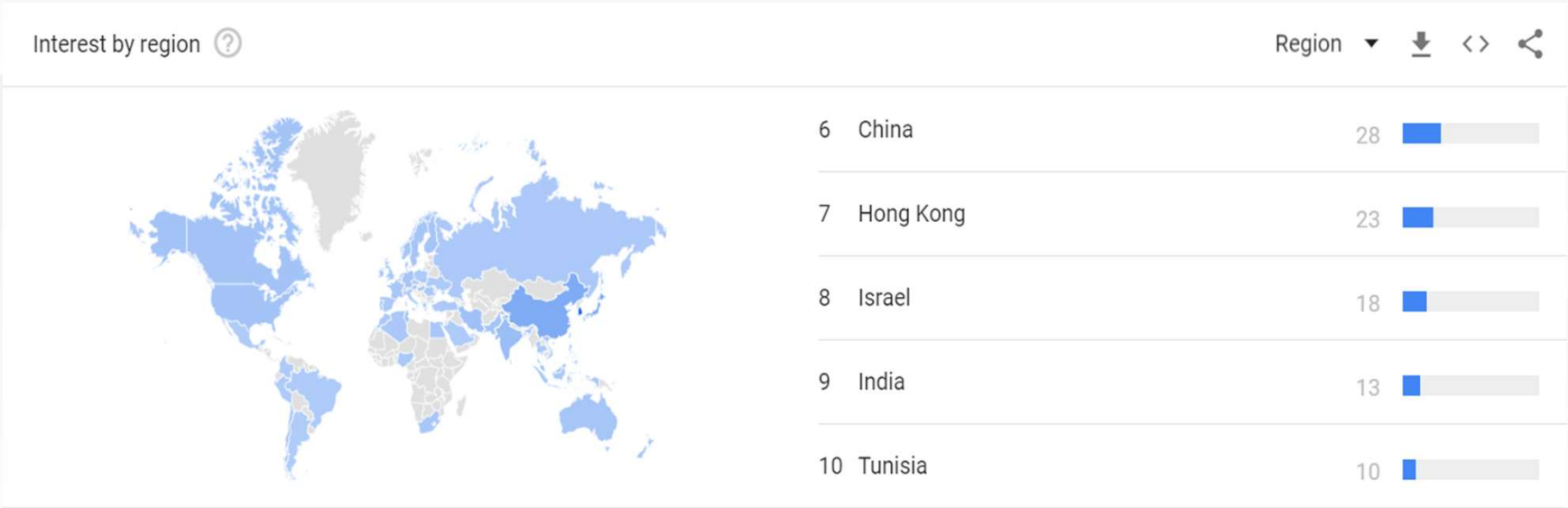
Mining What Developers Are Talking About Deep Learning

