Polylingual Topic Models (PLTM)

Introduction

On Day 3 we learned about LDA which is the most commonly used topic model. One of the greatest advantages of using topic models is their ability to represent documents as probability distributions θ into a low-dimensional latent space – the T - dimensional probability simplex. Representing documents as points in a shared latent space abstracts away from the specific words used in a document, thereby facilitating the analysis of relationships between documents written using different vocabularies. Polylingual Topic Models (PLTM) are a subset of extensions of LDA that offer the ability to model multilingual collections. Their underlying assumption is that documents that are translations of each other, while written in a different language, cover the same set of topics. Mapping multilingual documents into a common latent topic space provide means to analyze the relationship between documents written in a different language. Some of the typical applications of PLTM include: finding or detecting document translation pairs, creating translation lexicons, aligning passages and modeling comparable corpora and cross lingual clustering and classification. PLTM also offers tools for performing exploratory data analysis over other types of parallel data collections where parallel data objects share the same latent data structure.

Project Description

In this project we are going to explore PLTM and learn how PLTM could be used to find similarity across different types of representations of objects that share the same set of topics (e.g. document translation pairs, news stories in different languages). We are going to use an existing implementation of PLTM to train a topic model and infer latent data structure over a parallel collection or a comparable corpora of data objects. We'll use the representation of the collection in the shared topic space to perform exploratory data analysis and learn more about the relationships between the data objects in our collection.

Datasets Used

We encourage students to suggest datasets where PLTM may be useful to perform exploratory data analysis. We'll provide two types of data collections:

1. Translations of the proceedings of the European parliament sessions in the official languages of the European Union (Europarl collection). This collection is a well known parallel corpora, i.e. a collection where document translations are aligned.
2. News stories in different languages. This collection is an example of a comparable corpora, i.e. a collection of documents in different languages that share the same of topics but are not necessarily translations of each other.