WS 2019/20

Seminar Ökonometrie: Text Mining

Prof. Dr. Peter Winker
e–mail: Peter.Winker@wirtschaft.uni-giessen.de Tel.: 0641/99-22-640

David Lenz
e–mail: David.Lenz@wi.jlug.de Tel.: 0641/99-22-640

Das Seminar behandelt die Anwendung von computerlinguistischen Techniken für ökonomische Fragestellungen. Es werden unterschiedliche Methoden erarbeitet, die in diesem Anwendungsgebiet in der Praxis relevant und/oder Gegenstand aktueller Forschungsanstrengungen sind. Neben dem Verständnis der methodischen Grundlagen wird auch die empirische Umsetzung der Methoden kritisch diskutiert. Dazu sollen die Seminarteilnehmer weitere empirische Evidenz aus der Literatur und/oder eigenen empirischen Arbeiten beisteuern. Ziel des Seminars ist es, die TeilnehmerInnen zu einer selbstständigen Auseinandersetzung mit Methoden und Anwendungen aus dem Bereich des Natural Language Processing zu befähigen, die Lektüre und Auswertung wissenschaftlicher Beiträge einzuüben und die Darstellung und den Vortrag wissenschaftlicher Resultate zu trainieren. Fortgeschrittene Kenntnisse in Statistik sowie Programmierkenntnisse werden empfohlen.

Vorläufige Themenübersicht

15.10.2019	Einführungsveranstaltung: Organisatorisches / Themenvergabe
29.10.2019	Regulärer Termin um mögliche Fragen zu klären
19.11.2019	Thema 1: Text als Daten: Übersicht und Einführung
26.11.2019	Thema 2: Aufbereitung von Text Daten
03.12.2019	Thema 3: Text Repräsentationen
10.12.2019	Thema 4: Mehrsprachige Text Repräsentationen
17.12.2019	Thema 5: Wörterbuch basierte Sentiment Analyse
14.01.2020	Thema 6: Machine learning basierte Sentiment Analyse
21.01.2020	Thema 7: Topic Modeling
28.01.2020	Thema 8: Textklassifizierung
04.02.2020	Thema 9: Text-basierte Indikatoren

Basisliteratur

- Srivastava, A. M. Sahami (2009). Text Mining: Classification, Clustering, and Applications. Chapman & Hall/CRC, Boca Raton (verfügbar am Lehrstuhl & Semesterapparat)
- Zusätzliche Literatur wird zu den einzelnen Themen des Seminars angegeben.

Voraussetzungen

- regelmäßige aktive Teilnahme an den Seminarsitzungen
- Anfertigung einer mit mindestens "ausreichend" (4,0 bzw. 5 Punkten) bewerteten Seminararbeit (Gewichtung 2/3)
- Vortrag der Arbeit im Rahmen einer Seminarsitzung sowie kurzes Koreferat zu einem anderen Seminarthema (Gewichtung 1/3)
- **Gute Grundlagen in Statistik, Mathematik und Ökonometrie!**

Weitere Informationen

ECTS Punkte:	6
Sem.-Wochenstunden:	2
Umfang der Arbeit:	ca. 10 Seiten (Text inkl Abbildungen und Tabellen)

Thema 1: Text als Daten: Übersicht und Einführung

Wo liegen die Anfänge der computergestützten Analyse von Text Daten? Welches sind die relevanten Probleme? Wie haben sich die Fragestellungen im Laufe der Zeit verändert, wie die Methoden? Welche Fragestellungen werden zukünftig relevant sein? Diese und andere Fragen sollen in dieser Seminararbeit in Form einer Literaturübersicht beantwortet werden. Außerdem wird eine Übersicht über aktuelle Forschungsanstrengungen erwartet, welche Text als Datenquelle zur Beantwortung ökonmischer Fragestellungen benutzen.

