Modeling Community Question-Answering Archives

Zainab Zolaktaf, Fatemeh Riahi, Mahdi Shafiei and Evangelos Milios
Email : z.zolaktaf@gmail.com
Computer Science Department, Dalhousie University

1 Community Question-Answering Archives

Abstract

Community Question Answering (CQA) services contain large archives of previously asked questions and their answers. We present a statistical topic model for modeling Question-Answering archives. We refer to our model as Question-Answering Topic Model or QATM. The model explicitly captures relationships between questions and their answers by modeling topical dependencies. To analyze the model, we use it for Question Answering and Automatic Tagging. Our model achieves improved performance in retrieving the correct answer for a query question compared to the LDA model. In addition, for large numbers of topics, our model achieves better clustering performance in terms of F-measure.

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

Properties of Question-Answers

- Answer topics influenced by question topics
- Answer topics more technical and specific
- Answers may contain additional topics that are correlated with question topics

Goal

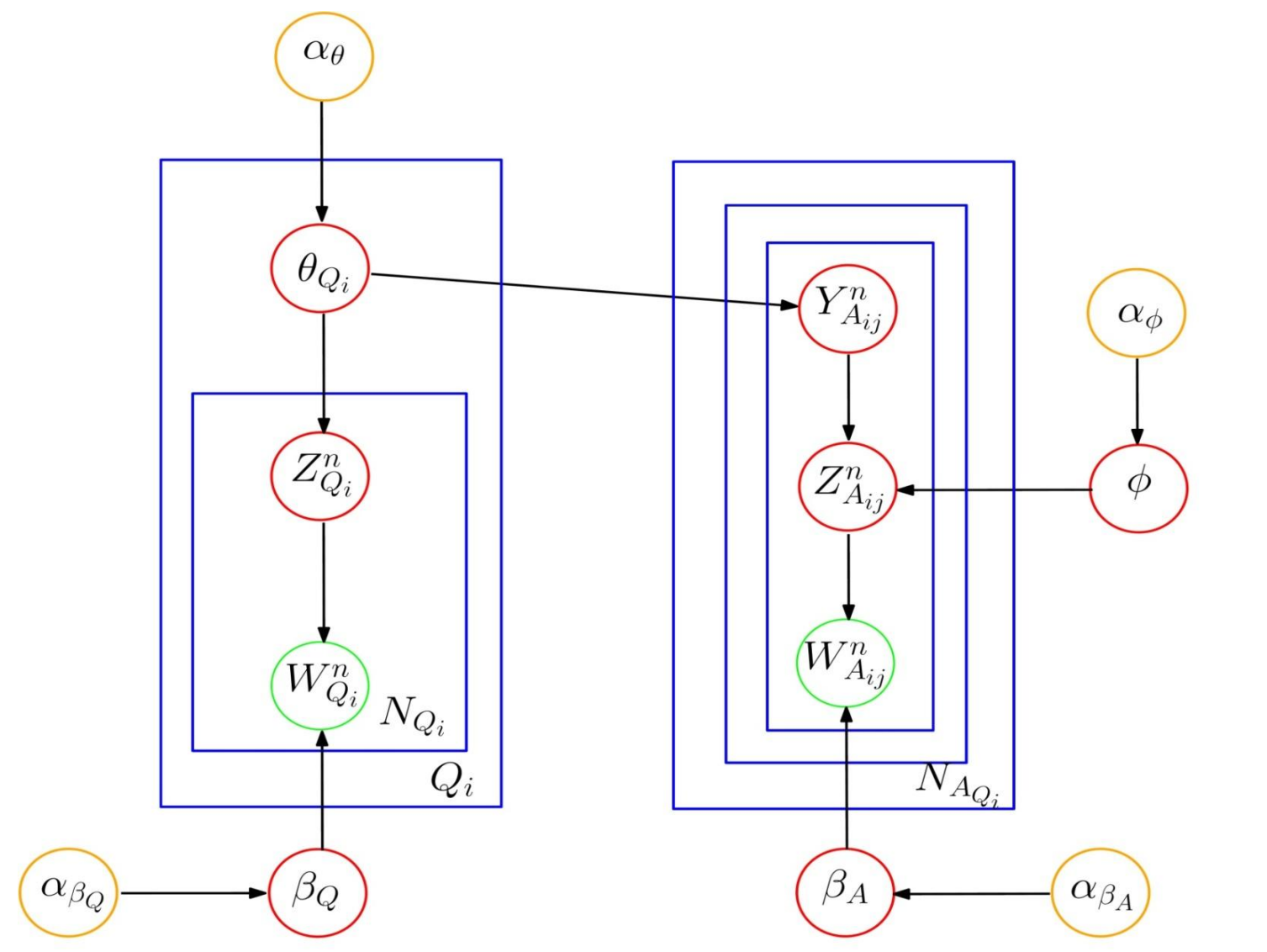
Discovering the topical structure of CQA archives

- Automatic Tagging: Annotating 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.

