Open Education Platform for Management Schools

Publikationstyp: Lehrmaterialien

Open Data Dokument-Tagging: Fallstudie für den Data Science Unterricht

Version Nr. 1, 27. Februar 2024

Tödtli, Beat
OST – Ostschweizer Fachhochschule
Wullschleger, Nicola
Stadt St. Gallen

Publiziert auf: www.oepms.org
Unter doi: 10.25938/oepms.394

Open Education Platform
for Management Schools

Open Data Dokument-Tagging: Fallstudie für den Data Science Unterricht

Version Nr. 1, 27. Februar 2024

Tödtli, Beat
OST – Ostschweizer Fachhochschule
Wullschleger, Nicola
Stadt St. Gallen

Publikationsform:	Fallstudie
Institution:	OST – Ostschweizer Fachhochschule
Schlüsselbegriffe:	Kaggle-Competition; Machine Learning; Dokument-Tagging; Data Mining; Data Science; Knowledge Discovery in Databases
Einsatzbereich:	Bachelorstudierende, Masterstudierende, Weiterbildung

Zitierweise nach APA:

Tödtli, B. & Wullschleger, N. (2024). Open Data Dokument-Tagging: Fallstudie für den Data Science Unterricht. *Open Education Platform*.
Doi: 10.25938/oepms.394

Open Education Platform
for Management Schools

Open Data Dokument-Tagging Fallstudie für den Data Science Unterricht

Beat Tödtli [a], Nicola Wullschleger[b]

[a] *Beat Tödtli, IPM- Institut für Informations- und Prozessmanagement, OST-Ostschweizer Fachhochschule, Rosenbergstrasse 59, 9000 St.Gallen, beat.toedtli@ost.ch*

[b] *Nicola Wullschleger, Fachspezialist Public Service Design und Innovation, Organisationsentwicklung Stadt St.Gallen, Poststrasse 28, 9001 St.Gallen,*

Abstract. Wir präsentieren eine Unterrichtssequenz, in welcher ein Data Science Anwendungsfall der St.Galler Open Data Plattform behandelt wird. Lernziel ist das vertiefte Verständnis der Data Mining Techniken zur Aufbereitung eines Datensatzes, wobei es vornehmlich um den Wissenstransfer in einen konkreten Anwendungsfall geht. InClass Kaggle Competitions eignen sich, um die Erarbeitung des Falls zu unterstützen.

Inhaltsverzeichnis

1. Einführung

Knowledge Discovery in Databases (KDD), Data Science, Data Mining und Machine Learning sind attraktive Themen und Fächer im Fachhochschulunterricht, werden aber von den Studierenden wegen ihrer Verwandtschaft zu Statistik und Informatik auch als Herausforderung wahrgenommen. Didaktisch minderwertige Online-Lehrmaterialien versuchen oft, Data Science ohne Bezüge zu Statistik zu vermitteln und arbeiten dabei oft mit vorgefertigten Codeblöcken und rezeptartigen Vorgehensdiskussionen. Dabei geht die Vermittlung einer zentralen Kompetenz verloren: Dem Urteilsvermögen in einer konkreten Situation.

KDD ist ein vielfältiger Prozess und erfordert eine Vielzahl von Kleinstentscheiden, welche oft auf einer inexakten Grundlage gefällt werden müssen. Zudem ergeben sich Bezüge zwischen diesen Einzelentscheiden, welche weitreichende Auswirkungen auf die Effizienz und Produktivität haben. Für ExpertInnen ist beispielsweise die effiziente Merkmalsauswahl eine Frage der Erfahrung, welche nur schwer vermittelt werden kann. Sie hat direkt Auswirkungen auf den Erfolg von nachgelagerten Schritten (etwa dem Modelltraining) und auf den ganzen Prozess und kann daher nicht isoliert betrachtet werden. Dieses schwer vermittelbare Urteilsvermögen ist beispielsweise gefragt, wenn die Evaluation des Gesamtprozesses unbefriedigend ist, aber noch keine Hinweise darauf bestehen, welche Komponenten verändert werden müssen. Nicht selten spielen kleine Änderungen am Lernprozess verhältnismässig grosse Performanzgewinne, jedenfalls, wenn an den richtigen Komponenten optimiert wird.

Ein anwendungsorientierter Machine Learning/Data Mining/KDD-Kurs sollte daher die praktische Auseinandersetzung mit diesen subtilen Fragestellungen betonen. Ein Fallbeispiel scheint ein geeignetes Instrument dafür zu sein: Erfahrungen mit den genannten Beurteilungsherausforderungen können fast nur durch Auseinandersetzung mit einer ganzheitlichen, anwendungsnahen und konkreten Aufgabenstellung erworben werden.

In Kapitel 2 präsentieren wir daher das Fallbeispiel einer Data Mining Anwendung der Open Data Plattform der Stadt St.Gallen. Die Open Data Plattform wünscht sich eine konkrete Erkennung bestimmter Dokumenttypen, um den Datensatz zugänglicher zu machen beispielsweise für politologische Analysen. Eine erste Kursdurchführung hat gute Resultate erbracht, das System wurde der Stadt St.Gallen betriebsbereit übergeben.

Kapitel 3 enthält die didaktische Reflexion der Fallstudie und widmet sich insbesondere den Erfahrungen mit einem speziellen Unterrichtsmittel für den Data Science Unterricht, den InClass-Kaggle Competitions. Einer der Autoren hat bereits viele Kurse auf Fachhochschulstufe mit diesem Mittel unterrichtet, und diese Publikation dient auch dazu, die dabei entwickelten Lehrmaterialien zu veröffentlichen. In Anhang C findet sich eine umfangreiche Liste mit frei zugänglichen Kaggle Competitions und zugehörigen Lehrmaterialien (sog. Notebooks).

2. Der Fall

2.1. Einführung

Die Stadt St.Gallen stellt nicht schützenswerte Verwaltungsdaten seit September 2019 auf einer Open Data Plattform (daten.stadt.sg.ch) zur Verfügung. Damit haben die Verwaltungseinheiten die Möglichkeit ihre öffentlichen Daten den Einwohnerinnen und Einwohnern, Unternehmen und Startups, oder Journalistinnen und Journalisten in strukturierter und maschinenlesbarer Form und zur freien Weiterverwendung zur Verfügung zu stellen. Damit leistet die Verwaltung einen Beitrag zur Förderung von Transparenz, Effizienz, Partizipation, und Innovation. Die Plattform verfügt bereits über eine Vielzahl an offenen Datensätzen und wird laufend erweitert.

Ein neuer Datensatz soll demnächst freigeschaltet werden, aber die Datenqualität sollte noch etwas erhöht werden. Der Datensatz ist Teil des Ratsinformationssystems der Stadtverwaltung und beinhaltet die traktandierten Geschäfte des Stadtparlaments St.Gallen. Den Sitzungen ist ein Downloadlink zum Protokoll und zu sämtlichen zugehörigen Dokumenten zugeordnet. Den einzelnen traktandierten Geschäften sind neben einem Downloadlink zum Traktandum, die Aktenplannummer, der Geschäftstyp, die Person, sowie die Partei zugeordnet. Über die Aktenplannummern sind den Geschäften Kategorien zugeordnet, welche die in den Parlamentssitzungen behandelten Themen codieren.

Die Aktenplannummern sind hierarchisch in 6 Ebenen gegliedert, jedoch ist die Annotationsqualität durchmischt: Bereits auf Aktenplanebene 2 gibt es eine allgemeine Sammelkategorie «Stadtparlament (früher Gemeinderat, Grosser Gemeinderat)» für sämtliche parlamentarischen Vorstösse (Interpellationen, Motionen und Einfache Anfragen), welche nicht inhaltlich kategorisiert wurden.

2.2. Zielvorgabe

Die Open Data Plattform der Stadt St.Gallen ist bestrebt, offene Datensätze möglichst hoher Qualität der Öffentlichkeit zur Verfügung zu stellen. Es soll untersucht werden, wie gut statistische Verfahren die Annotation von Texten in die Aktenplankategorien unterstützen können. Gelänge dies in einem grösseren Umfang automatisiert, könnte nicht nur der auf der Open Data Plattform bereitgestellte Datensatz mit dieser Zusatzinformation bereitgestellt werden- es wären auch ähnliche Analysen auf Parlamentsdokumenten anderer Städte möglich. Diese Technik könnte helfen, Codierungsfehler zu vermeiden. Darüber hinaus könnte eine Analyse des so erweiterten Datensatzes Aufschluss darüber geben, wie sich die Themen, mit denen sich das Parlament beschäftigt, über die Zeit verändern.

Das Ziel besteht insbesondere nicht nur in der Lösung eines Data Mining Problems, sondern in der Erarbeitung einer Lösung für die betreibende Organisation und die Nutzenden der Open Data Plattform der Stadt St.Gallen. Trotzdem kann die Fragestellung zunächst eingegrenzt werden auf die Erstellung und Optimierung eines Systems, welches Dokumenten «Tags», also Kategorien aus einer vorgegebenen Liste zuordnet. Ein Trainingsdatensatz ist dazu vorhanden, aber eben auch über zweitausend Dokumente, die noch getaggt, also klassiert werden müssen.