LYU1801
JIN Fenglei
Supervisor: Michael R. Lyu

CONTENTS

01 Introduction

Motivation



Search interest of deep learning

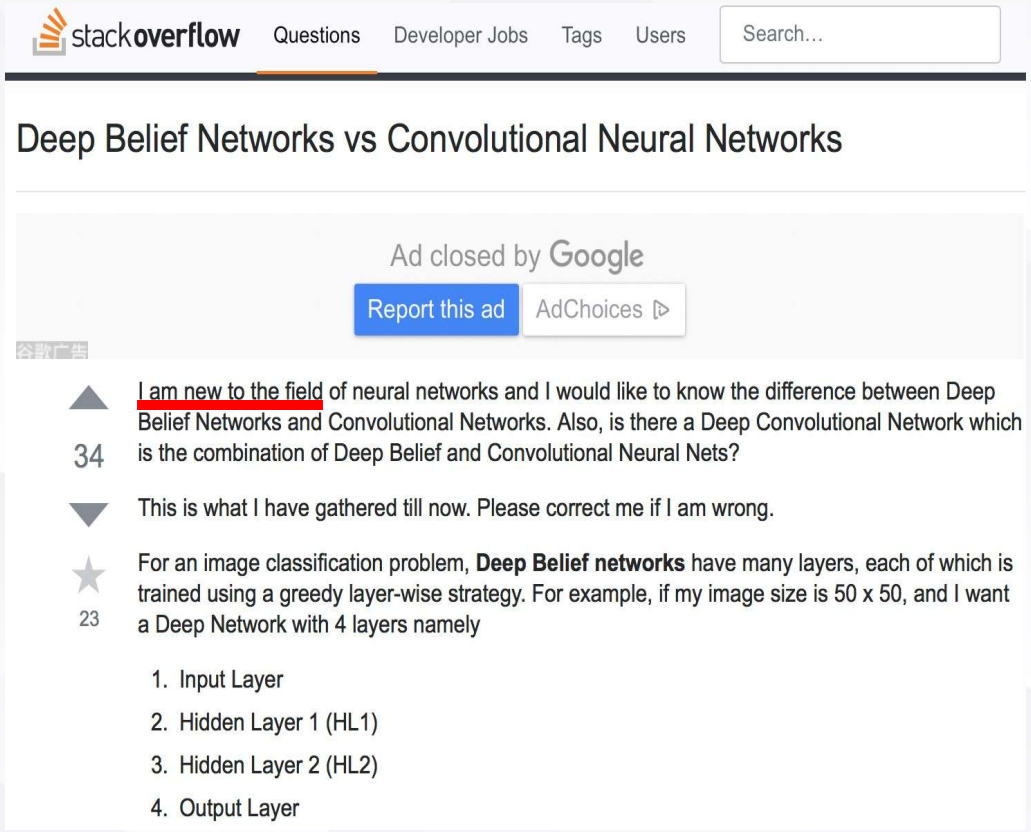
- Deep learning is popular!

Motivation

- **Lots of engineers and researchers are jumping into this area.**
  - More and more papers about deep learning
  - 36 FYP about deep learning this year!

Motivation

- Many new developers tend to enter this field and ask some basic questions.
- It is significant and necessary for the "newbies" to have a brief understanding about this field



Questions asked by "newbie"

Motivation

- Questions posted by developers directly reflect the focus of the deep leaning field
  - In October 2017, "Sophia", which is an AI robot and the first robot to receive citizenship at that time ,is popular.

- For experienced developers, knowing the newest information gives them inspiration.

Are the dialogs at Sophia's (the robot) appearings scripted?

7

2

I talk about the robot from: Hanson Robotics, which was granted the right to citizenship from Saudi Arabia.

I have found the following articles:

Your new friend is a humanoid robot

source: theaustralian.com.au

Like Amazon Echo, Google Assistant and Siri, **Sophia can ask and answer questions about discrete pieces of information**, such as what types of movies and songs she likes, the weather and whether robots should exterminate humans.

But her general knowledge is behind these players and she doesn't do maths. **Her answers are mostly scripted** and, it seems, from my observation, her answer are derived from algorithmically crunching the language you use.

Sometimes answers are close to the topic of the question, but off beam. Sometimes she just changes the subject and asks you a question instead.

She has no artificial notion of self. **She can't say where she was yesterday, whether she remembers you from before**, and doesn't seem to amass data of past interactions with you that can form the basis of an ongoing association.

Questions such as: *"What have you seen in Australia?"*, *"Where were you yesterday?"*, *"Who did you meet last week?"* and *"Do you like Australia?"* are beyond her.

Questions about "Sophia"

Contribution

- We propose a framework called IEDL to automatically track topic changes and Identify Emerging topics from Deep Learning-related posts in Q&A forum effectively
- We propose a novel topic interpretation method, which improve the topic coherence dramatically.
- We visualize the variations of the captured (emerging) topics along with time slices, with the emerging ones highlighted.

02 Related work

Related work

- Previous works for aspect extraction can be categorized into three approaches: **rule-based, supervised, and unsupervised**
  - LDA (Blei et al., 2003) and its variants are the most popular unsupervised approaches
  - Attention-based Aspect Extraction (ABAE) model (He et al., 2017)
  - On-line Latent Dirichlet Allocation (OLDA)
  - IDEA: with adaptively online latent Dirichlet allocation approach (AOLDA)

Previous Work

- Didn't extract phrases
- Didn't reconstruct sentences
- Didn't detect emerging topics
- No comparison

