MTP Progress

Evaluation of LDA

Outline

- Topic models
- Evaluation method
- Corpus
- Results
- Future Work

Topic modeling

- A document may consist of multiple topics
- Topic modeling aims to find out the proportion of each topic in the document

Evaluation method

- Tag every document with the topic having highest proportion
- Train the topic model using untagged corpus
- Find out the topic distribution of test documents
- Check whether this topic distribution confirms with the tag for the each document
- 5-fold cross validation was used

Corpus

- 1273 files of different topics from DMOZ
- Computers, films, real estate, cooking and sports were the 5 topics
- Total tokens (after stop word removal):221373

of files / topic

Topic	# of files
Computers	164
Sports	213
Cooking	251
Real Estate	261
Films	384
Total	1273

5-fold cross validation

- Training files / fold = 1018
- Testing files / fold = 255

Results

- Avg. accuracy = 20.867
- This is very low as a document contains multiple topics
- Weighted evaluation can be used

Weighted Evaluation

- We get a topic distribution for each testing document
- This topic distribution is arranged in descending order of topic proportions.
- We can assign weights according to the rank given to original tag of the document

Algorithm

- matches=0, counts=0
- For each document
  - Find topic distribution
  - Switch(tag):
    - case(topic1): matches += 1
    - case(topic2): matches += 0.8
    - case(topic3): matches += 0.6
    - case(topic4): matches += 0.4
    - case(topic5): matches += 0.2
  - counts++
- Accuracy = matches/counts

Results

Fold	Matches	Counts	Accuracy (in percentage)
Fold 1	156	255	61.4
Fold 2	156	255	61.2
Fold 3	170	255	66.9
Fold 4	152	255	59.6
Fold 5	161	253	63.9

Average Accuracy	62.6

High probability words / topic

Computers	Films	Cooking	Real Estate	Sports
site	film	recipes	services	reviews
software	information	recipe	real	news
free	offers	including	company	interviews
systems	production	collection	estate	information
programming	courses	tips	includes	features
research	links	source	commercial	current
resources	videos	production	based	tennis
code	television	baking	development	running
information	cinema	breakfast	title	tournament

Future Work

- Design a new joint-topic sentiment model
- Implementation of the model