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

2 QATM

Graphical Model



Input Variables: given

Latent Variables: posterior is estimated given the data

Observations: given

Simple intuition

- Documents exhibit multiple topics
- Topics in answers are influenced by topics in question

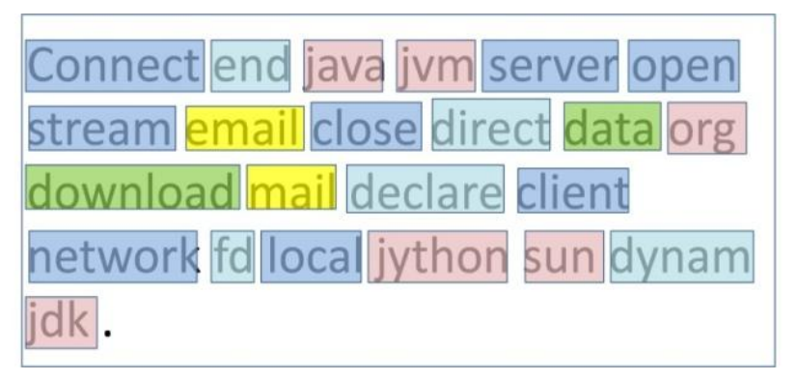
Generative Process

- 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

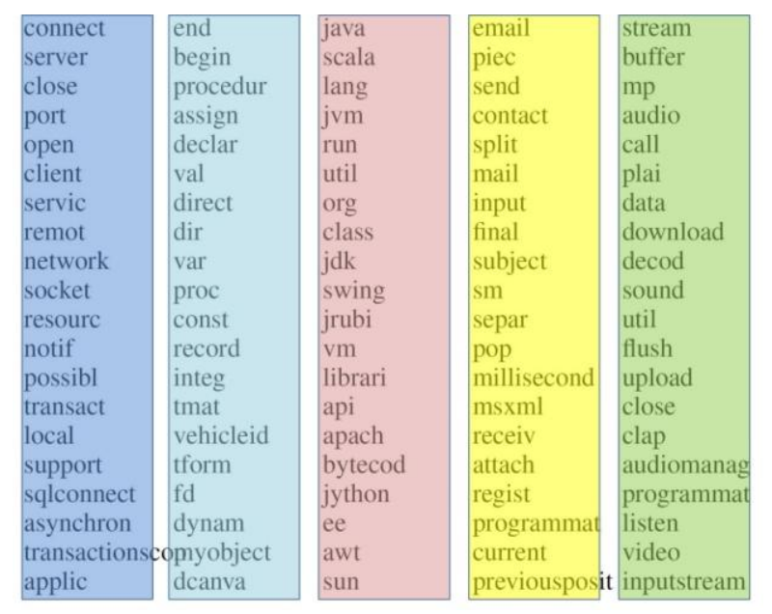
- Essentially, an optimization problem: Model specifies a joint probability distribution with some parameters. What are the optimal values of the model parameters?
- To fit the model to data, the generative process is inverted and parameter values are generated from given observations(the words)
- Exact inference is intractable. Gibbs sampling, an approximate inference algorithm is used.
- Inference Output

Per-word topic assignments

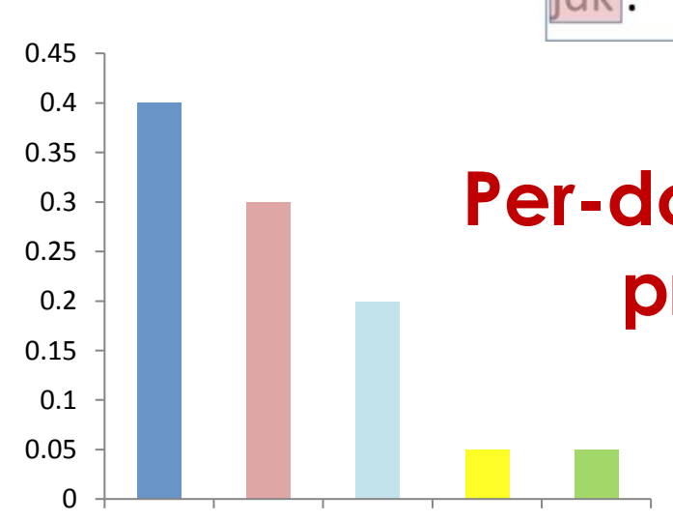


Connect end java jvm server open stream email close direct data org download mail declare client network fd local jython sun dynam jdk.