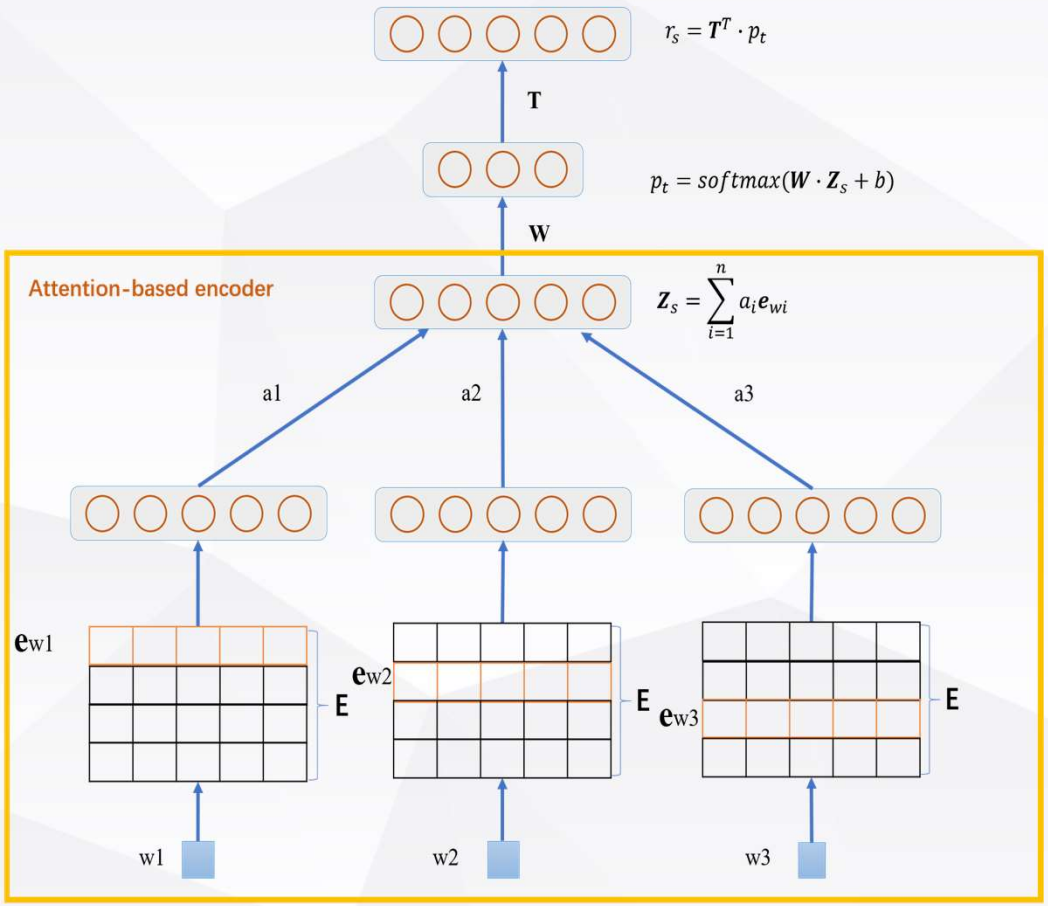
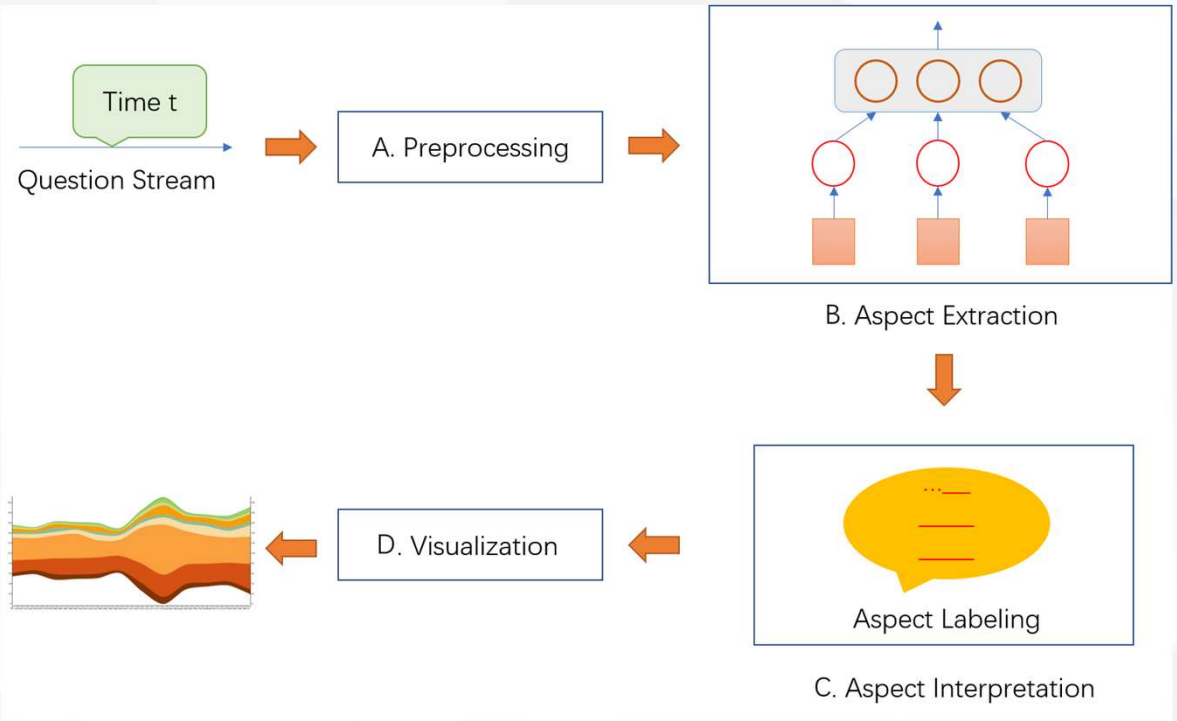


