Finding the optimal number of topics for emergency text classification in HO-LDA

Hannah Keiler

Mentors: William M. Pottenger and Christie Nelson

DIMACS REU 2012

Problem/Purpose

- **Problem** - Classifying text data by hand is time consuming and potentially subjective. It is especially problematic in a state of emergency, where looking through text and knowing what sort of help is needed is urgent.

- **General Purpose** – A potential discovery that topic modeling with Higher Order Latent Dirichlet Allocation (HO-LDA) outperforms LDA in providing a model used to classify text data that informs emergency responders.

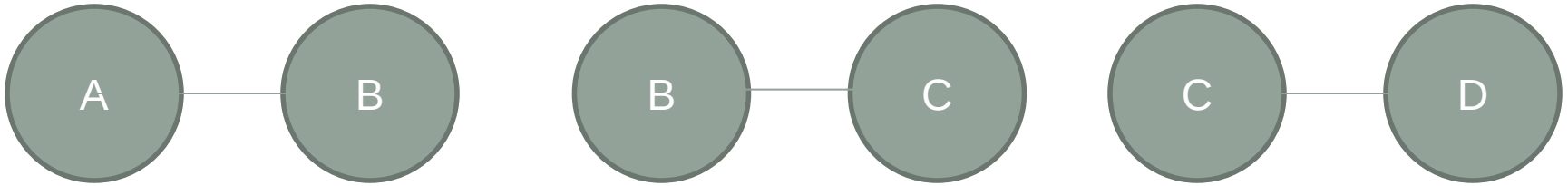
- **My Purpose** - Determining the optimal number of topics used in HO-LDA/LDA to provide a model – a novel area in topic modeling applications to emergency response.

Latent Dirichlet Allocation for Text

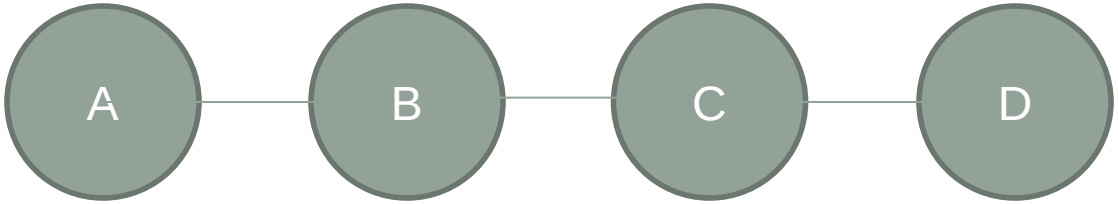
- LDA is a **topic model**, which is a statistical model for discovering the unobserved topics that explain why documents are similar
- Used on a set of documents each with a sequence of words
- From this, the distribution of words in a topic is computed, and the distribution of topics in a document is computed.

