OSINT and topic modelling

April 18, 2018

1 Introduction

In natural language processing, topic modelling is a method for documents classification and describing document content in large document collections. Probabilistic topic models are generally unsupervised generative models. Formally, a topic is a probability distribution over terms in a vocabulary. In other words, a topic is a set of statistically linked words. Topic models approach assumes that:

- each topic is a probability distribution over words
- the document is a probability distribution over topics
- words order in documents does not matter

Each document has contributions of multiple topics with some probabilities.

$$p(w|d) = p(w|d)p(t|d) = \sum_{t \in \theta} \phi_{wt}\theta_{td} = 1$$

1.1 Probabilistic latent semantic analysis (PLSA)

Probabilistic latent semantic analysis (PLSA)

1.2 Laten Dirichle allocation (LDA)

Laten Dirichle allocation (LDA)

1.3 Additive regularization topic models (ARTM)

Additive regularization topic models (ARTM)