Per-corpus topic distributions



connect	end	java	email	stream
server	begin	scala	piec	buffer
close	procedur	lang	send	mp
port	assign	jvm	contact	audio
open	declar	run	split	call
client	val	util	mail	plai
servic	direct	org	input	data
remot	dir	class	final	download
network	var	jdk	subject	decod
socket	proc	swing	sm	sound
resourc	const	jrubi	separ	util
notif	record	vm	pop	flush
possibl	integ	librari	millisecond	upload
transact	tmat	api	msxml	close
local	vehicleid	apach	receiv	clap
support	tform	bytecod	attach	audiomanag
sqlconnect	fd	jython	regist	programmat
asynchron	dynam	ee	programmat	listen
transactionscop	myobject	awt	current	video
applic	dcanva	sun	previousposit	inputstream



Per-document topic proportions

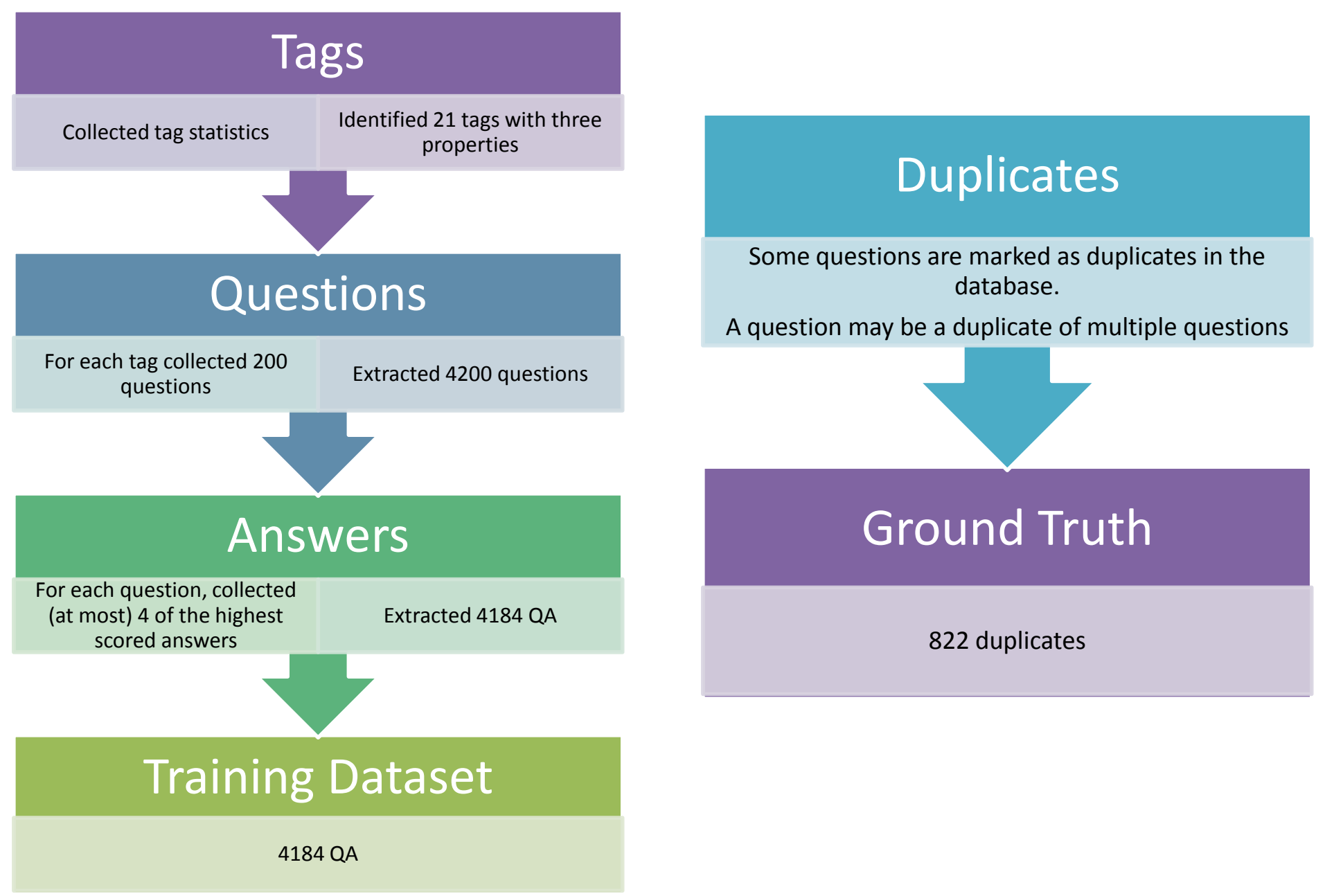
3 Dataset

Stackoverflow

Stackoverflow Statistics

#Questions	1,188,585
#Answers	2,939,767
#Tags	27,659
#Users	440,672
Avg #answer per question	2.4818
Avg #tags per question	2.9254
Avg score of questions	1.4967
Avg score of answers	1.8478

Dataset Creation



Tags
Collected tag statistics
Identified 21 tags with three properties

Questions
For each tag collected 200 questions
Extracted 4200 questions

Answers
For each question, collected (at most) 4 of the highest scored answers
Extracted 4184 QA