Einstiegsliteratur:

- Heyer, G., U. Quasthoff T. Wittig (2006). Text mining: Wissensrohstoff text. *W3l, Herdecke* 18
- Cleary, P., K. Garlock, D. Novak, E. Pullman S. Mann (2017). Text Mining 101: What You Should Know. *The Serials Librarian* 72(1-4), 156–159

Thema 2: Aufbereitung von Text Daten

Das Sammeln von Daten ist der erste Schritt bei der datengetriebenen Analyse einer gegebenen Problemstellung. Organisationen stellen häufig öffentliche Programmierschnittstellen für den Zugriff auf ihre Daten bereit, bspw. die Twitter API oder die NY Times API. Auch gibt es eine Fülle von frei verfügbaren Datensätzen, bspw. auf Github, Kaggle oder Reddit. Diese Daten sind in ihrer rohen Form häufig nicht direkt für ökonomische Analysen nutzbar. Gegenstand der Seminararbeit sollte die Darstellung aktueller Standards in der Datenaufbereitung sein, bspw. Stemming, Lemmatization, Stopwords filtern, Popularitäts-basiertes Vorfiltern von Wörtern etc.

Einstiegsliteratur:

- Anandarajan, M., C. Hill T. Nolan (2019). Text Preprocessing. Springer International Publishing, Cham, 45–59
- Vijayarani, S., M. Ilamathi M. Nithya (2015). Preprocessing Techniques for Text Mining-An Overview Dr

Thema 3: Text Repräsentationen

Text-Daten sind typischerweise hoch-dimensional und unstrukturiert. Die Art der Aufbereitung und Darstellung von Text-Daten für den Gebrauch mit Computern ist von entscheidender Bedeutung für den Erfolg der anschließenden Analysen. Traditionelle Darstellungen, bspw. die Repräsentation einzelner Wörter als Indizes in einem Wörterbuch, sind schnell und einfach zu reproduzieren. Aktuelle Forschungsanstrengungen zeigen, dass komplexere Methoden, bspw. Word2Vec oder fastText, die Ergebnisse jedoch deutlich verbessern können. Gegenstand der Seminararbeit sollte ein Vergleich verschiedener Methoden, bspw. BoW / Word2Vec / GloVe / fastText sein, oder die Anwendung selbiger, oder die detaillierte Darstellung einer der neueren Methoden.

Einstiegsliteratur:

- Mikolov, T., K. Chen, G. Corrado J. Dean (2013a). Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781
- Mikolov, T., I. Sutskever, K. Chen, G. Corrado J. Dean (2013b). Distributed Representations of Words and Phrases and their Compositionality. *CoRR* abs/1310.4546
- Joulin, A., E. Grave, P. Bojanowski, M. Douze, H. Jégou T. Mikolov (2016). FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*

Thema 4: Mehrsprachige Text Repräsentationen

Die gemeinsame Darstellung von Wörtern verschiedener Sprachen in einem gemeinsamen Vektorraum bietet faszinierende Möglichkeiten. Zum Beispiel kann bei der maschinellen Übersetzung das Problem der Übersetzung eines Wortes, das nie in parallelen Daten gesehen wurde, überwunden werden, indem seine Vektorraumnachbarn gesucht werden. Inhalt der Seminararbeit sollte die Darstellung aktueller Forschungsanstrengungen sein, oder die Implementation eines mehrsprachigen Word Embedding Systems.

Einstiegsliteratur:

- Ammar, W., G. Mulcaire, Y. Tsvetkov, G. Lample, C. Dyer N. A. Smith (2016). Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*

Thema 5: Wörterbuch basierte Sentiment Analyse

Im Allgemeinen zielt die Stimmungsanalyse darauf ab, die Einstellung eines Sprechers, Schreibers oder eines anderen Subjekts in Bezug auf ein Thema oder die gesamte kontextuelle Polarität oder emotionale Reaktion auf ein Dokument, eine Interaktion oder ein Ereignis zu bestimmen. Die Sentimentanalyse von Freitextdokumenten ist eine gängige Aufgabe im Bereich Text Mining. In der klassischen Sentiment-Analyse werden Texten vordefinierte Sentiment-Labels wie ”positiv” oder ”negativ” zugeordnet. Texte (hier als Dokumente bezeichnet) können Rezensionen über Produkte oder Filme, Artikel usw. sein. Traditionelle Methoden benutzen Wörterbücher mit positiven und negativen Wörtern, um von der Häufigkeit der jeweiligen Wortkategorie auf den Sentiment eines Dokuments zu schließen. Inhalt der Seminararbeit sollte die Darstellung Wörterbuch basierter Methoden sein, oder die Anwendung einer solchen Methode.