Illustration of ABAE of last semester

03 Methodology

Overview

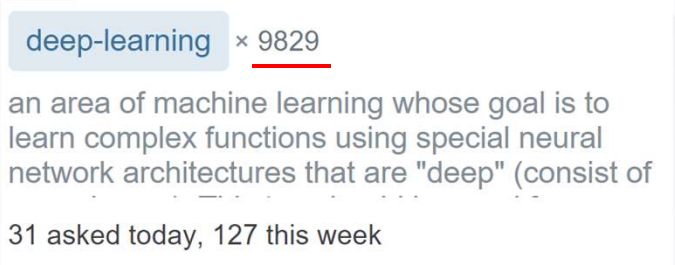
- Part A: preprocesses the raw posts
- Part B: extract aspects
- Part C: interpret the topic
- Part D: visualization



Framework of IEDL

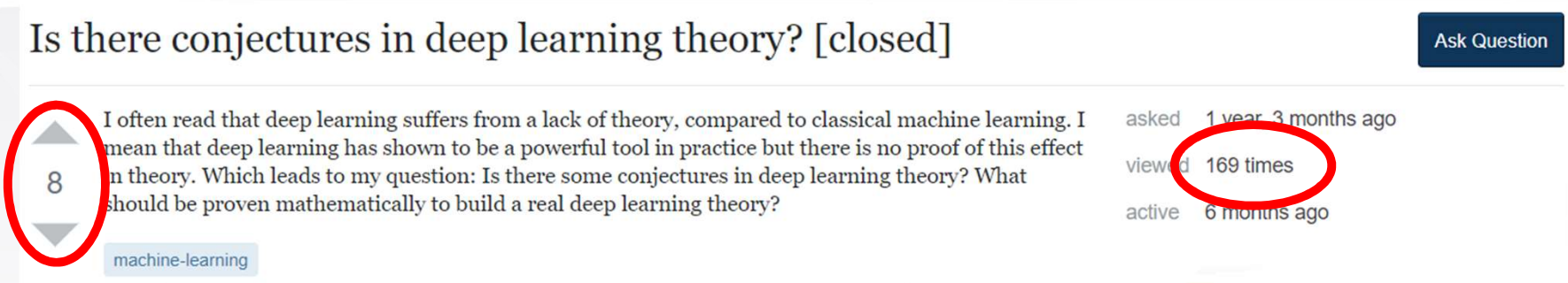
Data Crawling

- Over 7,000 questions provided by StackExchange
- Over 9,000 questions under the tag of deep-learning in StackOverflow
- Use a python package called scrapy to crawl the data in StackOverflow
- Enter the website of every question to crawl the detailed information



StackOverflow deep-learning tag

Data Analysis



StackExchange deep-learning questions

- **Votes** and **Views** are important attributes!