Anschliessend ist allerdings zu prüfen, wie das erstellte Data Mining Verfahren sinnvoll eingesetzt werden kann. Die Tagbestimmung kann entweder automatisiert oder in Form eines Empfehlungssystems für eine beschleunigte manuelle Zuordnung realisiert werden.

Open Education Platform
for Management Schools

Didaktische Reflexion

Open Data Dokument-Tagging Fallstudie für den Data Science Unterricht

Beat Tödtli [a], Nicola Wullschleger[b]

[a] *Beat Tödtli, IPM- Institut für Informations- und Prozessmanagement, OST-Ostschweizer Fachhochschule, Rosenbergstrasse 59, 9000 St.Gallen, beat.toedtli@ost.ch*

[b] *Nicola Wullschleger, Fachspezialist Public Service Design und Innovation, Organisationsentwicklung Stadt St.Gallen, Poststrasse 28, 9001 St.Gallen,*

Abstract. Wir präsentieren eine Unterrichtssequenz, in welcher ein Data Science Anwendungsfall der St.Galler Open Data Plattform behandelt wird. Lernziel ist das vertiefte Verständnis der Data Mining Techniken zur Aufbereitung eines Datensatzes, wobei es vornehmlich um den Wissenstransfer in einen konkreten Anwendungsfall geht. InClass Kaggle Competitions eignen sich, um die Erarbeitung des Falls zu unterstützen.

3. Didaktischer Reflexionsbericht

3.1. Einführung

Der in Kapitel 2 beschriebene Fall stellt eine typische Data Science Problemstellung dar. Wichtige Bearbeitungsschritte liegen in der Orientierung über die Datengrundlage, der Erarbeitung eines genauen Problemverständnisses, einer iterativen Lösungsentwicklung, Evaluierung und Optimierung sowie schliesslich dem Deployment. Die enge Verflechtung von Problemverständnis, Anwendungskontext und der Beurteilung der Lösungsqualität machen die Kompetenzen aus welche an Hand dieser Fallstudie eingeführt, geübt und reflektiert werden können.

3.2. Didaktisches Setting: Zielgruppe und Vorwissen

Dieses Fallbeispiel richtet sich an Studierende der Informatik, Wirtschaftsingenieurwissenschaften und der Wirtschaftsinformatik, welche in das Gebiet der Data Science einsteigen und interessiert sind, Data Science bis zur Bloom-Stufe «anwenden können» (Bloom et al., 1956) zu erlernen. Es eignet sich für Unterricht auf Bachelor- und Masterstufe oder in einem MAS in Data Science an einer Fachhochschule oder Universität. Die Fallstudie wurde bisher erst einmal eingesetzt. Die Kaggle-Competitions in Anhang C wurden in unterschiedlichen Kursen eingesetzt, etwa in einem Machine Learning-Kurs einer Informatik-Klasse und einer Wirtschaftsingenieursklasse auf Bachelorstufe und in zwei Durchführungen in der Master-Ausbildung der Wirtschaftsinformatik im Rahmen eines Data Science Moduls. Zudem kamen sie in einem CAS-Kurs zu Data Science zum Einsatz.

Vorausgesetzt werden Grundkenntnisse in Python. Zusätzliche Kenntnisse des Scientific Python Stacks müssen allerdings erarbeitet werden, und deren Vorhandensein kann den Arbeitsaufwand deutlich verringern. Vorwissen in Knowledge Discovery in Databases (KDD) sowie von Information Retrieval ist ebenfalls nötig, kann aber gut auch parallel zur Fallstudie eingeführt werden.

An statistischen Vorkenntnissen verlangt diese Fallstudie wenig: Eine «Hands-On»-Mentalität, wie sie in Data Science Tutorials oft anzutreffen ist, wird priorisiert, und die Bedienung von Application Programming Interfaces (APIs) ersetzt weitgehend ein detailliertes Verständnis von statistischen Verfahren. Beispielsweise wurden bedingte Wahrscheinlichkeiten im Rahmen der theoretischen Behandlung des Naive Bayes Verfahrens erläutert- in der Fallstudie hingegen wurde einfach eine neue Klasse (BernoulliNB von Scikit-Learn) instanziiert und benutzt, und Wissen über bedingte Wahrscheinlichkeiten wurde nicht weiter benötigt.

Das Thema Dokumentklassifikation ist im Gebiet des Natural Language Processing (NLP) zu verordnen, und theoretisches Vorwissen dazu ist sehr hilfreich und in engeren Grenzen unumgänglich. Tatsächlich aber kann das Fallbeispiel auch mit wenigen Vorkenntnissen angegangen werden: Der vorliegende Datensatz erzeugt auch mit den einfachsten NLP-Verfahren passable Resultate. Die fundierte Auseinandersetzung mit dem Thema erlaubt es zwar, in diesem Kontext leicht bessere Resultate zu erzielen, allerdings sind diese nicht entscheidend für die Erreichung der Lernziele, die in Abschnitt 3.3 genannt sind.

Dementsprechend eignet sich die Fallstudie in den Wirtschaftsinformatik- und Wirtschaftsingenieurstudiengängen vornehmlich im Masterstudiengang (MAS und MSc) sowie in spezialisierten Bachelorkursen. In Fachhochschul-Informatikstudiengängen waren ähnliche Kaggle-Competitions (vgl. Anhang C) im Bachelorstudiengang problemlos einsetzbar.

3.3. Lernziele der Fallstudie

Die folgenden Lernziele sollen mit dieser Fallstudie erreicht werden:

- Die Studierenden können den CRISP-DM-Prozess zur Umsetzung der Aufgabenstellung im Fallbeispiel selbständig abarbeiten
    - Sie können das Vorgehen für jeden Teilschritt erklären und die Besonderheiten des Fallbeispiels nennen insbesondere bezüglich der Anwendung und der Datenlage.
    - Sie verstehen, wie die Teilschritte miteinander zusammenhängen und können beispielsweise erläutern, wie Entscheidungen in den einzelnen Teilschritten Auswirkungen auf andere Teilschritte haben.
- Die Studierenden können die Performanz eines Dokumentenklassifikationssystems im Hinblick auf eine konkrete Anwendungssituation (wie jene des Fallbeispiels) beurteilen.
    - Sie können relevante Performanzkenngrössen zur Evaluation der Leistung des erstellten Tagging-Systems nennen und aus Ihrer Definition heraus begründen, weshalb diese relevant sind.
    - Sie können die Praxistauglichkeit eines Tagging-Systems in einer konkreten Situation (wie jener des Fallbeispiels) angemessen beurteilen.
    - Sie können die Konsequenzen von (fiktiven) niedrigen Performanzkennwerten für den Anwendungsfall des Fallbeispiels aufzeigen.

Der CRISP-DM-Prozess ist in Abbildung 1 visualisiert. Um obige Lernziele messbarer zu machen, werden für jeden der Prozessschritte präzisere Lernziele formuliert:

- Business Understanding: Die Studierenden verstehen die Wichtigkeit einer genauen Auseinandersetzung mit den Bedürfnissen der Kundschaft. Sie können erläutern, welche Konsequenzen eine falsche Erfassung der Bedürfnisse der Kundschaft auf den Projektverlauf, insbesondere das Zeitmanagement, hat.
- Data Understanding: Die Studierenden kennen einige der Datenqualitätsprobleme und können erläutern, welche Konsequenzen diese mit sich führen. Sie können aufzeigen, wie ein mangelndes Verständnis der Datenstruktur den weiteren Verlauf des Projekts beeinflussen kann.
- Data Preparation: Die Studierenden können weitgehend unstrukturierte Daten in einen Standardformat überführen. Sie können erläutern, inwiefern ein unsorgfältiges Vorgehen bei diesem Schritt Probleme in den nachfolgenden CRISP-DM-Prozessschritten mit sich führt.
- Modelling: Die Studierenden können erläutern, wie Anwendungsfälle, welche nicht unmittelbar als Klassifikations- oder Regressionsprobleme formuliert sind, in diese Form überführt werden können. Sie können insbesondere erläutern, wie Textklassifikatoren für ein Tag-Empfehlungssystem genutzt werden können.
- Evaluation: Die Studierenden verstehen, dass Performanzmetriken auf einen Anwendungsfall abgestimmt gewählt werden müssen
    - Die Studierenden können Performanzkennwerte effektiv (mit Visualisierungen) kommunizieren und dabei die (voraussichtlichen) Auswirkungen für den vorliegenden Anwendungsfall aufzeigen.
    - Die Studierenden können begründete, dem Fachkenntnisstand der Kundschaft angemessen formulierte Einschätzungen abgeben, ob das gebaute System eingesetzt werden soll.
    - Die Studierenden sind fähig, eine Kundschaft, welche die obige Performanzmetriken nicht kennt, verständlich über die Tauglichkeit eines DM-Systems zu informieren.

