Community Question-Answering Archives

Goal

Discovering the topical structure of CQA archives

Automatic Tagging: Annotation of the content with topics discovered to help browse and understand the archive.

Question Answering: Given a newly submitted question, use estimated topics to retrieve relevant question-answers from the archive.

Properties of Questions-Answers

Answer topics influenced by question topics

Answer topics more technical and specific

Answers may contain additional topics that are correlated with question topics

Properties of Questions-Answers

Documentation exhibit multiple topics

Topics in answers are influenced by topics in question

Documents are generated by realizations of random variables that are sampled from probability distributions

Corpus contains two types of topics. Question topics (Q-topics) and Answer topics (A-topics)

Each question is a mixture of Q-topics

Each answer is a mixture of A-topics

Model captures the dependency between question topics and answer topics by conditioning each A-topic in an answer on Q-topics drawn from the topic distribution of the corresponding question.

Inference and Parameter Estimation

Essentially, an optimization problem: Model specifies a joint probability distribution with some parameters. What are the optimal values of the model parameters?

To fit the model to data, the generative process is inverted and parameter values are generated from given observations (the words)

Inference Output

Per-word topic assignments

Per-corpus topic distributions

Per-document topic proportions

This research was supported by an NSERC Strategic project and the Business Intelligence Network, an NSERC Strategic Network.

Dataset

Stackoverflow

Stackoverflow Statistics

#Questions: 1,188,585

#Answers: 3,009,767

#Tags: 27,659

#Users: 445,672

Avg. Answer per question: 2.4828

Avg. Tags per question: 2.0354

Avg. score of questions: 1.4927

Avg. score of answers: 1.9478

Dataset Creation

Input Variables: given

Latent Variables: posterior is inferred given the data

Observations: given

Example Question-Answer Pair

Question:

How do I replace all occurrences of a word in a document with another word? Any solution is welcome!

Answer:

In some languages, such as Perl, regular expressions (regex) allow you to search and manipulate text based on patterns. In other languages, standard string manipulation functions are available (e.g., `replace()` method from Java's string class). In other languages, such as Perl, regex's are integrated into the language itself. Utilities such as grep, vi, etc can also perform this type of pattern matching and replacement.

Experiments – Question Answering

QATM retrieval performance compared to LDA model in terms of Mean Average Precision and TopN measures

Results on the Stackoverflow dataset

Topical dependencies captured by QATM with examples of Q-topics and A-topics represented by their first 20 most probable words

Experiments – Automatic Tagging

Clustering results in terms of Precision, Recall and F-measure.

Tag Distribution in Training Set and Test Set