Training Dataset
4184 QA

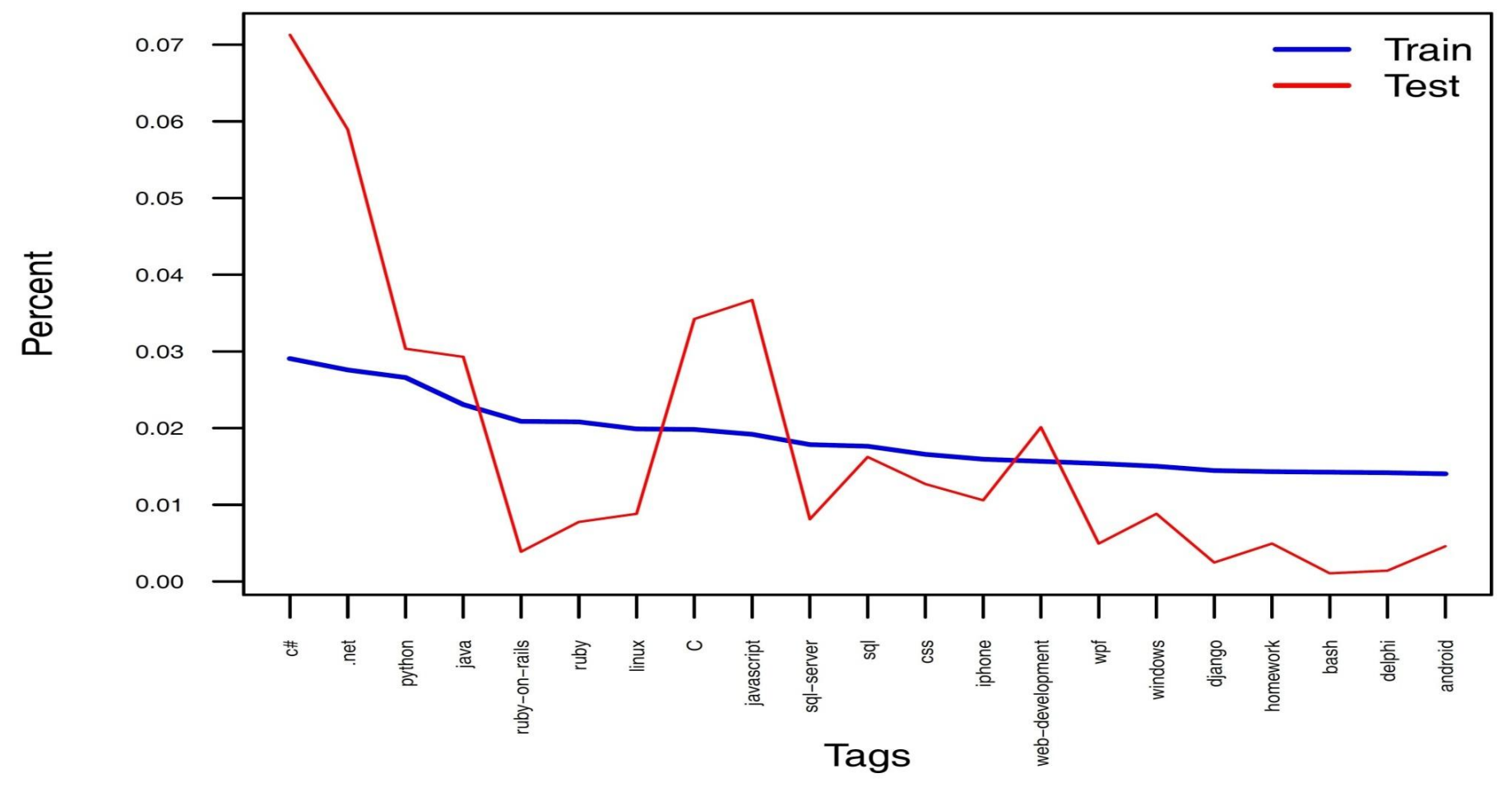
Duplicates
Some questions are marked as duplicates in the database.
A question may be a duplicate of multiple questions

Ground Truth
822 duplicates

Selected Tags

C#	.net	Sql
Sql-server	Java-script	css
Java	Ruby	Ruby-on-rails
Wpf	Iphone	Web-development
Android	Windows	Delphi
Django	Python	C
Bash	Linux	Homework