- Deployment:
    - Die Studierenden können den Nutzen von virtuellen Umgebungen erläutern und relevante Schwierigkeiten des Deployments (Testing, Kompatibilität, Dokumentation) an Hand ihrer eigenen Erfahrung erklären.
    - Die Studierenden können die praktischen Auswirkungen einer ungenügenden Testung eines Systems auf die Zufriedenheit der Kundschaft erläutern.
    - Die Studierenden können praktische Faktoren nennen, die den zeitlichen Aufwand für die Inbetriebnahme eines Softwaresystems beeinflussen.

Insgesamt ergibt sich die folgende Gegenüberstellung von Kompetenzen und Vermittlungsansätzen in der Fallstudie:

Kompetenzen	**Vermittlungsansätzen in der Fallstudie**
Fachkompetenzen	Indem ein komplettes Projekt bearbeitet wird, sollen (in variabler Tiefe) alle relevanten Kompetenzen angesprochen und unter Anleitung einmal durchgeführt werden
Alltägliche Beurteilungs- und Entscheidungskompetenzen in der Arbeit an einem KDD-Projekt	Die Fallstudie mit einem konkreten Kontext bietet die Gelegenheit, an vielen unterschiedlichen Stellen und insbesondere in den sechs CRISP-DM-Prozessschritten (vgl. **Abbildung *1***) auf den Einfluss einer Eigenart des Anwendungsfalls auf die Vorgehensweise zu verweisen.
Methodenkompetenzen	Angeleitetes Programmieren und die Bearbeitung von Softwareengineeringaufgaben zur Umsetzung der Aufgabenstellung des Fallbeispiels führt zur Vertiefung der Anwendungskompetenz der Instrumente wie Python, Data Mining-Tools wie Scikit-Learn oder Visualisierungsbibliotheken.
Selbstkompetenz	Ein längeres Projekt schult die Kompetenz, den aktuellen Status des Projekts zu beurteilen und daraus die konkreten nächsten individuellen Aktivitäten und Aufgaben abzuleiten. Fehlermeldungen oder Quantitative Resultate erfordern konstant ein flexibles Anpassen des weiteren Vorgehensplans.
Sozialkompetenz	Die Projektarbeit an diesem Fallbeispiel erfordert eine Aufgabenteilung und nicht erst für das Deployment eine enge Abstimmung der Teilaufgaben im Hinblick auf das Gesamtziel.

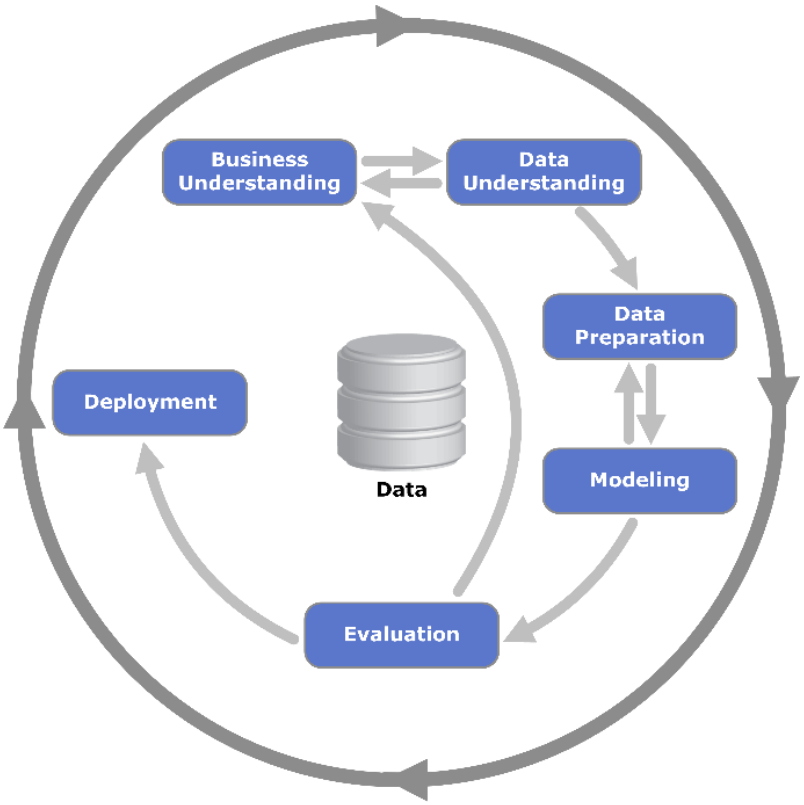


Abbildung 1: Der CRISP-DM Prozess
(Jensen 2012)

3.4. Analyse

3.4.1. Datengrundlage

Die zur Verfügung stehenden Daten bestehen zunächst aus einer Excel-Datei, welche Geschäfte, die im Stadtparlament behandelt wurden, auflistet. Relevant sind hier die folgenden Informationen zu den Geschäften:

- der Traktandumstitel
- die Aktenplannummer: auf Ebene 2 sind neben der Sammelkategorie 28 Kategorien vergeben
- Die URL zu den Traktandumsdokumenten: Ein FTP-Server stellt zip-Dateien mit den Traktandumsdokumenten zur Verfügung. Darin sind PDF-Dokumente vorhanden. Typischerweise beinhalten diese die Motionen der Politikerinnen und Politiker und allfällige weitere Unterlagen wie Stadtpläne etc.

Es gibt 2151 Parlamentsgeschäfte in der Sammelkategorie «Stadtparlament (früher Gemeinderat, Grosser Gemeinderat)». Für weitere 1721 Geschäfte ist eine geeignete Aktenplannummer vorhanden.

3.4.2. Lösungsansatz und Vorgehen für das Fallbeispiel

Um die Dokumente in der Sammelkategorie zu kategorisieren, können die Titel und Textdokumente der restlichen Parlamentsgeschäfte sowie deren bekannte Aktenplannummer (auf Aktenplanebene 2) benutzt werden, um zu lernen, welche statistischen Zusammenhänge zwischen dem Text und der Aktenplannummer bestehen. Ein solches System kann dann benutzt werden, um eine automatisierte

Zuordnung einer Kategorie vorzunehmen, oder alternativ, um Empfehlungen für die beschleunigte manuelle Zuordnung zu generieren.

Für die Bearbeitung der Data Mining-Aufgabenstellung bietet sich eine Bearbeitung in zwei Etappen an:

- Lernen einer Kategorisierung der Dokumente basierend auf dem Geschäftstitel
- Lernen einer Kategorisierung der Dokumente basierend auf den Parlamentsdokumenten. Diese müssen über die URL beschafft und aufbereitet werden. Anschliessend können umfangreichere Textanalysen z.B. der Motionstexte bessere Merkmale für die Kategorienvorhersage ermitteln.

Als didaktisches Mittel für die Umsetzung der Data Mining-Aufgabenstellung wird eine sogenannte Kaggle Competition eingesetzt. Dabei handelt es sich um eine Webplattform, welche den Studierenden ein Fallbeispiel mit den Trainingsdaten des Data Mining Problems zur Verfügung stellt und sie auffordert, eine Vorhersage für weitere Testdaten zu erstellen und auf die Plattform hochzuladen. Die Qualität jeder sog. Submission wird geprüft und es wird ein Leaderboard angezeigt, wo also die Rangliste der bisher besten Beiträge angezeigt wird. Nach Abschluss der Competition wird eine Siegerin bzw. ein Sieger gekürt. Weitere Details zu Kaggle Competitions sind im Anhang A (Kapitel 4) gegeben. Eine Liste der in dieser Fallstudie oder in ähnlichen Settings bereits benutzten Kaggle-Competitions mit deren Zugangslinks sind im Anhang C gegeben.

Im Anhang B (Kapitel 5) ist die Ausgangslage auf der Willkommensseite der Open Data St.Gallen Kaggle Competition wiedergegeben, wie sie den Studierenden präsentiert wird. Nach der Schilderung der Motivation für ein Tag-Empfehlungssystem und der Erklärung der Struktur des Datensatzes wird auf die Kriterien eingegangen, welche der Beurteilung von Wettbewerbsbeiträgen und letztlich zur Bestimmung der/des Gewinner*in herangezogen werden.

Für die Bearbeitung des Falls (über die Kaggle Competition hinaus) erfolgt anschliessend ein dritter Schritt:

- Die Erstellung eines Funktionsmusters, welches Vertreter*innen der Stadt St.Gallen präsentiert und ihnen dann übergeben wird. Zentral sind in diesem Bearbeitungsschritt die Bearbeitung der Evaluations-, Umsetzungs- und Inbetriebnahmefragen, welche sich aus den Zielvorgaben der Open Data Plattform der Stadt St.Gallen ergeben.