Einstiegsliteratur:

- Mäntylä, M. V., D. Graziotin M. Kuutila (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review* 27, 16 – 32
- Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies* 5(1), 1–167
- Cambria, E., D. Das, S. Bandyopadhyay A. Feraco (2017). A practical guide to sentiment analysis, 5. Springer
- Bannier, C. E., T. Pauls A. Walter (2018). Content Analysis of Business Communication: Introducing a German Dictionary. *Journal of Business Economics*

Thema 6: Machine-learning basierte Sentiment Analyse

Im Allgemeinen zielt die Stimmungsanalyse darauf ab, die Einstellung eines Sprechers, Schreibers oder eines anderen Subjekts in Bezug auf ein Thema oder die gesamte kontextuelle Polarität oder emotionale Reaktion auf ein Dokument, eine Interaktion oder ein Ereignis zu bestimmen. Die Sentimentanalyse von Freitextdokumenten ist eine gängige Aufgabe im Bereich Text Mining. In der klassischen Sentiment-Analyse werden Texten vordefinierte Sentiment-Labels wie ”positiv” oder ”negativ” zugeordnet. Texte (hier als Dokumente bezeichnet) können Rezensionen über Produkte oder Filme, Artikel usw. sein. Aktuelle Methoden benutzen Machine Learning, um auf den Sentiment eines Doukments zu schließen. Inhalt der Seminararbeit könnte die Darstellung Machine Learning-basierter Methoden sein, oder die Anwendung einer machine learning Methode zur Sentiment Bestimmung.

Einstiegsliteratur:

- Mäntylä, M. V., D. Graziotin M. Kuutila (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review* 27, 16 – 32
- Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies* 5(1), 1–167
- Cambria, E., D. Das, S. Bandyopadhyay A. Feraco (2017). A practical guide to sentiment analysis, 5. Springer

Thema 7: Topic Modeling

Beim maschinellen Lernen und bei der Verarbeitung natürlicher Sprache ist ein Topic Modell eine Art statistisches Modell zum Auffinden der abstrakten "Themen", die in einer Sammlung von Dokumenten vorkommen. Topic Modelling ist ein häufig verwendetes Text-Mining-Werkzeug zum Auffinden verborgener semantischer Strukturen in einem Textkörper. Inhalt der Seminararbeit sollte die detaillierte Darstellung einer der Methoden, bspw. LDA, sein, oder ein Vergleich verschiedener Topic Modelle, oder die Anwendung einer Methode zur Identifikation latenter Topics in einem Corpus.

Einstiegsliteratur:

- Srivastava, A. M. Sahami (2009). Text Mining: Classification, Clustering, and Applications. Chapman & Hall/CRC, Boca Raton (verfügbar am Lehrstuhl) Kap. 4
- Blei, D. M., A. Y. Ng M. I. Jordan (2003). Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3, 993–1022
- Hisano, R., D. Sornette, T. Mizuno, T. Ohnishi T. Watanabe (2013). High Quality Topic Extraction from Business News Explains Abnormal Financial Market Volatility. *PLOS ONE* 8(6), 1–12
- Mizuno, T., T. Ohnishi T. Watanabe (2017). Novel and topical business news and their impact on stock market activity. *EPJ Data Science* 6(1), 26

Thema 8: Text Klassifizierung

Die Klassifizierung von Textdokumenten nach Kategorien oder Themen ist ein wichtiger Bestandteil vieler Textverarbeitungssysteme. Die Aufgabe der Textklassifizierung besteht darin, Dokumente in vordefinierte Themen wie Wirtschaft, Politik und Sport einzuteilen. Zum Beispiel ist die automatische Zuweisung mehrerer klinischer Codes (fachspezifische medizinische Ausdrücke) zu klinischem Freitext ein typisches Problem der Textklassifikation mit mehreren Themen. Spam-Erkennung ist ein weiteres Beispiel für Analysen dieser Kategorie. Inhalt der Seminararbeit könnte die Darstellung verschiedener Anwendungen und Methoden der Textklassifizierung sein, oder eine eigene Untersuchung, bei der Texte in Kategorien unterteilt werden.