Higher Order Latent Dirichlet Allocation

First Order



Higher Order

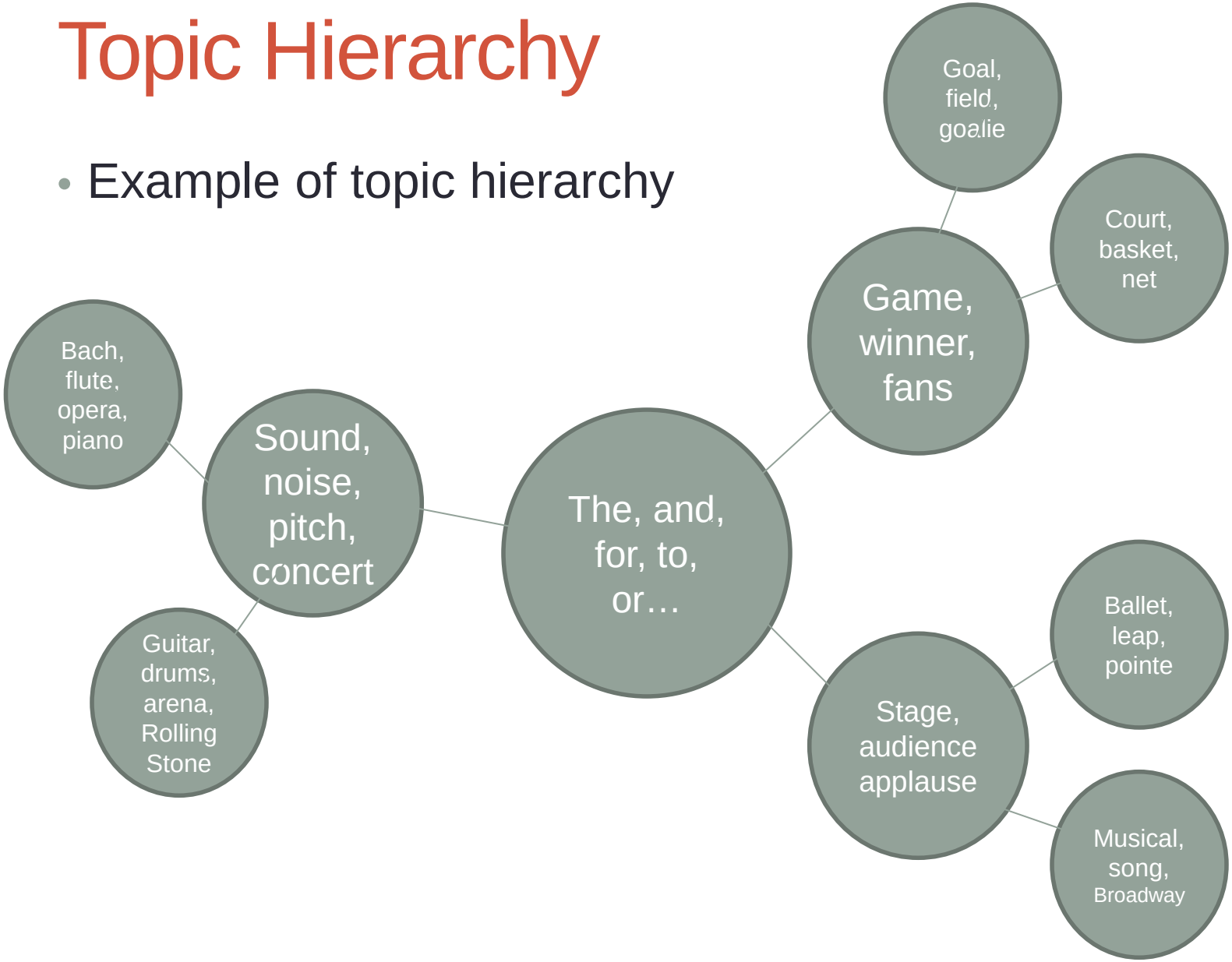


Number of Topics - Background

- The number of topics is one of the parameters in LDA and HO-LDA
- **Possible approaches**: looking at a “distance” between topics and maximizing it, looking at topic hierarchies.
  - Maximizing some function of the distance between clusters has been a method for determining number of clusters in cluster analysis
  - Using inferred hierarchies rather than predetermining number of topics has been proposed as an option in topic modeling

Topic Hierarchy

- Example of topic hierarchy

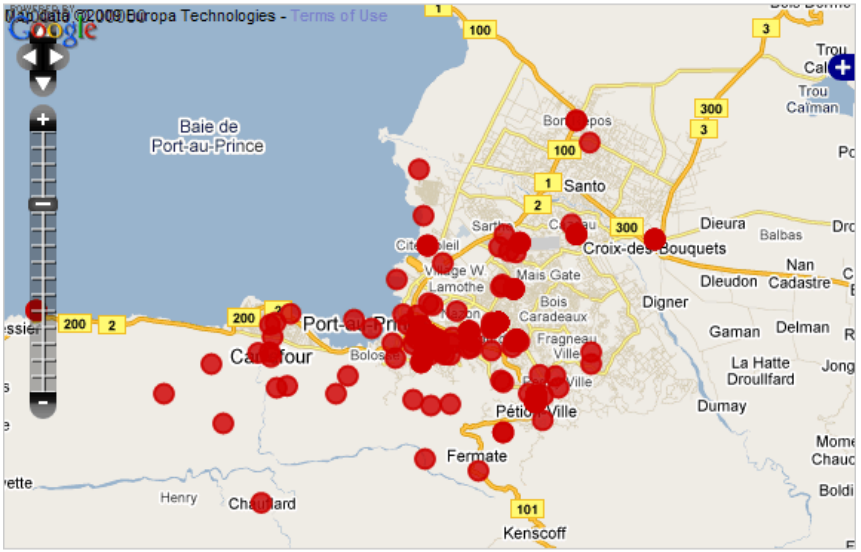


The 2010 Earthquake in Haiti

- Magnitude of 7 – the worst to strike Haiti in 200 years and resulted in severe infrastructure damage and over 200,000 deaths
- Many international organizations came to help as emergency responders, search and rescue teams and aid distributors, but had difficulty collectively prioritizing and physically locating areas of need.