3.4.3. Erfolgskriterien für die Fallbearbeitung

Die Evaluation des Tag-Recommenders erfolgt in der Kaggle-Competition an Hand des Masses «Mean Average Precision @k»[1], das kurz als «MAP@k» bezeichnet wird. Dabei wird das System, welches k (Kategorie- oder Tag-)Empfehlungen bereitstellt, dafür belohnt, wenn die vorgeschlagene Kategorie gemäss Trainingsdatensatz korrekt ist. Es wird stärker belohnt, je besser die relevanten Kategorien priorisiert wurden.

[1] Die «Mean Average Precision at k» (Kurzform «MAP@k») ist eine gängige Metrik im Information Retrieval. Sie mittelt Precision@k-Werte für mehrere k-Werte, und über alle Benutzer des Systems. Vgl. z.B. (Mu Zhu, 2004)

Für den Erfolg im Rahmen der Fallstudie ist dieses Kriterium allein ungenügend. Weitere wichtige Kriterien sind:

- Wie angemessen ist die Performanz-Evaluation? Wurden beispielsweise statistische Unsicherheiten angegeben?
- Wurde ein plausibles und konkretes Szenario geschildert, wie das System in der Praxis nutzbringend eingesetzt werden kann? Erfolgt eine angemessene Interpretation der Performanzmetriken (z.B. MAP@3) in Bezug auf ihre Auswirkungen im Annotationsalltag des Anwendungsfalls?
- Vollständigkeit und Dokumentationsqualität: Wurde ein vollständiges, lauffähiges System abgeliefert? Wie gut ist es dokumentiert?
- Usability- und Stabilitäts-Kriterien: Wie ist die Einsatzfähigkeit und Bedienungsfreundlichkeit zu bewerten? Läuft das System? Lässt es sich neu installieren? Wie kann es gewartet werden?

Diese Punkte werden in der Abschlussbesprechung mit den fachlichen und technischen Ansprechpersonen der Stadt St.Gallen besprochen (siehe auch Abschnitt 3.5).

Zu welchem Grad die Fehler des entwickelten Systems toleriert werden können bestimmt weitgehend die Einsatzart des Funktionsmusters. Es muss also abhängig von der Leistung der entwickelten Systeme diskutiert werden, ob die vorgeschlagenen Kategorienzuordnungen den Dokumenten automatisch zugewiesen werden können, oder ob sie manuell nochmals überprüft werden müssen.

In der Tat ist davon auszugehen, dass eine manuelle Überprüfung notwendig ist. Das Fallbeispiel zeigt an dieser Stelle schön auf, dass Entscheidungsunterstützende sowie semi-automatisierte Data Mining-Anwendungen in diesem Fall sehr viel realistischer sind als vollautomatische, selbstentscheidende Systeme.

3.5. Unterrichtsdisposition

Die Unterrichtsinhalte und die parallellaufende Bearbeitung des Falls für ein 14-wöchiges Semester ist in untenstehender Disposition skizziert. Der Unterrichtsumfang beträgt 4 ECTS. Für die Bearbeitung der Fallstudie wird wöchentlich in der Präsenzveranstaltung ca. eine Lektion für die Diskussion von auftretenden Problemen und Koordinationsaufgaben zur Erreichung des Projektziels eingerechnet.

Woche	Aktivitäten im Präsenzunterricht	Aktivitäten im Selbststudium für die Fallstudie
1	Kurze Bekanntgabe des Modus des Leistungsnachweises Einführung in Data Science, Data Mining (CRISP-DM-Prozess), Machine Learning, Python IDE sowie Versionierung mit git	Einarbeitung in eine Python IDE, git und in die Data Science Bibliotheken Numpy und Pandas
2	Inputvortrag zu Open Data, Besuch der auftragnehmenden Partei und Besprechung der Projektziele Einführung von Kaggle Competitions und Kaggle Notebooks. Vorbereitung des Selbststudiums (Teilnahme an einer sehr einfachen Kaggle Competition, C5). Data Understanding und Data Preparation in CRISP-DM	Teilnahme an der Kaggle Competition C5, Studium der Datengrundlage und erste konzeptionelle Überlegungen zum Vorgehen
3	Einführung in Empfehlungssysteme und die Evaluation von Information Retrieval Systemen. Vorstellen und Einarbeiten in die Open Data St.Gallen Kaggle Competition	Teilnahme an der Open Data St.Gallen Kaggle Competition nur an Hand der Dokumenttitel

4	Abschluss der ersten Kaggle Competition Besprechung der Erfolgsmetriken der Kaggle-Competition. Einführung eines ersten Data Mining Klassifikationsverfahrens (Naive Bayes)	Teilnahme an der Open Data St.Gallen Kaggle Competition an Hand der Dokumenttitel
5	Data Preparation in CRISP-DM, Einführung von ersten Text-Merkmalsextraktionsverfahren (Bag of Words, TF-IDF). Besprechung des APIs zur Benutzung von Data Mining Klassifikationsverfahren	Teilnahme an der Open Data St.Gallen Kaggle Competition an Hand der Dokumenttitel und mit verschiedenen Klassifikatoren
6	Besprechung von NLP-Bibliotheken und des Baus und der Evaluation von Recommendersystemen. Besprechung der Möglichkeiten von Wahrscheinlichkeitsvorhersagen für den Bau und die Evaluation des Tag-Recommenders	Verbesserung der eigenen Leistung an der Kaggle-Competition, insbesondere der zweit- und drittbesten Tag-Empfehlung Visualisierungen der Recommenderleistung
7	Organisation des Deploymentprozesses: Dokumentation, virtuelle Umgebungen (pyenv und pipenv), Testung, Naive Bayes zur Spam-Mail-Erkennung	Erste Deployment-Anstrengungen, Code-Dokumentation
8	Entscheidungsbäume und Random Forests	Testen des Deployments, Evaluation der Eignung der Kaggle-Competition-Lösung für die Fallstudie. Arbeit an Verfahren, welche den Volltext der Parlamentsprotokolle berücksichtigt. Visualisierungen der Leistung des Prototyps
9	Perzeptron und Supportvektormaschinen	Evaluation weiterer Ansätze, insbesondere von BERT-basierten Datenvorbereitungsansätzen
10	Fragelektion mit der Kundschaft zur Übergabe des erstellten Funktionsmusters, Klärung der Anwendungsart und Evaluationskriterien für die Benutzungsfreundlichkeit einfache Neuronale Netze	Erstellen eines Fragekatalogs und weiteres Vorbereiten der Codebasis und des Deployments
11	Coaching zur Finalisierung der Fallstudie Neuronale Netze	Aufbereiten und Testen der Code-Basis und Erstellung der Präsentation und der Demonstration
12	Präsentation der Lösung (kollektiv und individuell), Beurteilung und Benotung der Resultate Deep Learning, Transformer	Nachbesserungen am Code

Tabelle 1: Unterrichtsdisposition im Rahmen eines 14-wöchigen Kurses

Der Präsenzunterricht soll allgemeine theoretisch Data Mining Themen und konkrete Machine Learning Verfahren behandeln. Die obige Disposition ist so aufgebaut, dass Themen, welche aus Sicht der Projektarbeit der Studierenden am Fallbeispiel gerade aktuell sind, gleichzeitig im Unterricht theoretisch behandelt werden.

In Woche 3 wird die Thematik der Empfehlungssysteme eingeführt und deren Bezug zu Standard-Klassifikationsverfahren beleuchtet. Gleichzeitig wird die Kaggle Competition C5 eingeführt, an welcher die Studierenden praktische Erfahrungen mit Data Mining-Werkzeugen sammeln und die Mechanik von Kaggle Competitions kennen lernen.

Ab Woche 4 beginnt der Übergang zur zentralen Kaggle Competition C1 und der Zuwendung zu den spezifischen Herausforderungen von Textrecommendersystemen: In Woche 4 wird mit Naive Bayes ist ein Klassifikationsverfahren behandelt, das sich sehr gut für Textdaten eignet. Ab Woche 5 wird im Präsenzunterricht auf ein vertieftes Verständnis von weiteren, komplexeren Machine Learning Verfahren hingearbeitet. Die Absicht dabei ist, dass die Studierenden im Selbststudium diese Verfahren ausprobieren und an der Kaggle Competition C5 direkt Erfahrungen machen können, inwiefern die gelernten Methoden einen Nutzen für das Fallbeispiel haben.

Viele dieser Verfahren benötigen nummerische Features als Input, weshalb in Woche 5 zunächst mal auf Textvektorisierungstechniken wie Bag of words und TF-IDF eingegangen wird. Random Forests (ein Klassifikationsverfahren) bieten dann erstmals die Möglichkeit, Wahrscheinlichkeiten vorherzusagen. Es muss anschliessend (in Woche 6) behandelt werden, wie diese zu Empfehlungen verarbeitet werden und wie unterschiedliche Wahrscheinlichkeitszuordnungen bezüglich der Qualität der Empfehlungen zu beurteilen sind.