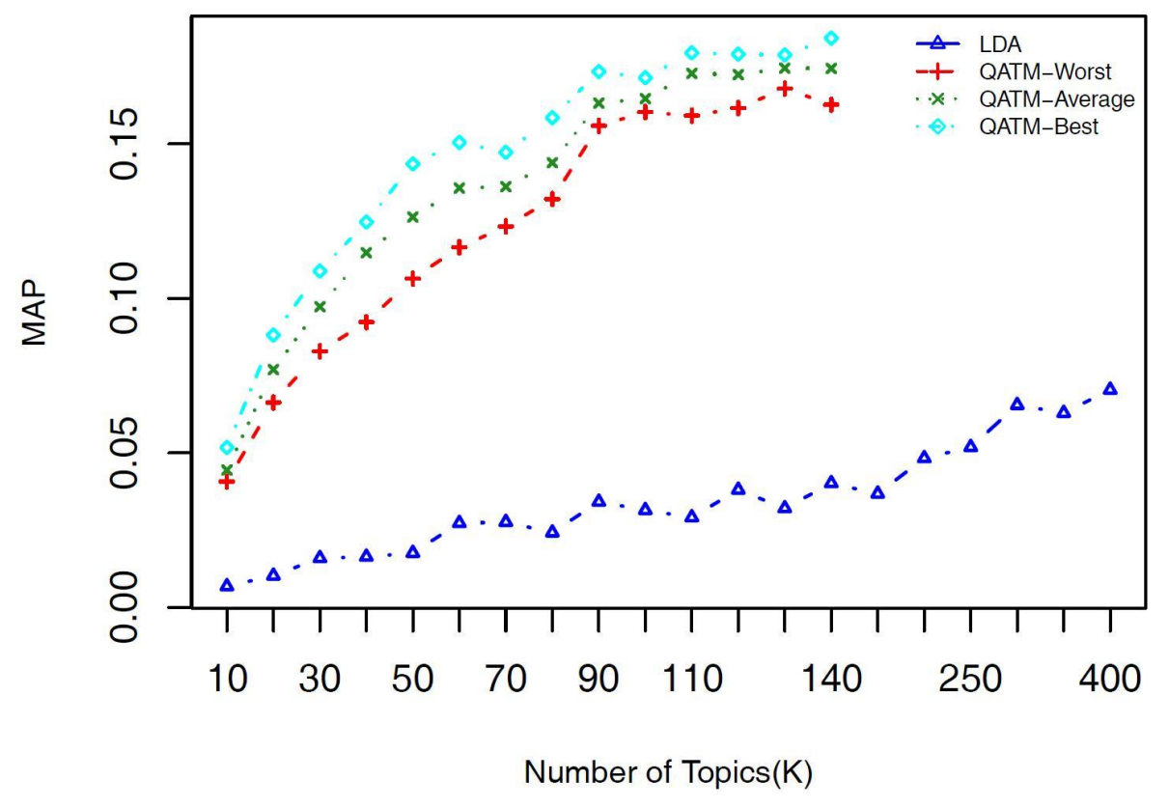
Tag Distribution in Training Set and Test Set



4 Experiments – Question Answering

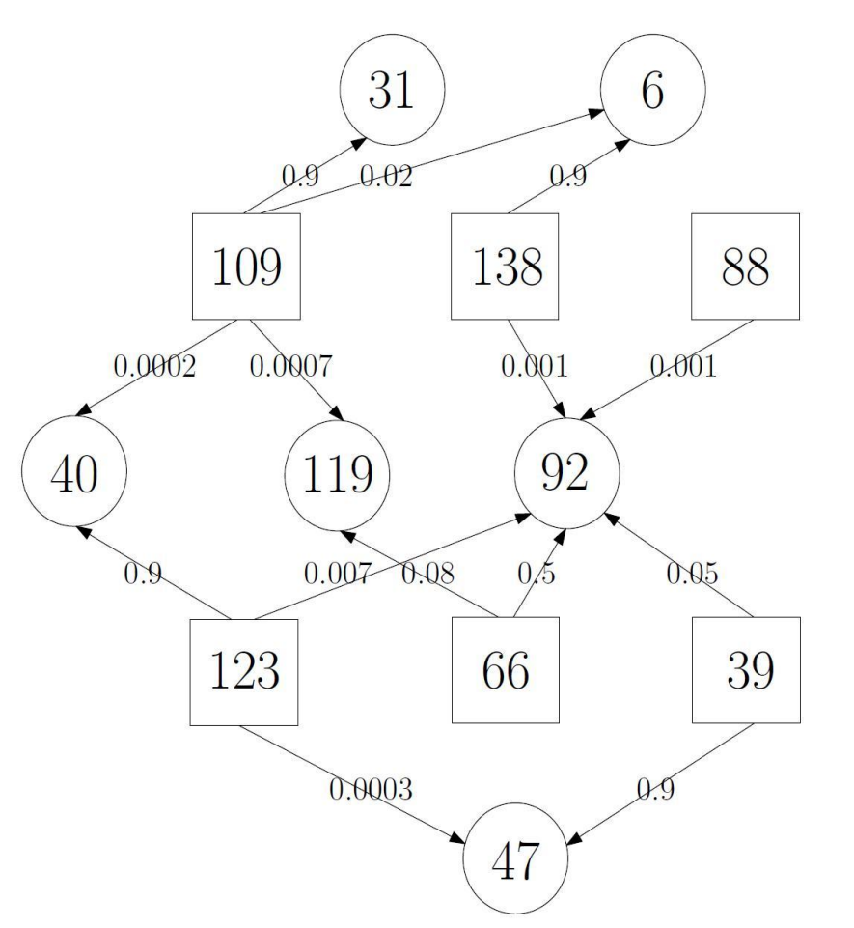
Results on the Stackoverflow dataset

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



	Top1	Top2	Top3	Top4	Top5
LDA	0.023	0.026	0.029	0.03	0.032
QATM-Worst	0.108	0.127	0.138	0.144	0.148
QATM-Average	0.122	0.142	0.151	0.156	0.16
QATM-Best	0.131	0.152	0.161	0.164	0.168

