AM 207 Project Proposal

Cole Diamond, Wei Dai, Raphaël Pestourie

April 9, 2015

1 What is the problem you are addressing?

- Our area of interest is in **Information Retrieval** systems.
- The general project idea is to **identify latent topics** in a corpus of documents.
- Our corpus will be 10 books from Project Gutenberg
- We will use our model to perform inference on a randomly selected page from our corpus to predict which book it originated from.
- We will employ **bayes theorem** and **gibbs sampling** to perform inference

2 What has been done already?

Significant progress has been made on this problem by researchers in the field of Information Retrieval

2.1 TF-IDF

Description

- The text-frequency inverse-document frequency scheme [1] uses a vocabulary of words acros all documents, and, for each document in the corpus, a count is formed of the number of occurrences of each word.
- After normalization, the term frequency count is compared to an inverse document frequency count, which measures the number of occurrences of a word in the entire corpus #### Advantages
- Tf-idf reduction can perform basic identification of sets of words that are discriminative for documents in the collection ##### Shortcomings
- Provides a relatively small amount of reduction in description length
- Reveals little in the way of inter or intradocument statistical structure.

2.2 LSI

- Latent Semantic Indexing [2] uses a singular value decomposition of the X matrix to identify a linear subspace in the space of tf-idf features that captures most of the variance in the collection.
- This addresses the reduction problem between does not give information on the inter and intra-document structure

2.3 pLSI

- The probabilistic Latent Semantic Indexing [3] model attempts to relax the simplifying assumption made in the mixture of unigrams model that each document is generated from only one topic.
- pLSI posits that a document label d and a word w_n are conditionally independent given an unobserved topic z.

$$p(d, w_n) = p(d) \sum_z p(w_n|z)p(z|d)$$

- Each word is generated from a single topic, and different words in a document may be generated from different topics.
- Each document is represented as a list of mixing proportions for these mixture components and thereby reduced to a probability distribution on a fixed set of topics #### Shortcomings

(1) The number of parameters in the model grows linearly with the size of the corpus, which leads to overfitting
(2) It is not clear how to assign probability to a document outside of the training set.

2.4 LDA

- LDA overcomes both of these problems by treating the topic mixture weights as a k-parameter hidden random variable rather than a large set of individual parameters which are explicitly linked to the training set [4]
- LDA generalizes easily to new documents.
- Furthermore, the parameters in a k-topic LDA model do not grow with the size of the training corpus
- Gibbs sampling can be used to perform learning with LDA [5]

2.5 Papers

2.6 What are the questions you are trying to answer?

Given an unseen page from a book, can we predict the book title using its topical composition?

2.7 What methodology are you planning to use?

- Since LDA seems to be more robust than TF-IDF, pLSI and LSI in topic modeling, we plan to use this technique.
- In LDA, documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.
- We assume that there are a fixed universe of topics that produce words in a corpus.
- The figure below illustrates this model.

LDA assumes the following generative process for each document w in a corpus C:

- Choose $N \sim poisson(\xi)$.
- Choose $\theta \sim Dir(\alpha)$.
- For each of the N words w_n:
  - Choose a topic $z_n \sim Mult(\theta)$.
  - Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

Given the parameters α and β, the joint distribution of a topic mixture θ, a set of N topics z, and a set of N words w is given by:

$$p(\theta, z, w, \alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta),$$

1. We use Gibbs sampling to sample from the conditionals of the posterior of Latent Dirichlet Allocation.

2. We perform inference on a new, unseen document to predict the topics that it references.

3 What data do you have available?

1. Frankenstein - https://www.gutenberg.org/cache/epub/84/pg84.txt
2. The Adventures of Sherlock Holmes - https://www.gutenberg.org/cache/epub/1661/pg1661.txt
3. A tale of two cities - https://www.gutenberg.org/cache/epub/98/pg98.txt
4. Moby Dick - https://www.gutenberg.org/cache/epub/2701/pg2701.txt
5. Beowulf - https://www.gutenberg.org/cache/epub/16328/pg16328.txt
6. Dracula - https://www.gutenberg.org/cache/epub/345/pg345.txt
7. The Adventures of Huckleberry Finn - https://www.gutenberg.org/cache/epub/76/pg76.txt
8. Ulysses - https://www.gutenberg.org/cache/epub/4300/pg4300.txt
9. The Republic - https://www.gutenberg.org/cache/epub/1497/pg1497.txt
10. The Divine Comedy - https://www.gutenberg.org/cache/epub/8800/pg8800.txt

References

[1] G. Salton and M. McGill, editors. Introduction to Modern Information Retrieval. McGraw-Hill, 1983.

[2] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. Indexing by latent semantic analysis. Journal of the American Society of Information Science, 41(6):391–407, 1990.

[3] T. Hofmann. Probabilistic latent semantic indexing. Proceedings of the Twenty-Second Annual International SIGIR Conference, 1999.

[4] Blei, David M and Ng, Andrew Y and Jordan, Michael I, Latent dirichlet allocation. The Journal of machine Learning research, (3):993–1022, 2003

[5] Darling, W. M. A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, (Portland, Oregon, USA, 2011).