Ab Semesterwoche 7 wird die Gewinner-Lösung der Open Data St.Gallen Kaggle-Competition evaluiert und zu einer vorgeschlagenen Lösung des Fallbeispiels weiterentwickelt. Die weiteren Inputs aus dem theoretischen Teil des Präsenzunterrichts fliessen danach als Ansätze für Verbesserungsoptionen in die Weiterentwicklung der Lösung des Fallbeispiels ein. Wie erwähnt ist hier darauf zu achten, dass das primäre Ziel, am Ende des Semesters eine funktionierende Lösung präsentieren zu können, nicht beeinträchtigt wird. Projektorganisationsfragen stellen sich nun verschärft: Es ist sinnvoll, dass sich ein Teil der Studierenden auf die Finalisierung eines ersten Funktionsmusters konzentriert (wobei auch Dokumentation, Testung und Deployment bei der Kundschaft in den Vordergrund rücken), während ein anderer Teil das Potential von möglichen Verbesserungs- und Weiterentwicklungsansätzen ermittelt. Wenn ein Verbesserungsansatz erfolgreich ist, kann die Finalisierung auf Grund der Erfahrungen der ersten Gruppe oft effizient nachvollzogen werden. Zu beachten ist dabei der Mehraufwand, der durch unterschiedliche Versionen und Funktionsmuster entstehen kann. Beispielsweise kann muss die Bedienungsanleitung sowohl das erste, einfachere System wie auch das System mit dem verbesserten Ansatz dokumentieren.

In der letzten Semesterwoche erfolgt die Präsentation der Lösung sowie die Übergabe des Codes. Zentral ist hier, die angetroffenen transferorientierten Herausforderungen zu thematisieren, und einen Konsens über die Performanzevaluation zu erzielen. Insbesondere geht es um die Kommunikation des tatsächlichen Mehrwerts, welcher Data Mining in diesem Fallbeispiel gebracht hat, sowie um die Herausforderungen der Inbetriebnahme des Systems z.B. für zukünftige, ähnliche Problemstellungen. Bei zusätzlich vorhandenen Lektionen könnte hier noch verstärkt auf die Visualisierung, auf GUI-Applikationen mit Visualisierungen der Data Science Resultate eingegangen werden. Ebenfalls relevant sind Fragen über das Vorgehen bei ev. nötigen Weiterentwicklungen (z.B. Fehlerbehebungen), nachdem aus Studierendensicht das Projekt abgeschlossen wurde.

3.6. Breite und Tiefe der Fallstudie

Diese Fallstudie beginnt mit der allgemein formulierten Zielsetzung des Open Data Portals der Stadt St.Gallen, Mehrwert aus den von Ihnen zur Verfügung gestellten Daten zu ziehen. Um die einzelnen Herausforderungen aufzuzeigen, die sich bei der Konkretisierung dieser Zielstellung ergeben, wird die Fragestellung der Fallstudie anschliessend stark eingeschränkt auf eine limitierte Anwendung, deren Mehrwert abnimmt, je überschaubarer und realistischer die Herausforderungen werden. Die

Beleuchtung aller konkreten Aspekte von Text Data Analytics wird dann an diesem limitierten, aber konkret erfahrbaren Anwendungsfall aufgezeigt.

Der Zwischenschritt, in dem die Lösung der Tag-Zuordnung als Kaggle-Competition aufgesetzt wird, engt den Fokus der Fallstudie zwischenzeitlich stark ein. Danach wird aber durch den Fokus auf die Abgabe eines lauffähigen Funktionsmusters rasch auf die ursprüngliche breite Zielsetzung, einen echten Mehrwert für die betreibende Abteilung des Open Data Portals zu bieten, zurückgewechselt.

Insgesamt deckt dieses Projekt den gesamten KDD-Prozess ab. Der gegebene Datensatz bewirkt eine Fokussierung auf Data Mining Techniken im Textbereich. Bild- oder andere Sensordaten sowie deren Klassifikationsverfahren beispielsweise stellen Themen dar, welche nicht behandelt werden.

3.7. Didaktische Hinweise

In diesem Abschnitt sollen einige didaktische Aspekte dieser Fallstudie diskutiert werden.

3.7.1. Technische und didaktische Herausforderungen von Kaggle-Competitions für den Data Science Unterricht

Auf Fachhochschulstufe ist für die Eignung von Kaggle-Competitions für den Data Science Unterricht hauptsächlich die Selbstorganisationsfähigkeit und das Data Science Vorwissen der Studierenden ausschlaggebend. Gute Vorkenntnisse in Python und dem Scikit-Stack (insbesondere den Bibliotheken Pandas, Scikit-Learn, Matplotlib, Numpy, Seaborn etc.) sind hilfreich. Mit reduzierten Kenntnissen verlagern sich die Herausforderungen zunächst in Richtung syntaktische Beherrschung dieser Pakete. Ein Kurs in Python für Data Science ist auf jeden Fall zur Vorbereitung empfehlenswert.

Studierende mit wenig Programmiererfahrung hatten zunächst Mühe, den Code einer gesamten Wettbewerbseingabe zu überblicken und sich selbständig anzueignen. In solchen Fällen kann der Betreuungsaufwand gross werden, was im Präsenzunterricht zu unproduktiven Zeiten führen kann. Es hat sich in solchen Fällen gezeigt, dass es Sinn macht, eine erste, sehr einfache Kaggle-Competition einzusetzen (vgl. Competition C5 «Schnell-Mal-Klassifizieren», Anhang C). Daran wurde der Datenvorverarbeitungs- und Data Mining-Aufwand auf ein Minimum reduziert, und es wurde ein vorbereitetes Notebook (vgl. Notebook N3, Anhang C, Kapitel 6) abgegeben, welches die Studierenden durch eine erste Wettbewerbseingabe führt.

Technische Schwierigkeiten wurden beobachtet, wenn mehr als 10 Studierende gleichzeitig ein Kaggle Notebook ausführen. Ebenfalls ist zu beachten, dass Competitions nicht mehr rekonfiguriert werden können, wenn sie einmal gestartet sind.

3.7.2. Kaggle Competitions als didaktisches Instrument

Generell bestätigte sich wiederholt die Erfahrung, dass Kaggle Competitions ein hoch motivierendes, aber fortgeschrittenes Instrument zum Erlernen von Data Science und insbesondere Data Mining ist. Die Motivation ergibt sich meist aus dem kompetitiven Format, welches aber –wie hier geschehen– gut in kollaboratives Lernen umgesetzt werden kann.

Für Studierende mit wenig Python-Vorkenntnissen hat sich gezeigt, dass gut vorbereitete Arbeits-Notebooks sehr wichtig sind. Ziel sollte es sein, dass diese Studierende rasch «selbständig» eine erste Kaggle-Submission vornehmen können. Ein so vorgefertigtes Code-Beispiel birgt aber auch die Gefahr, dass die Studierenden den Code laufen lassen, ohne wirklich zu verstehen, was passiert. Trotzdem können die Studierenden auf diese Weise an den Code herangeführt werden (wie beispielsweise für die

«titanic-privat»-Competition C10 mit den Notebooks N10.1 und N10.2, vgl. Anhang C). Nach einer ersten erfolgreichen Submission empfiehlt es sich, auf die Hyperparameter der Klassifikatoren hinzuweisen (beispielsweise der Tiefe eines Entscheidungsbaums) und vorzuschlagen, diese Parameterwerte mal zu optimieren. Ein zweites, fortgeschritteneres Notebook kann dann auf diagnostische Visualisierungen eingehen (und z.B. Validierungs- und Lernkurven anzeigen), um die Klassifikatoren zu optimieren.

3.7.3. Einsatz von Kaggle Competitions aus Sicht der Lehrperson

InClass Kaggle Competitions aufzusetzen ist technisch ein weitgehend reibungsloser Prozess. Aufwändig ist insbesondere das Finden einer geeigneten Klassifikationsaufgabe und die Erstellung des zugehörigen Datensatzes. Es ist die Situation wie in der «Kreditkartendatensatz»-Competition zu vermeiden (vgl. Competition C6, Anhang C), wo mit wenig Aufwand ein Score von 79% erreicht werden kann, während gleichzeitig ein Wert von 83% kaum erreichbar ist.

Die Dokumentation von fortgeschrittenen Aspekten des Aufsetzens von Kaggle Competitions ist teilweise mangelhaft. Beispielsweise findet sich nirgends eine genaue Definition der einsetzbaren Erfolgsmetriken, etwa der mean average precision in der Competition C1 (vgl. Anhang C).

Technische Schwierigkeiten haben sich ergeben, wenn viele Studierende gleichzeitig eine Kaggle Notebook-Instanz starten. Die Aufwände zum Erlernen der Mechanismen, wie an einer Kaggle Competitions teilgenommen werden kann, erfordern ca. 1 Lektion sowie 1-2 Lektionen Selbststudium.