Einstiegsliteratur:

- Srivastava, A. M. Sahami (2009). Text Mining: Classification, Clustering, and Applications. Chapman & Hall/CRC, Boca Raton, (verfügbar am Lehrstuhl) Kap. 3, 7
- Luss, R. A. d'Aspremont (2015). Predicting abnormal returns from news using text classification. *Quantitative Finance* 15(6), 999–1012
- Abrahams, A. S., W. Fan, G. A. Wang, Z. Zhang J. Jiao (2015). An integrated text analytic framework for product defect discovery. *Production and Operations Management* 24(6), 975–990

Thema 9: Konstruktion Text-basierter Indikatoren

Die Messung (& Prognose) mikro- und makro-ökonomischer Größen wie BIP, Zins oder Wertpapierkurse ist von großer Relevanz, wichtige Merkmale sind eine schnelle Verfügbarkeit und hohe Treffsicherheit. Traditionelle Indikatoren haben verschiedene Probleme, bspw. langsame Verfügbarkeit oder große Prognosefehler bei sich verändernden Marktbedingungen. Text Daten, bspw. Zeitungsartikel, beinhalten eine große Menge relevanter Informationen, die Marktakteure in Echtzeit zur Steuerung ihrer wirtschaftlichen Aktivitäten nutzen. Die Konstruktion Text-basierter Indikatoren ist daher Gegenstand aktueller Forschungen. Inhalt der Seminararbeit könnte die Darstellung / Gegenüberstellung verschiedener Text-basierter Indikatoren sein, oder die Konstruktion eines Text-basierten Indikators.

Einstiegsliteratur:

- Levenberg, A., S. Pulman, K. Moilanen, E. Simpson, S. Roberts (2014). Predicting Economic Indicators from Web Text Using Sentiment Composition. *International Journal of Computer and Communication Engineering* 3(2), 109–115
- Lüdering, J. P. Winker (2016). Forward or backward looking? The economic discourse and the observed reality. *Jahrbücher für Nationalökonomie und Statistik* 236(4), 483–515
- Rönnqvist, S. P. Sarlin (2017). Bank distress in the news: Describing events through deep learning. *Neurocomputing* 264, 57 – 70. Machine learning in finance
- Thorsrud, L. A. (2018). Words are the New Numbers: A Newsy Coincident Index of the Business Cycle. *Journal of Business & Economic Statistics* (online), 1–35
- Tillmann, P. A. Walter (2018). ECB vs Bundesbank: Diverging Tones and Policy Effectiveness. *MAGKS Discussion Paper* 20

References

[Abrahams et al.(2015)Abrahams, Fan, Wang, Zhang and Jiao] Abrahams, A. S., W. Fan, G. A. Wang, Z. Zhang and J. Jiao (2015). 'An integrated text analytic framework for product defect discovery'. *Production and Operations Management* 24(6), 975–990.

[Ammar et al.(2016)Ammar, Mulcaire, Tsvetkov, Lample, Dyer and Smith] Ammar, W., G. Mulcaire, Y. Tsvetkov, G. Lample, C. Dyer and N. A. Smith (2016). 'Massively multilingual word embeddings'. *arXiv preprint arXiv:1602.01925* .

[Anandarajan et al.(2019)Anandarajan, Hill and Nolan] Anandarajan, M., C. Hill and T. Nolan (2019). Text Preprocessing. Springer International Publishing, Cham, S. 45–59.

[Bannier et al.(2018)Bannier, Pauls and Walter] Bannier, C. E., T. Pauls and A. Walter (2018). 'Content Analysis of Business Communication: Introducing a German Dictionary'. *Journal of Business Economics* .

[Blei et al.(2003)Blei, Ng and Jordan] Blei, D. M., A. Y. Ng and M. I. Jordan (2003). 'Latent Dirichlet Allocation'. *J. Mach. Learn. Res.* 3, 993–1022.

[Cambria et al.(2017)Cambria, Das, Bandyopadhyay and Feraco] Cambria, E., D. Das, S. Bandyopadhyay and A. Feraco (2017). A practical guide to sentiment analysis, volume 5. Springer.