Data Labeling

- **We manually label 507 posts**
  - Categories: Image, NLP, Game-ai, Self-driving, Programming-languages, and Reinforcement-learning.
  - The labels are determined based on the tags provided by Stack Exchange and to maximize their distinguishability

Preprocessing

- massive noisy words
- codes, terminologies and websites
- HTML tags


```
{"title": "Reduce image dimensions in python",
 "question": "<div class=\"post-text\" itemprop=\"text\">\r\n\r\n<p>I have in input an
            image with dimensions (28, 28, 3).
            I trained a keras model with several images with dimensions (28, 28, 1). I
            want \n   to check a single test image with this model, but every time I get
            a dimension error. How can I reduce original dimensions (28, 28, 3) to (28,
            28, 1)?</p>\n\n<pre><code>test_image = image.load_img('test/number3.png' ,
            target_size = (28, 28))\ntest_image = image.img_to_array(
            test_image)\ntest_image = np.expand_dims(test_image, axis = 1)\nresult =
            classifier.predict(test_image)\n</code></pre>\n    </div>", "answer": "<div
            class=\"post-text\" itemprop=\"text\">\r\n<p>Depending on how you would like
            to reduce dimensionality you can just choose one of the colour channels like
            this</p>\n\n<pre><code>one_channel_image =
            test_image[:,:,0]\n</code></pre>\n\n<p>or you could find use the mean across
            the colour channels</p>\n\n<pre><code>one_channel_image = np.mean(
            test_image, axis=2)\n</code></pre>\n\n<p>In my experience of ML image
            problems just taking one channel works fine.</p>\n\n<p>If you need to
            increase dimensionality from (28, 28) to (28, 28, 1) you can use
            numpy.reshape</p>\n\n<pre><code>one_channel_image = test_image.reshape((28,
            28, 1))\n</code></pre>\n    </div>"},
```


Massive question

Preprocessing

- Word Formatting:
  - lowercase
  - lemmatization
- Word Filtering:
  - reduce the non-informative words
- Word Replacement:

Non-informative parts	Replacing words
Websites (eg: http://..., https://...)	url
All numbers	<digit>
Image html tag	img
Code, pseudocode	code
Unknown words in dictionary	<unk>

Preprocessing

- HTML Tags Summarization:

Tags	Description	Tags	Description
 	new line		ordered list
<hr>	thematic change in the content	<blockquote>	a section that is quoted from another source
	stress emphasis	<pre>	a preformatted text
	important text	<code>	a code or pseudocode (handled before)
<h1>, <h2>, <h3>	define HTML headings	<img src=…>	image (handled before)
	unordered (bulleted) list	…	

Preprocessing

Phrase Extraction :

$$PMI(w_i, w_j) = log \frac{p(w_i w_j)}{p(w_i)p(w_j)}$$

Extracted phrases: output_layer, dot_product, neural_network, initial_state, hide_layer,cross_entropy, cross_validation, computer_science,tic_tac_toe, activation_function.