3.7.4. Erfahrungen mit der Open Data St.Gallen Kaggle Competition

Während viele der bisher erwähnten Kaggle Competitions in 1-2 Präsenzeinheiten behandelt werden können, ist die Open Data St.Gallen Kaggle Competition anspruchsvoller, dafür auch wesentlich realitätsnaher als andere referenzierte, von einem der Autoren im Unterricht eingesetzten InClass Kaggle Competitions. Die Open Data St.Gallen Kaggle Competition (Competition C1, vgl. Anhang C) lässt sich in der Variante «Trainieren auf dem Geschäftstitel» innert 2-3 Semesterwochen lösen. Die Studierenden erreichen damit einen MAP@3-Score von ca. 88-91%, während das einfachste Demonstrationsbeispiel der dozierenden Person (Notebook N1, vgl. Anhang C) einen Score von 83% erreicht. Fortgeschrittene Studierende, welche eine Kategorisierung der Dokumente basierend auf den Parlamentsdokumenten versuchen, erreichten einen Score von 93%. Diese Werte können als hinreichend gut bezeichnet werden, um darauf aufbauende Anwendungsfälle zu motivieren.

Ein möglicher Nachteil der Open Data St.Gallen Kaggle Competition ist das verhältnismässig aufwändige Aufbereiten der gegebenen Rohdaten. Eine ähnliche Situation ergab sich bei einer Bildklassifikations-Competition (Competition C2, vgl. Anhang C, Beispielbilder siehe **Abbildung 2**), bei welcher von Studierenden erstellte Bilder von handgezeichneten Kreuz-, Kreis- und Plus-Symbolen erkannt werden sollten. Die Bilddateien wurden in einer Vielzahl von Formaten erstellt (Unterschiede im Dateiformat, in der Auflösung, dem Farbraum, dem Vorhandensein eines Alpha-Kanals etc.). In einer späteren Durchführung wurden diese Daten vorverarbeitet (Notebook N2.2, vgl. Anhang C) und als .csv-Datei abgegeben (vgl. Notebook N2.1, Anhang C). Damit konnte im Unterricht der Fokus stärker auf Erkennungsverfahren wie der Hauptkomponententransformation oder convolutional neural networks gelegt werden. Eine ähnliche Entwicklung ist auch für die zweite Durchführung der Open Data Kaggle Competition zu erwägen.

Abbildung 2: Trainingsbilder der Kreuz-Kreis-Plus-Kaggle Competition
(Eigene Darstellung aus Bildern der Kaggle Competition)

3.7.5. Erfahrungen in einem 14-wöchigen Semesterkurs

Die bisherige Durchführung der Fallstudie in einer Machine Learning Klasse für WirtschaftsingenieurInnen auf Bachelorlevel hat gezeigt, dass der Fall selbst dann bearbeitet werden kann, wenn die Studierenden zu Kursbeginn keine Python-Vorkenntnisse besitzen. Es wurde aber auch klar, dass die zeitliche Belastung hoch sein kann, wenn mit wenigen Vorkenntnissen gearbeitet wird. Die Studierenden meldeten, dass die Anforderungen des realen Anwendungsfalls den Druck auf sie merklich erhöht hat, eine funktionierende und präsentable Lösung abzuliefern. Für viele Studierende blieb die Motivation während des Kurses hoch, den Fall erfolgreich abzuschliessen und zu präsentieren. Es soll nicht verschwiegen werden, dass die Koordinations- und Finalisierungsarbeiten für eine präsentable Lösung in den letzten 1-2 Wochen auch für die dozierende Person aufwändig werden kann. In dieser Phase haben sich aber auch die Studierenden mehrheitlich stark für einen erfolgreichen Abschluss eingesetzt.

Insgesamt ergibt sich daher i.A. eine sehr positive Erfahrung. Das zentrale Lernziel, ein vertieftes Verständnis des Data Science Entwicklungsprozesses zu erreichen und diesen auch praktisch teilweise selbständig durchführen zu können, sind teilweise erreichbar. Die Fallstudie zeigt in der Tendenz eher den vollständigen Prozess auf, als dass auf die Beherrschung einzelner Aspekte durch die Studierenden fokussiert würde. Teilweise kann die Evaluation der Qualität eines Entscheids für die eine oder andere Vorgehensoption zurücktreten gegenüber dem Ziel, zumindest eine Lösung lauffähig und fristgerecht abliefern zu können.

Es bleiben wichtige Verbesserungsmöglichkeiten. Der eher grosse zeitliche Umfang der Fallstudie führt zum Wunsch, Teile der Fallbearbeitung zu kürzen oder besser vorzustrukturieren. Die zeitliche Planung der Aktivitäten zur Inbetriebnahme der Lösung darf nicht unterschätzt werden. Der Tendenz, dass die Studierenden zu lange an den Verfahren optimieren, um gute Resultate für die Kaggle-Competition zu erzielen, muss mit einer klareren zeitlichen Strukturierung entgegengewirkt werden, damit das Testen, die Dokumentation und die Inbetriebnahme der finalen Lösung nicht zu kurz kommen. In zukünftigen Durchführungen muss auch deshalb früher mit diesen Aktivitäten begonnen werden, damit genügend Zeit bleibt für die Einforderung und Bearbeitung von Nachbesserungsanforderungen seitens des Dozierenden oder des Kunden/der Kundin.

Anhang A: Kaggle Competitions

Kaggle.com ist eine Webseite, welche eine frei verfügbare Data Science Infrastruktur[2] sowie viele Lernmaterialien bietet. Eine Lernform, die sich in bisherigen Durchführungen von Data Mining Kursen für Informatikstudierende sehr bewährt hat, sind InClass Kaggle Community Competitions[3]. Sie erlauben das Stellen einer konkreten Data Science Klassifikationsaufgabe, in welcher ein Trainingsdatensatz und ein Wettbewerbsdatensatz bereitgestellt werden. Eine Beispielseite ist die «Titanic privat» Kaggle-Competition (Competition C10, vgl. Anhang C). Die Willkommensseite ist in **Abbildung 3** gezeigt.

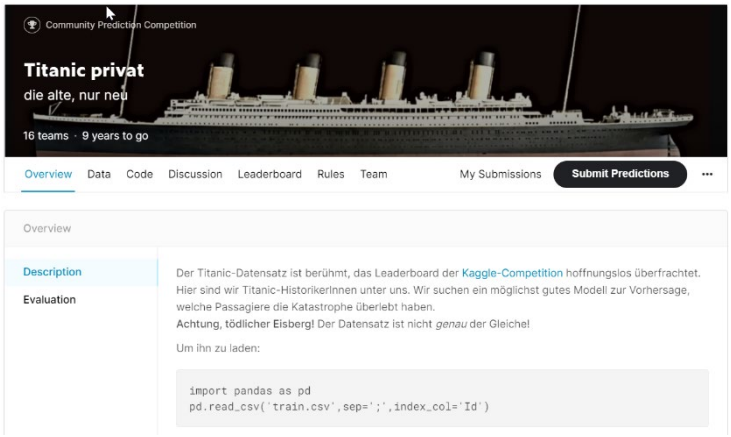


Abbildung 3: Willkommensseite einer InClass Kaggle Competition
(Kaggle.com, 2021)

In der einfachsten Form steht den Studierenden nach Einrichtung durch die Lehrperson eine Kaggle-Wettbewerbsseite zur Verfügung (Competition C10, vgl. Anhang C), auf welcher die folgenden Funktionen bereitgestellt sind:

- Das Fallbeispiel mit den relevanten Informationen zur Aufgabenstellung
- Die zur Teilnahme an der Competition benötigten Daten sowie deren Erläuterungen (Merkmalsbeschreibungen, Beispielformate etc.)
- Eine Submissionsseite, auf welcher die Wettbewerbsteilnehmenden ihren eigenen Wettbewerbsbeitrag hochladen können
- Ein Leaderboard, auf welchem die Wettbewerbsteilnehmenden nachsehen können, wie lange der Wettbewerb noch läuft, und auf welcher Position in der vorläufigen Rangliste sie und ihre Mittbewerber stehen.

[2] Siehe https://www.kaggle.com/docs/notebooks

[3] Siehe auch den Community Setup Guide, sowie die allgemeinen Einführungen zu Kaggle Competitions und Kaggle Community Competitions.

