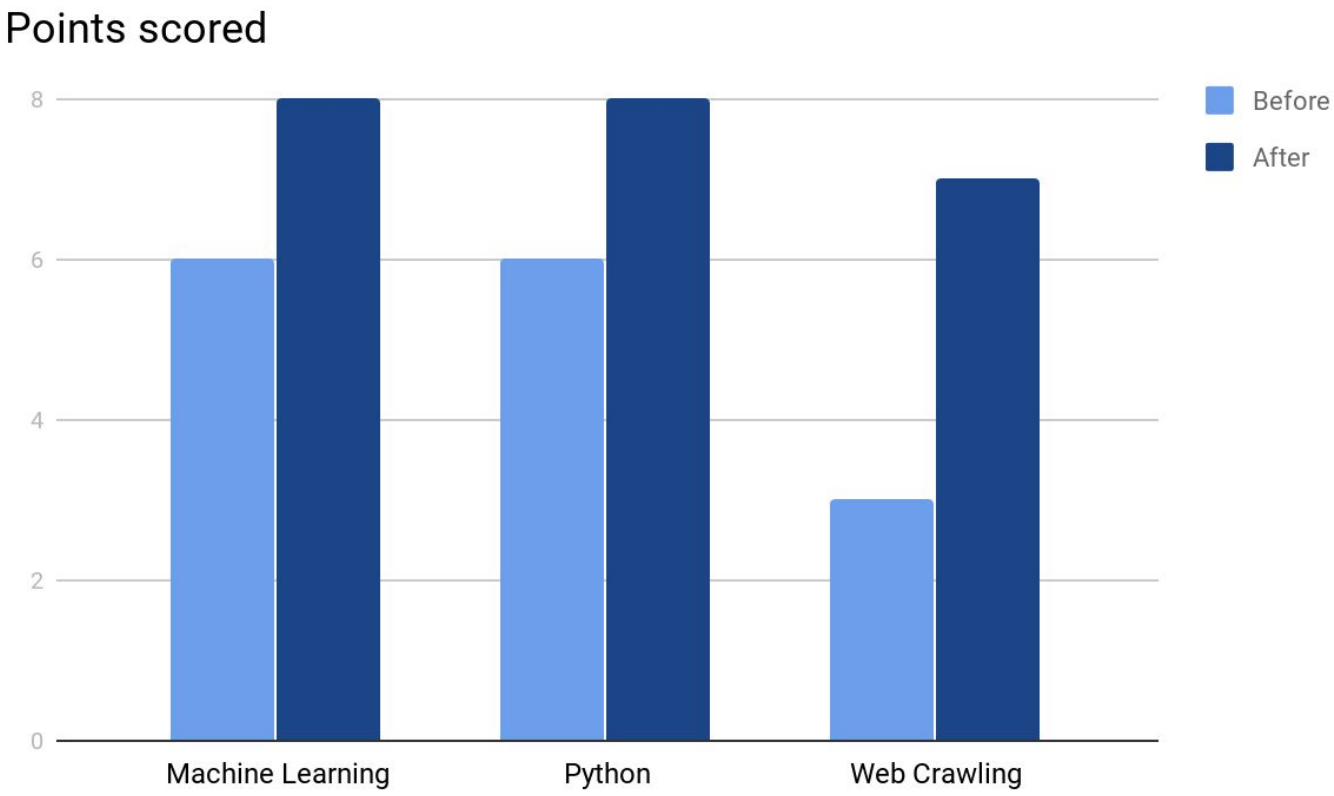
Learnings – Individual Learning Curve

Arvind Somasundaram



Learnings – Individual Learning Curve

Anshul Sharma



Challenges

- Searching for APIs that are cost effective as well as productive.
- Building recommender systems that can rank any data.
- Building a dataset large enough to encompass a wide variety of study fields and not be biased to one study area
- Researching state of the art models for classification and its implementation

Learnings

- Researched about cost effective techniques to maximize profits.
- Researched about existing APIs for gathering video links.
- Researched different online resources for building a dataset
- Used multiple factors to decide the ones to scrape to build a dataset for document classification
- Explored different strategies for building recommender systems.
- Built state of the art architecture for document classification from research papers.
- Trained model and evaluated results.
- Building an easy integration into the app.

Organization & Management

Project Communication and Organization

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

- Repositories for frontend and backend development
- Code reviews

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

Planning - Trello

That's all Folks!

Questions & Answers

IDP - StudySmarter

Thank you for your attention!

Technical University Munich

Supervisor Prof. Dr. Nicola Breugst

Munich, August 13, 2018

StudySmarter