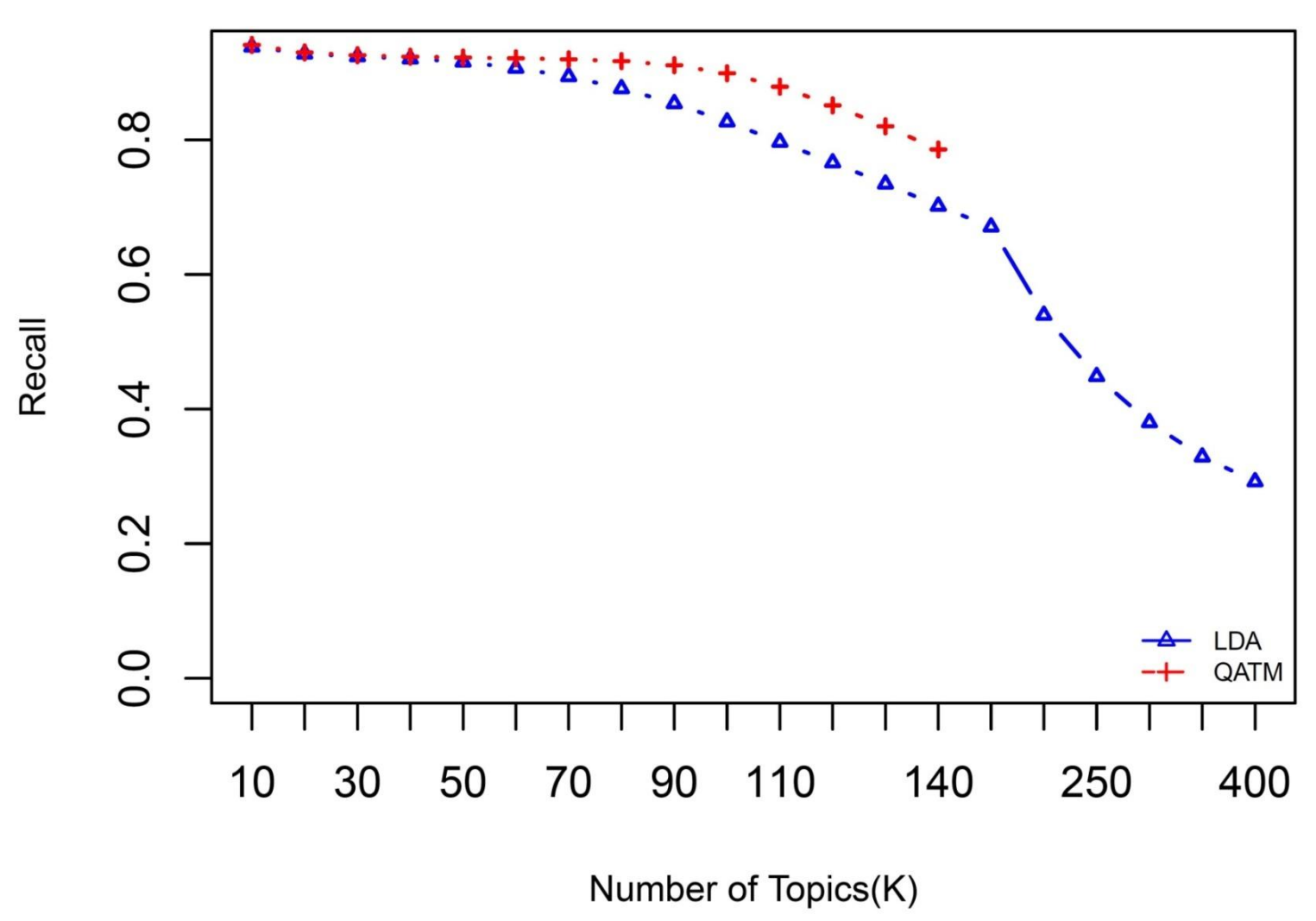
Topical dependencies captured by QATM with examples of Q-topics and A-topics represented by their first 20 most probable words