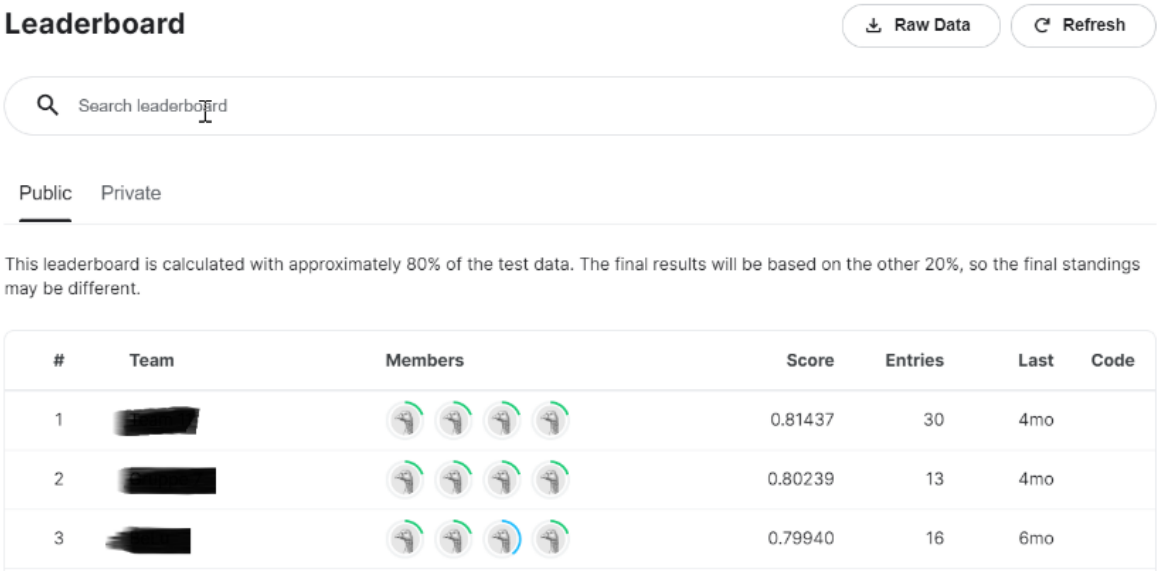


Abbildung 4: Leaderboard einer InClass Kaggle Competition
(Kaggle.com, 2021)

Abbildung 4 zeigt das Leaderboard einer InClass Kaggle Competition. Angezeigt werden für alle Teilnehmenden deren aktuelle Punktezahl (oftmals ist dies die erreichte Klassifikationsgenauigkeit) und deren Rang. Ein zentrales didaktisches Element besteht in der Motivierung der Teilnehmenden durch diese Darstellung, die Competition zu gewinnen. Wichtig ist es zu betonen, dass der/die GewinnerIn des Wettbewerbs auf einem separaten Teil des Testdatensatzes ermittelt wird. Das «private Leaderboard», welches den Wettbewerbsgewinner bestimmt, wird erst nach Abschluss des Wettbewerbs freigegeben. Die Implikationen dieser Regel für das Design einer möglichst erfolgsversprechenden Wettbewerbseingabe sind wichtige Inhalte des Präsenzunterrichts.

Nebst dem obigen kompetitiven Element bietet die Kaggle Competition-Seite auch mehrere kollaborative Elemente (siehe Menubalken in **Abbildung 3**):

- Kaggle Discussion Boards: Diskussionsseiten für Fragen zur Competition.
- Kaggle Notebooks: Eine integrierte Programmierumgebung mit vielen vorinstallierten Data Science Bibliotheken, in welcher direkt auf die Wettbewerbsdaten zugegriffen werden kann, und von wo aus auch Wettbewerbsbeiträge direkt abgeschickt werden können. Besonders hilfreich ist die Möglichkeit, ein (öffentliches) Notebook einer anderen Person zu kopieren und anzupassen.

Anhang B: Die Formulierung des Falls in der Open Data St.Gallen Kaggle Competition

Die Stadt St.Gallen Open Data Kaggle Competition findet sich auf einer Kaggle-Webseite und beinhaltet Angaben zur Ausgangslage, zu den zur Verfügung gestellten Daten, der Vorgehensweise für die Wettbewerbsteilnahme etc. Im Folgenden wird hier der das Fallbeispiel schildernde Text der Wettbewerbsseite wiedergegeben (vgl. Competition C1, Anhang C):

Ausgangslage

Die Stadt St.Gallen betreibt seit September 2019 ein Open Data Portal auf dem unpersönliche, unkritische Daten der Stadtverwaltung veröffentlicht werden. Mit der Veröffentlichung von sogenannten Open Government Data (OGD) als maschinenlesbare Datensätze will die öffentliche Verwaltung einen Beitrag zur Förderung von Transparenz, Partizipation, und Innovation leisten. Neben Statistikdaten, Geodaten, Verkehrszähldaten oder Sensordaten gehören dazu auch Daten aus dem Politik- und Verwaltungs-Bereich, wie beispielsweise öffentliche Vergaben oder die städtischen Budgets.

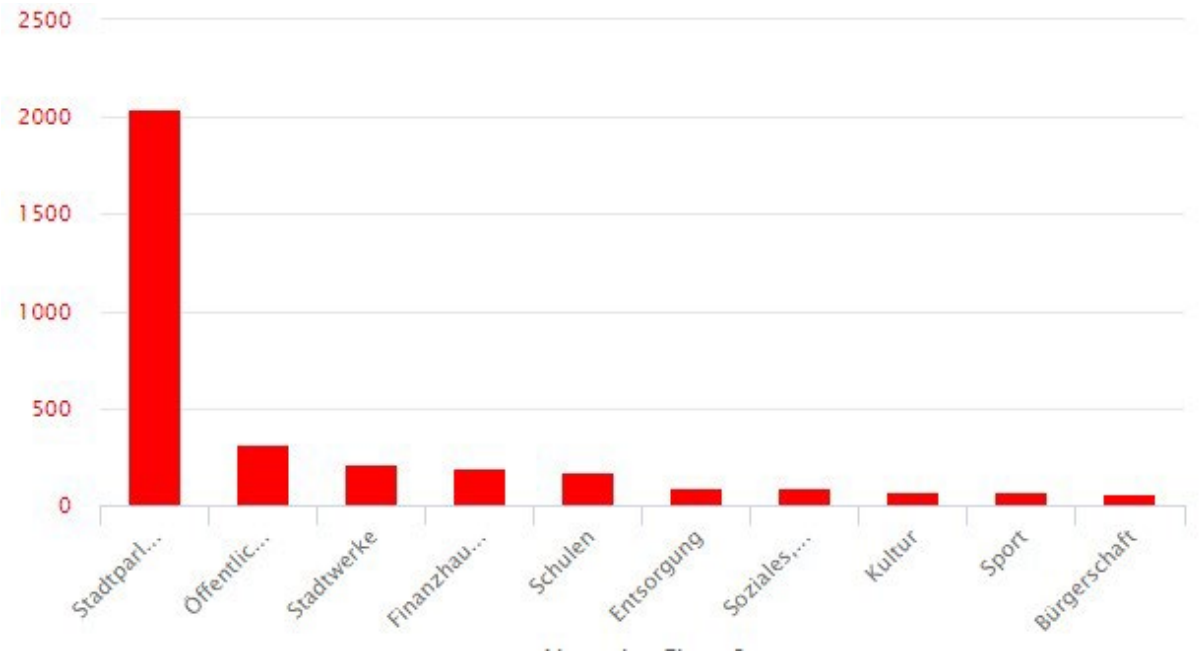
Nun plant die Stadt St.Gallen die Veröffentlichung eines umfangreichen Datensatzes des städtischen Ratsinformationssystems. Darin beinhaltet sind die traktandierten Geschäfte des Stadtparlaments St.Gallen seit dem Jahr 2001. Den behandelten Geschäften sind verschiedenen Eigenschaften wie das Sitzungsdatum, die Art des Geschäfts, oder auch die Aktenplannummer (sprich das Thema) zugeordnet. Damit lassen sich die rund 4000 Geschäfte nicht nur thematisch filtern, sondern es lassen sich auch politikwissenschaftliche Analysen bspw. über die Themenschwerpunkte im Verlauf der Jahre durchführen.

Problemstellung

Die Geschäfte des Stadtparlaments können grossmehrheitlich in zwei Arten unterteilt werden:

- Die Sachgeschäfte, welche von Seiten der Verwaltung angestossen, bearbeitet und traktandiert werden (1750 Geschäfte).
- Die Parlamentarischen Vorstösse, bestehend aus Interpellationen, Motionen und Einfachen Anfragen, welche von Seiten der Stadtparlamentarierinnen und Stadtparlamentariern eingebracht werden (1840 Geschäfte).

Während die Sachgeschäfte einer thematischen Aktenplannummer wie Schule, Kultur, Entsorgung, etc. zugeordnet werden, erhalten die parlamentarischen Vorstösse eine Aktenplannummer nach der entsprechenden Geschäftsart. Dies führt zu folgender thematischer Häufung der Geschäfte.



Erläuterung: Die erste Säule "Stadtparlament" wird durch die grosse Anzahl an parlamentarischen Vorstössen bestimmt, während die Sachgeschäfte in die restlichen Kategorien eingeteilt werden.

Diese Verzerrung der Themen führt zu einem geringeren Mehrwert des Datensatzes, weil

1. die Filterung nach Aktenplannummer (Thema) unvollständige Resultate ergibt,
2. politische Themenanalysen nur die Sachgeschäfte und somit die Verwaltungsseite abdecken, und
3. damit ein politikwissenschaftlicher Vergleich der Geschäftsthemen zwischen Verwaltungsseite und Parlamentsseite verunmöglicht wird.

