Extended Topic Models with Numerical Features

Gökhan Çapan, Ali Caner Türkmen

March 23, 2016

Introduction: Topic Models

- *Unsupervised* learning, recover *latent* topics in documents
- Can be thought of as *clustering*.
- **Key Assumption:** Topics lead to distinct word distributions. Intuitive.

Latent Dirichlet Allocation

- Rich, probabilistic mixed membership model due to (Blei et al., 2003)
- Widely adopted and extended
- **Key Idea:**

Comparison of Topic Models

- (Blei et al., 2003)

Some Examples of LDA Extensions

- Gokhan lit survey

We will propose a simple extension to LDA for working with Count Features

- Jointly model both word distributions and some count features in probabilistically sound framework
- Can then easily extend to other distributions for numeric features
- Close to the state of the art

Data Set and Features

- **Data Set:** News articles sampled from Anadolu Agency website. 1337 documents (can be expanded), 3000 tokens after adjusting for document frequency.
- **Features:** Complexity features such as word count, sentence count, average sentence length, comma count. (TBD)
- **Novel Features:** Etymological counts. Count the number of words from their etymological origins. Number of Arabic, Farsi, French words, etc. Source: TR Wiktionary Database Dump.

Learning

- **Variational inference:** Derive approximate inference algorithms based on a decoupling of the original model OR Variational EM-like procedures to find parameter estimates.
- Gibbs sampling assuming appropriate priors (tentative, out of scope for this project)

Conclusion

We propose two key contributions

- Put the topic modeling problem in an extended LDA framework, with numerical features
- Use etymological counts for the Turkish language