Cultural Trends in Tyumen Region: National and Global Contexts

Abstract (continued)

The initial stage of the project resulted in the full estimation of the available dataset in terms of its completeness and quality, including possible biases and imbalances. The key task was to explore the textual attributes of the datapoints, which contained descriptions (announcements) of cultural events registered on seven web-resources.
Particularly, we developed the machinery for automatic semantic analysis to produce a reliable categorisation of the events. For the purposes of this project, a cultural event is understood broadly as any social (public) activity in the spheres of contemporary arts, education, sports, music, theatre, including social activities arranged by restaurants, computer clubs, and private enthusiasts (the so-called everyday culture, such as yoga classes, guided tours, etc).
This preliminary step, including failed attempts, was necessary to re-consider directions for the subsequent research.

Our dataset has some inherent specificities, which reflect the type of the web-resources scraped (government-curated, commercial or independent), their geographic coverage (a strong imbalance towards English-speaking Europe and North America), and size (84% of the raw datapoints come from user-generated content on Meetup and Behance social networks; TED local and e-flux combined account for less than 1% of the dataset).
The novelty of the approach proposed for cultural studies here consists in leveraging Natural Language Processing (NLP) methods to access thematic and conceptual information relevant for investigating cultural trends. We developed a preprocessing pipeline, which addresses the issues typical for web-scraped texts, including filtering out the datapoints with empty or faulty text attributes. Quite unexpectedly, preprocessing required by NLP methods claimed 45% of the original dataset.
The impact of text preprocessing on, and its necessity for, NLP tasks was evaluated in a text categorization setting, using machine learning methods.

Another task that was tackled at the initial stage of the project was testing various approaches to numeric text representation to identify the best method to produce semantic partition of the data using clustering. This is needed to produce a system of thematic categories that cuts across all data harvested from a range of diverse web-resources, some of which lack thematic sections.
To identify the best-performing setting experimentally, we tested two multilingual embedding frameworks: ElMO and BERT. A number of internal evaluation measures and correlation with gold labels were used to compare clustering results for the two types of numeric representation. The results for BERT were better in the external evaluation experiment, and for the metrics which consider the global quality of clustering solution.

On the philosophical level, it is important to establish a non-comparative approach. Any comparison of obviously different cities leads to the confirmation of incommensurability. Therefore, our idea is to show the cultural developmental and contact traits across the regions, possible similarities, and, in the second step, if possible, influences and authentic characteristics of the cultural production and cultural exchange. On the practical level, the idea of the project is to offer the public a set of resources which would advise the creators of cultural policy and inform the planning of cultural production in the Tyumen region.