Ziel der Competition

Während eine manuelle Neuzuordnung der parlamentarischen Vorstösse möglich wäre, ist sie ressourcentechnisch unrealistisch. Aus diesem Grund soll dieser Mangel des Datensatzes im Rahmen dieser Competition mittels Data Mining adressiert werden. Ziel ist es, die parlamentarischen Vorstösse den bestehenden thematischen Kategorien zuzuordnen und die Genauigkeit der maschinellen Zuordnung auszuweisen. Fehlerhafte Zuordnungen sollen möglichst vermieden werden, um die Qualität der Filterfunktion des Datensatzes zu gewährleisten. Mittels der ausgewiesenen Genauigkeit könnten ungenaue Zuordnungen durch Verwaltungsmitarbeitende überprüft oder manuell zugeordnet werden. Damit kann die Machine Learning Competition den Nutzen des Datensatzes massiv steigern, während der manuelle Aufwand innerhalb der Verwaltung minimal gehalten wird.

Datenstruktur

Die Themen, sprich die Aktenplannummern enthalten verschiedene Ebenen, welchen die Sachgeschäfte und parlamentarischen Vorstösse zugeordnet wurden. Die hierarchische Struktur des Aktenplans sieht wie folgt aus:

- Ebene 1: 9 Kategorien
- Ebene 2: 35 Kategorien
- Ebene 3: 175 Kategorien
- Ebene 4: 1094 Kategorien
- etc.

Aus diesem Grund ist eine Zuordnung auf Basis von Ebene 2 realistisch. Der Aktenplan auf Ebene 2 beinhaltet 35 Kategorien, wovon 29 im vorliegenden Datensatz verwendet wurden. 12 Kategorien enthalten weniger als 20 Geschäfte, weshalb sie in der Kategorie "sonstige Themen" zusammengefasst wurden. Neben der Kategorie "Stadtparlament", in der die parlamentarischen Vorstösse enthalten sind, verbleiben somit die 17 folgenden Kategorien, welchen die Geschäfte zugeordnet werden sollen:

- Bürgerschaft
- Entsorgung
- Finanzhaushalt
- Kultur
- Organisation der Verwaltung
- Schulen
- Soziales, Sozialhilfe
- Sport
- Stadtrat
- Stadtwerke
- Städtisches Personal
- Umweltschutz
- Verkehr, Telekommunikation
- Verkehrsbetriebe
- Öffentliche Ordnung und Sicherheit
- Öffentliches Bauen
- *Sonstige Themen*

Aufgabe der Competition

Trainieren Sie auf Basis der bereits kategorisierten Geschäfte und deren Titel ein Machine Learning Modell und testen Sie es auf dessen Genauigkeit (siehe Data und Evaluation).

Danksagung

Wir danken der Stadt St.Gallen und insbesondere Nicola Wullschleger für die Bereitstellung der Daten.

Evaluation

Die Evaluation wird anhand der "mean average precision at k" (kurz MAP at k) mit Relevanzniveau k=3 vorgenommen. Geben Sie drei Empfehlungen ab (mit einem Leerschlag getrennt). Pro Dokument ist jeweils nur ein Thema relevant. Wenn das relevante Thema an erster Stelle liegt, ergibt dies 1

Punkt. Wenn es an zweiter Stelle liegt, 1/2 Punkt, an dritter Stelle 1/3 Punkt.
Anschliessend wird über alle Testdokumente gemittelt.

Datenbeschreibung

Die für das Training zur Verfügung stehenden Daten bestehen aus den Titeln der Sachgeschäften (X) sowie der zugehörigen Kategorie (Aktenplannummer, als y bezeichnet).

Dateien

- **train.csv** -die Trainingsdaten mit Spalten X1 (dem Titel) und y (dem zu lernenden Label)
- **test.csv** - die Testdaten- hier ist nur X1 vorhanden, y muss vorhergesagt werden
- **sample_submission.csv** - eine Datei, welche das Format für die Submission angibt. Für jedes Testbeispiel müssen 3 Vorschläge für die korrekte Kategorie (nur eine ist korrekt!) gemacht werden.
- **Kategorienbeschreibung.xlsx** - hier werden die Kategorien beschrieben. Werte von y können so Themengebieten zugeordnet werden.

Anhang C: Kaggle-Competitions für den Unterricht

Wir listen im Folgenden einige Kaggle-Competitions auf, welche im Data Science Fachhochschulunterricht von einem der Autoren eingesetzt wurden.

- C1: Open Data St.Gallen Kaggle Competition– Tagempfehlungen für das St.Galler Ratsinformationssystem:
  https://www.kaggle.com/t/dbc7a9341366404eb665c64a4b57ea61
  Dies ist die Kaggle Competition zur im Haupttext diskutierten Fallstudie.
  N1: Einführung in die Textklassifikation:
  https://www.kaggle.com/code/toedtli/textklassifikation-einf-hrung
- C2: Daan-Kreuz-Kreis-Plus– Bildklassifikation auf von Studierenden erstellten Kreuz-, Kreis- oder Plus-förmigen Bildern:
  https://www.kaggle.com/t/af5c2cfab877461da72490e1ef1cf59c
  N2.1: Eine Beispielsubmission unter Verwendung von Hauptkomponentenanalyse:
  https://www.kaggle.com/code/toedtli/kreuzkreispluskagglesubmission
  N2.2: Eine Demonstration von Bildvorverarbeitungstechniken
  https://www.kaggle.com/code/toedtli/preprocessingimagekkpdata
- C3: «Kreuz, Kreis oder Plus?»– Klassifikation auf binären Vektoren, welche Kreuz- Kreis- oder Plus-Bilder codieren:
  https://www.kaggle.com/t/2c436344e26e4609be77243ff0b1d086

  N3: Einsteigernotebook: https://www.kaggle.com/code/toedtli/techie-starthilfe
- C4: Digit Recognizer Privat– Klassifikation auf dem Chinese MNIST-Digits Datensatz:
  https://www.kaggle.com/t/34c0ffb88778434aa2c3fd916d74d9f9
  N4: https://www.kaggle.com/code/toedtlifhsg/eigenfaces-chinese-mnist
  Bilderkennung von geschriebenen Ziffern, mit Deep Learning Beispielcode
- C5: Schnell-Mal-Klassifizieren– Feature Engineering auf einem einfachen 2D-Datensatz:
  https://www.kaggle.com/t/91e496b7508d473d9ab6b6ff194ef5a7
  Besonders einfacher 2D-Datensatz für den Einstieg in Python/Numpy/Pandas sowie in Kaggle Competitions
  N5: Ein Einführungsnotebook zur «Schnell-Mal-Klassifizieren» findet sich hier:
  https://www.kaggle.com/code/toedtli/vorlage-competition
- C6: UCI Kreditkartendatensatz – Schätzung des Insolvenzrisikos von Kreditkartennutzern:
  https://www.kaggle.com/competitions/kreditkartendatensatz/
- C7: OED19Classification–einfacher 2D-Datensatz mit einführenden Folien:
  https://www.kaggle.com/competitions/oed19classification/
  N7.1: Demo-Notebook OED2019:
  https://www.kaggle.com/code/toedtli/demo-notebook-oed2019
  N7.2: einführende Folien:
  https://www.kaggle.com/competitions/oed19classification/data?select=Open_Education_Day_2019_Folien_Beat_Toedtli.pdf
- C8: Daan-nicht-MNIST– Klassifizieren von 16 Bildklassen:
  https://www.kaggle.com/t/f301522dae3248aa928b5e04f0022003

Beispielvisualisierungen finden sich hier:
N8: https://www.kaggle.com/code/toedtlifhsg/daten-berblick

- C9 : Twitter-Airline-Sentiment Analysis– Textklassifikation von Airline-Review-Artikeln : https://www.kaggle.com/t/2968658cd5ce4c0db575b115251fede7
  N9 : Beispiellösung : https://www.kaggle.com/code/toedtlifhsg/vorlage-sentiment-analysis
- C10: Titanic privat– die Titanic-Competition mit zusätzlichen fehlenden Werten: https://www.kaggle.com/competitions/titanic-privat/
  N10.1: Vorlage für Übung 1: https://www.kaggle.com/code/toedtlifhsg/titanic-vorlage-f-r-bung-1
  N10.2: Vorlage für Übung 2: https://www.kaggle.com/code/toedtli/titanic-vorlage-f-r-bung-2

Literaturverzeichnis

Bloom, B.S. (Ed.), Engelhart, M.D., Furst, E.J., Hill, W.H., & Krathwohl, D.R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. New York: David McKay.

Jensen, K. (2012). English: A diagram showing the relationship between the different phases of CRISP-DM and illustrates the recursive nature of a data mining project. Own work based on: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.0/en/ModelerCRISPDM.pdf (Figure 1). https://commons.wikimedia.org/wiki/File:CRISP-DM_Process_Diagram.png

Kaggle. (2021). Titanic privat Overview. https://www.kaggle.com/competitions/titanic-privat/overview

Kaggle. (2021). Titanic privat Leaderboard. https://www.kaggle.com/competitions/titanic-privat/leaderboard

Zhu, Mu. (2004). Recall, Precision and Average Precision. https://web.archive.org/web/20110504130953/http://sas.uwaterloo.ca/stats_navigation/techreports/04WorkingPapers/2004-09.pdf