Hessisches Kompetenzzentrum
für Hochleistungsrechnen

Combining Text Mining and Multivariate Time Series Modelling

Project Manager
Viktoriia Naboka

Principal Investigator
Prof. Dr. Peter Winker

Project Term
2021 - 2022

Clusters
justHPC Gießen

Institute
Chair of Statistics and Econometrics

University
Justus Liebig University Giessen

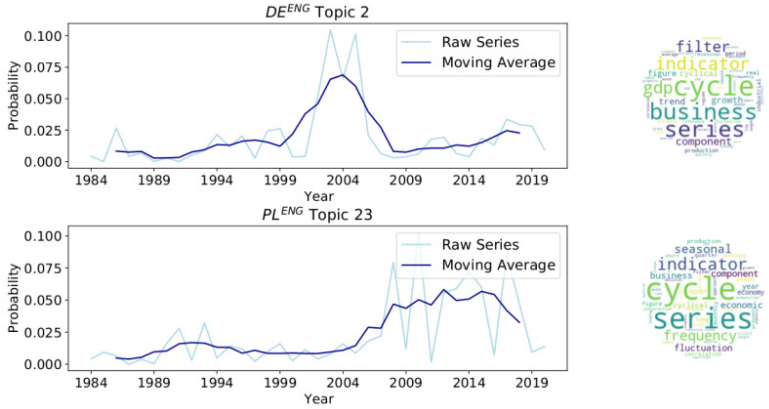


Figure 1: Topic Match “Business cycle”. DE^{ENG} represents the German set of scientific publications in the English language. PL^{ENG} represents scientific publications in the English language published in Poland.

Introduction

Collections of texts are considered as a valuable source of information for applied economic analysis. Recent developments in the access to large sets of documents such as scientific abstracts, articles, news items, social media messages or statements of different institutions increase the interest in this type of data. In this project, we use text corpora as the main source of data and build on recently developed and widely used natural language processing (NLP) techniques, especially topic modelling. The contribution of this project is threefold. First, we propose methods for comparing (matching) identified topics from various corpora using scientific economic publications from two countries, namely Germany and Poland. Second, we propose to use a new measure for selecting an optimal number of topics, namely singular Bayesian information criterion (sBIC). We evaluate the performance of sBIC and other model selection criteria commonly used in the literature in a comprehensive Monte Carlo (MC) simulation, for which the data are generated by a well defined data generation process (DGP) with the known number of topics. Third, we propose methods for integrating trends in topics and real indicators in multivariate time series analysis.

Methods

Probabilistic topic models allow to automatically analyze and uncover the underlying structure of large text collections without human intervention. In the current project, we focus on one of the mostly used techniques, namely Latent Dirichlet Allocation (LDA). We estimate LDA models using the Gibbs sampler as implemented in the Python package lda. To generate random sequences used in the text generation stage, we use the random

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN
TECHNISCHE UNIVERSITÄT DARMSTADT
UNIKASSEL VERSITÄT
Philipps Universität Marburg

number generator from Pythons' numpy package.

Results

In our first paper, it has been shown that the proposed topic matching approach allows to compare the topic-word distributions of two different LDA models and to identify suitable topic pairs across text corpora [2]. This could be useful when, for example, comparing topic trends in different countries, as in the present case. It could be also useful when analyzing the emergence and evolution of topic trends over time for sub-samples of one corpus.
Our second paper is dedicated to the comparison of different selection criteria with regard to the optimal number of topics in LDA models [1]. The comparison is based on Monte Carlo simulations and carried out for several alternative settings, varying with respect to the number of topics, the number of documents and the size of documents in the corpora. Simulation results showed that the singular Bayesian information criterion performed relatively well for all data generating processes considered in the experiments. The performance of different model selection procedures was evaluated by not only examining the accuracy of estimating the actual number of topics but also by analyzing the structure and contents of the estimated topics.

Outlook

Based on the results of our first work [2], we want to examine more closely the links between the corresponding topic time series and real macroeconomic variables with a focus on potential differences across countries. Thereby, we aim to consider different methods for deriving trends in topics and include these text based indicators in time series models, e.g., the widely used vector autoregressive model.

Publications

[1] Victor Bystrov et al.: "Choosing the Number of Topics in LDA Models – A Monte Carlo Comparison of Selection Criteria", 2022
https://doi.org/10.48550/arXiv.2212.14074

[2] Victor Bystrov et al.: "Cross-Corpora Comparisons of Topics and Topic Trends", Journal of Economics and Statistics 242.4, pp. 433–469, 2022
https://doi.org/10.1515/jbnst-2022-0024

Last Update: 2023-02-02 14:44