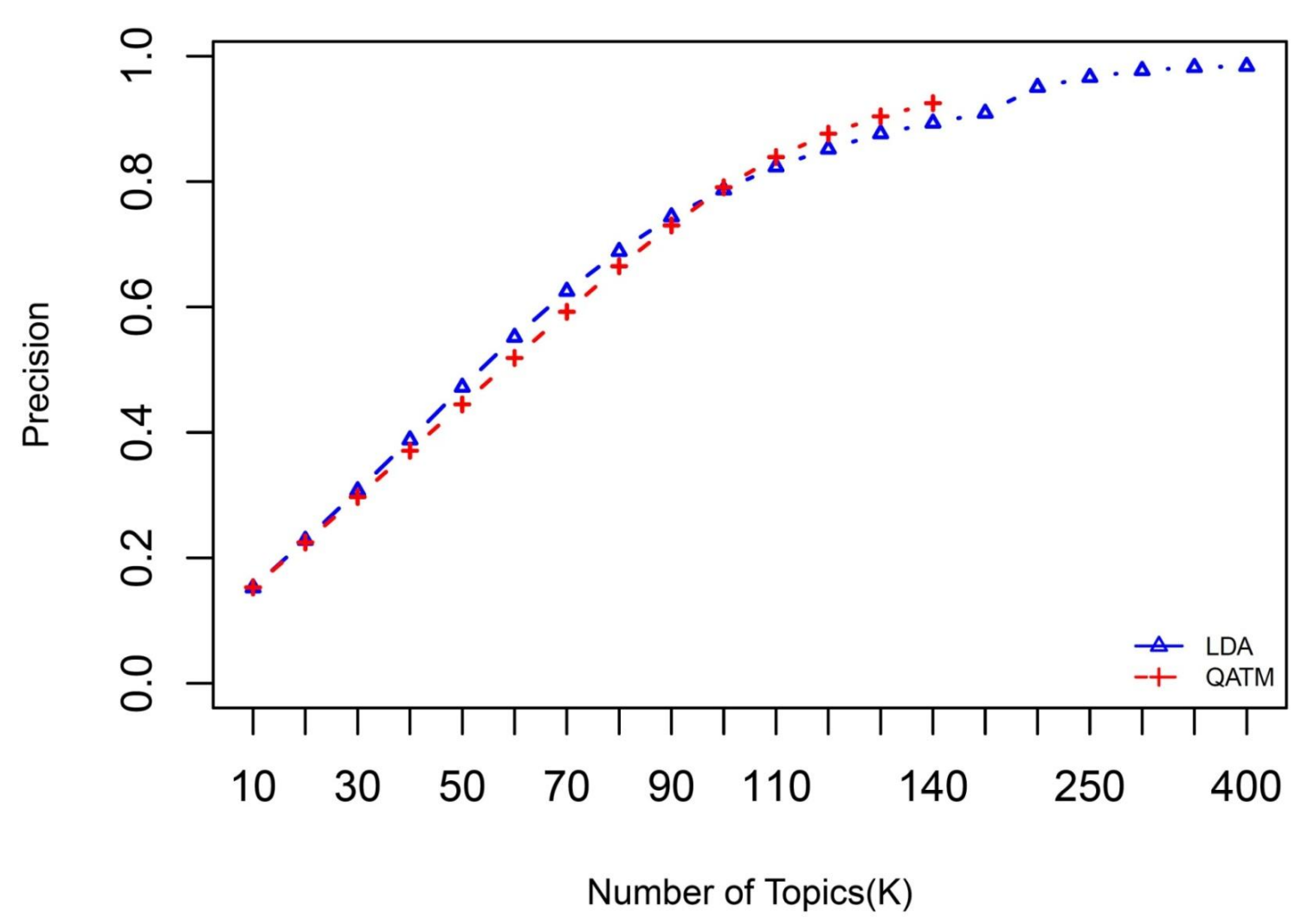
Topic Dependencies	Q-topic 39	Q-topic 66	Q-topic 88	Q-topic 109	Q-topic 123	Q-topic 138	A-topic 6	A-topic 31	A-topic 40	A-topic 47	A-topic 92	A-topic 119
	connect	end	java	email	stream	http	http	part	stream	connect	thi	answer
	server	begin	scala	piec	buffer	googl	www	email	buffer	socket	wai	question
	close	procedur	lang	send	mp	www	googl	send	data	port	ani	point
	port	assign	jvm	contact	audio	blog	org	home	length	network	onli	case
	open	declar	run	split	call	site	html	target	byte	send	doe	order
	client	val	util	mail	plai	org	blog	work	close	server	make	actual
	servic	direct	org	input	data	stackoverflow	articl	contact	mp	data	someth	follow
	remot	dir	class	final	download	en	en	messag	flush	client	check	give
	network	var	jdk	subject	decod	html	link	mail	reader	receiv	exampl	ye
	socket	proc	swing	sm	sound	sampl	aspx	sourc	upload	remot	doesn	expect
	resourc	const	jrubi	separ	util	wiki	wiki	sm	audio	summari	code	sinc
	notif	record	vm	pop	flush	place	post	relpath	output	proxi	befor	assum
	possibl	integ	librari	millisecond	upload	link	stackoverflow	split	memorystream	close	wa	consid
	transact	tmat	api	msxml	close	twitter	doc	address	payment	conn	note	happen
	local	vehicleid	apach	receiv	clap	msdn	archiv	smtp	dest	timeout	find	otherwis
	support	tform	bytecod	attach	audiomanag	home	msdn	gmail	bankhead	param	possibl	howev
	sqlconnect	fd	jython	regist	programmat	forum	librari	receiv	input	packet	abov	side
	asynchron	dynam	ee	programmat	listen	articl	github	subject	result	open	don	correct
	transactionscop	myobject	awt	current	video	aspx	found	spam	assert	implement	problem	inform
	applic	dcanva	sun	previousposit	inputstream	archiv	amp	bodi	filepath	host	put	sai