Ushahidi in Haiti

- Allows users to take information from social media outlets that has a location associated with it and put it on a map
- 85% of households had access to cell phones and the towers were repaired in the days following the earthquakes
- Volunteers were responsible for manually translating and categorizing the incoming information



Examples

Text Messages

- *National Palace has extensive damage. President and First Lady are safe.*
- *THERES A BRIDGE IN DELMAS 31 THAT IS DAMANGED, THERES NO FOOD AND NO GAS YET*

Categories

- Medical Emergency, People Trapped, Fire, Water Shortage, Collapsed Structure, etc…
- Messages can be classified into more than one category

The Data

- Originally 3,598 Text Messages, Tweets, and other text from social media sent in the wake of the 2010 Haitian Earthquake
  - http://haiti.ushahidi.com/
- 3,561 of these are in English or were translated into English. The rest are in French or Haitian Creole.

Hypothesis

- Hypothesis: The use of Higher Order Topic Modeling of text from social media will efficiently inform disaster responders.
  - Hypothesis: The use of HO-LDA for topic modeling will outperform LDA
  - Hypothesis: The use of HO-LDA for topic modeling will outperform LDA for small training sample sizes
    - I will be helping run trials in this area

For My Project:

- **Hypothesis: Determining optimal number of topics to be used in LDA/HO-LDA will improve the classification metrics: accuracy, precision, recall, F-measure**

Methodology

- Preprocess
  - Done in Python
  - Remove common words (and, the, or,…), punctuation, and words occurring no more than once to make a dictionary
  - Returns a vector for each document containing the dictionary words in the sentence represented as their numerical labels
- LDA/ HO-LDA
  - Choose number of topics
  - Choose Sample Size
    - Randomized and Stratified
- J48 in Weka
  - Accuracy, Precision, Recall, F-Score

Methodology - Sampling

- Preprocess in Python
- Given a topic size:
  - For training size 100%
    - Run LDA and HO-LDA
    - Run J48 30x with randomly generated seed values
  - For training size ≠ 100%
    - Run LDA and HO-LDA for 30 randomized, stratified samples
    - Run J48 for each LDA or HO-LDA run using the same seed value each time

To Work Out – Number of Topics

- Determine metrics for judging the optimal number of topics
- Determine where in our process the optimal number of topics will be evaluated. Can it be part of the preprocessing?
- Develop a method for determining optimal number of topics

Conclusion

- Ultimately, we aim to show that HO-LDA outperforms LDA for small sample sizes and develop a method for determining the optimal number of topics.
- We will investigate this using text data from text messages and social media in the wake of the 2010 Haitian Earthquake.

Citations

- Blei, D. M. ,Griffiths, T. L., Jordan, M. I., *The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies*
- Blei, D., and Lafferty, J. *Topic Models*
- Celikyilmaz, A., Hakkani-Tur, D., *A hybrid Hierarchical Model for Multi-Document Summarization*
- Chaturvedi, S. K., Swami, D. K., Singh, G., *Dirichlet Distribution with Centroid Model (DDCM) based Summarization Technique for Web Document Classification*
- Ganiz, M. C., Lytkin, N. I., Pottenger, W. M., *Leveraging Higher Order Dependencies Between Features for Text Classification*
- Grün, B., Hornik, K., *topicmodels: An R package for Fitting Topic Models*
- Hoffman, M., Blei, D., Cook, P., *Content-based musical similarity computation using the Hierarchical dirichlet process*
- Hong, L., Davison, B. D., *Empirical Study of Topic Modeling in Twitter*
- Jung, Y., Park, H., Du, D., Drake, B. L., *A Decision Criterion for the Optimal Number of Clusters in Hierarchical Clustering*
- Sugar, C. A., James, G. M., *Finding the Number of Clusters in a Dataset: An Information-Theoretic Approach*
- Tibshirani, R., Walther, G., Hastie, T., *Estimating the Numbers of Clusters in a Data Set Via the Gap Statistic*
- haiti.ushahidi.com

J48 in Weka

- Used for classification
- Training data: set of samples each with a vector of attributes. Also associated to each sample is a class.
  - Uses decision trees
    - At each node, an attribute is chosen to split the sample
    - Continues this process on the subsets

Metrics

- Accuracy $\frac{\text{True positives}}{\text{Total}}$
- Precision $\frac{\text{True positives}}{\text{True positives + False positives}}$
- Recall $\frac{\text{True positives}}{\text{True positives + False negatives}}$
- F-Score $\frac{\text{2*precision*recall}}{\text{precision + recall}}$