**Masterarbeitstitel**

Vergleich der Vorhersagegüte von neuronalen Netzen und baumbasierten Algorithmen bei Multiklassifikation basierend auf Worteinbettungen der Verwendungszwecke von finanziellen Transaktiondaten.

Comparison of prediction accuracy between neural nets and tree-based algorithms based on wordembeddings of purpose titles of financial transaction data

**Preprocessing:**

* **Bag of words**
* **Andere**
* **dimensionsreduktion**

**Datensatz beschreibung:**

* **In Raw**
* **Textmining preprocessing**

**Exploration:**

* **Mengengerüst der klassen**
* **Häufigste Wörter pro Klasse**

**Features:**

* **Number of Camel words**
* **Number of camel words**
* **Number of total words**
* **Bag of words**

**Train/Test Split**

* **Kategorien gebalanced im Train Set**
* **Nachschauen über binomialverteilung wieviele Klasse im Testset nötig sind**

**Methoden**

* **Multinomial logit mit /ohne spline: als eta alles mögliche, zB auch predictor vom neuronalenNetz**
* **Bamlss paket: erweiterungen**
* **Classification RandomForest**
* **XGBoost**

**Performance maß:**

* **rank probability score,**
* **difference between multinomial likelihood and true classes**
* **brier-score**

**Next Steps:**

* **Buch zu categorial data holen und multinomial logit lesen**
* **Bamss Paper lesen**
* **Datensatz einlesen und Framework schreiben**
* **Marcus/Conni bescheid sagen mitte nächster Woche.**
* **David Middelbeck empfehlung:** Sandbox transaktionsdaten schauen
* **Performance Maße nachschauen**

tags:

textmining word embeddings multicategorial transaction data neural nets comparison xgboost

.

daten müssen enthalten:

transaktionsdaten:

* Verwendungszweck, auftraggeber, betrag
* empfänger: zb amazon marketplace
* Zielvariable: categorie zb. online einkauf güter (40 Klassen +)
* Zielvariable: unterkategorie: elektronik

Idee:

news datensatz für die Masterarbeit nehmen aber Anwendung im Datalab auf Transaktionsdaten.

<https://www.kaggle.com/rmisra/news-category-dataset>

Fragen:

* Wie sehr ist es dem Datalab wichtig, beschrieben zu werden (also als Rolle der Firma, da die Daten ja nochmal von Kunden kommen). Vertagung, eher egal für wdl
* doch lieber früher input von Profs holen, mit dem Überblick welche Daten grade im WDL vorlliegen und was skalierbar ist?

2tes Thema als Alternative für Prof mitgeben.

edyoucated

2tes Thema

NBO zeitbezogen

Praktikums Feedback

* globale Variablen im merge
* slides mehr in Projektleiterposition