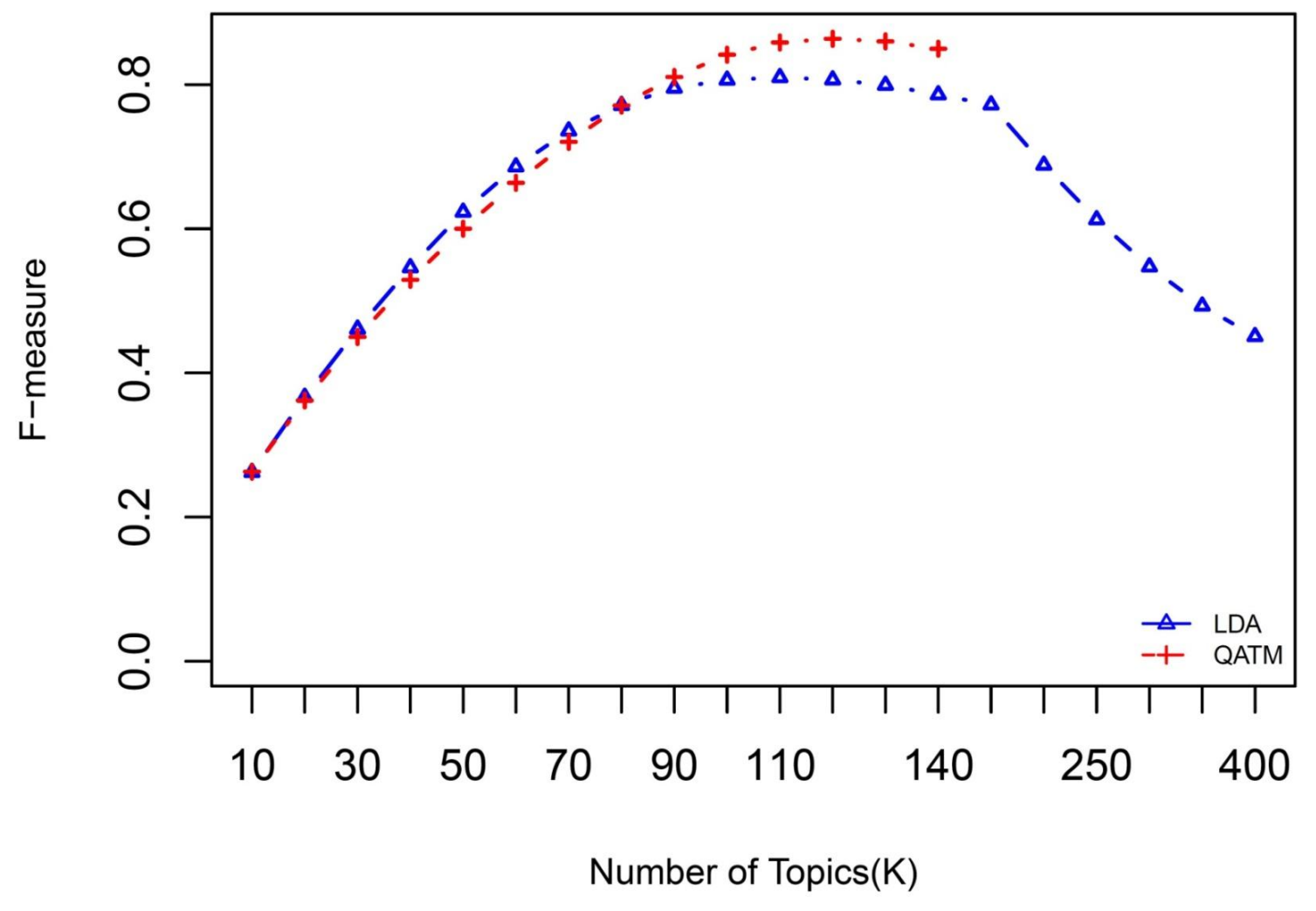
4 Experiments – Automatic Tagging

Results on the Stackoverflow dataset

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


Recall
LDA
QATM
Number of Topics(K)


Precision
LDA
QATM
Number of Topics(K)


F-measure
LDA
QATM
Number of Topics(K)