[Cleary et al.(2017)Cleary, Garlock, Novak, Pullman and Mann] Cleary, P., K. Garlock, D. Novak, E. Pullman and S. Mann (2017). 'Text Mining 101: What You Should Know'. *The Serials Librarian* 72(1-4), 156–159.

[Heyer et al.(2006)Heyer, Quasthoff and Wittig] Heyer, G., U. Quasthoff and T. Wittig (2006). 'Text mining: Wissensrohstoff text'. *W3l, Herdecke* 18.

[Hisano et al.(2013)Hisano, Sornette, Mizuno, Ohnishi and Watanabe] Hisano, R., D. Sornette, T. Mizuno, T. Ohnishi and T. Watanabe (2013). 'High Quality Topic Extraction from Business News Explains Abnormal Financial Market Volatility'. *PLOS ONE* 8(6), 1–12.

[Joulin et al.(2016)Joulin, Grave, Bojanowski, Douze, Jégou and Mikolov] Joulin, A., E. Grave, P. Bojanowski, M. Douze, H. Jégou and T. Mikolov (2016). 'FastText.zip: Compressing text classification models'. *arXiv preprint arXiv:1612.03651* .

[Levenberg et al.(2014)Levenberg, Pulman, Moilanen, Simpson, and Roberts] Levenberg, A., S. Pulman, K. Moilanen, E. Simpson, and S. Roberts (2014). 'Predicting Economic Indicators from Web Text Using Sentiment Composition'. *International Journal of Computer and Communication Engineering* 3(2), 109–115.

[Liu(2012)] Liu, B. (2012). 'Sentiment Analysis and Opinion Mining'. *Synthesis Lectures on Human Language Technologies* 5(1), 1–167.

[Lüdering and Winker(2016)] Lüdering, J. and P. Winker (2016). 'Forward or backward looking? The economic discourse and the observed reality'. *Jahrbücher für Nationalökonomie und Statistik* 236(4), 483–515.

[Luss and d'Aspremont(2015)] Luss, R. and A. d'Aspremont (2015). 'Predicting abnormal returns from news using text classification'. *Quantitative Finance* 15(6), 999–1012.

[Mikolov et al.(2013a)Mikolov, Chen, Corrado and Dean] Mikolov, T., K. Chen, G. Corrado and J. Dean (2013a). 'Efficient Estimation of Word Representations in Vector Space'. *CoRR* abs/1301.3781.

[Mikolov et al.(2013b)Mikolov, Sutskever, Chen, Corrado and Dean] Mikolov, T., I. Sutskever, K. Chen, G. Corrado and J. Dean (2013b). 'Distributed Representations of Words and Phrases and their Compositionality'. *CoRR* abs/1310.4546.

[Mizuno et al.(2017)Mizuno, Ohnishi and Watanabe] Mizuno, T., T. Ohnishi and T. Watanabe (2017). 'Novel and topical business news and their impact on stock market activity'. *EPJ Data Science* 6(1), 26.

[Mäntylä et al.(2018)Mäntylä, Graziotin and Kuutila] Mäntylä, M. V., D. Graziotin and M. Kuutila (2018). 'The evolution of sentiment analysis—A review of research topics, venues, and top cited papers'. *Computer Science Review* 27, 16 – 32.

[Rönnqvist and Sarlin(2017)] Rönnqvist, S. and P. Sarlin (2017). 'Bank distress in the news: Describing events through deep learning'. *Neurocomputing* 264, 57 – 70. Machine learning in finance.

[Srivastava and Sahami(2009)] Srivastava, A. and M. Sahami (2009). Text Mining: Classification, Clustering, and Applications. Chapman & Hall/CRC, Boca Raton.

[Thorsrud(2018)] Thorsrud, L. A. (2018). 'Words are the New Numbers: A Newsy Coincident Index of the Business Cycle'. *Journal of Business & Economic Statistics* (online), 1–35.

[Tillmann and Walter(2018)] Tillmann, P. and A. Walter (2018). 'ECB vs Bundesbank: Diverging Tones and Policy Effectiveness'. *MAGKS Discussion Paper* 20.

[Vijayarani et al.(2015)Vijayarani, Ilamathi and Nithya] Vijayarani, S., M. Ilamathi and M. Nithya (2015). 'Preprocessing Techniques for Text Mining-An Overview Dr'.