Model

- Goal: learn a topic distribution and detect emerging topics

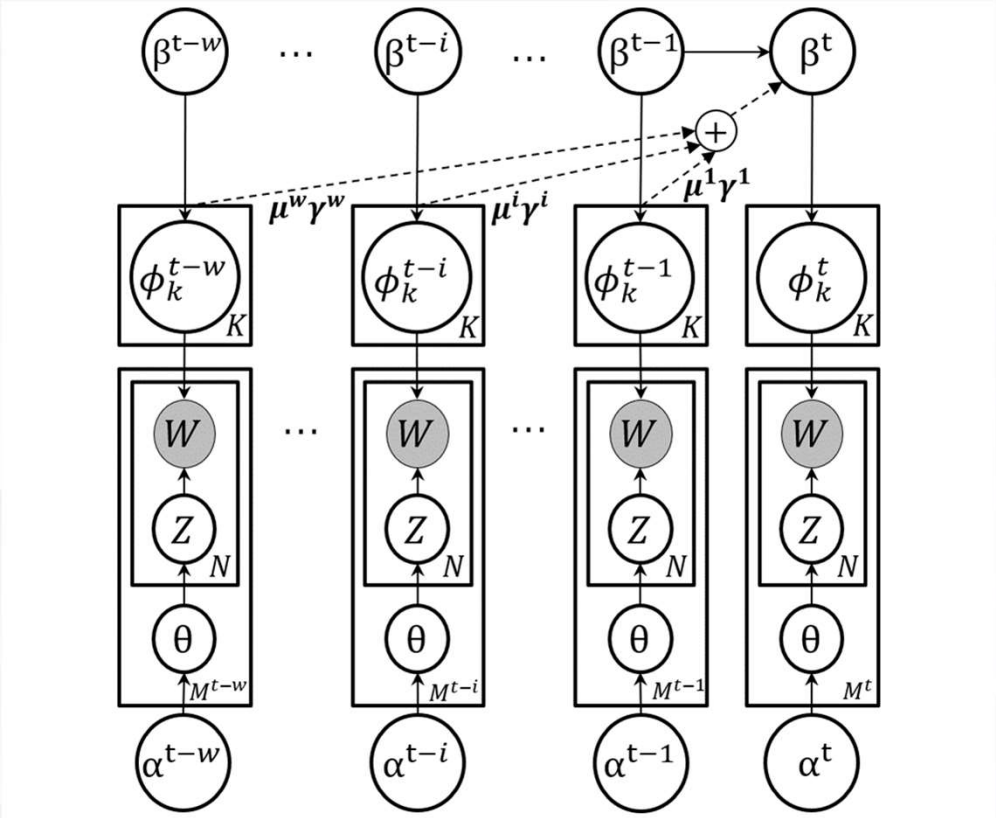


Illustration of IEDL

Model

- Normal LDA

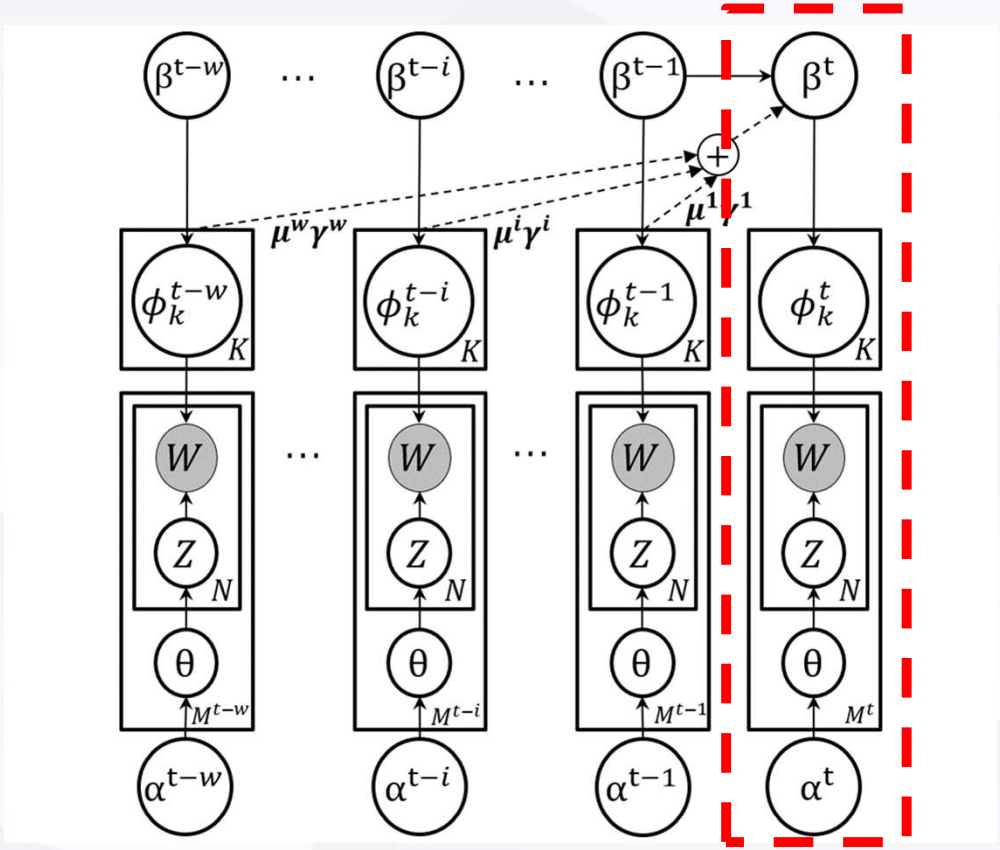


Illustration of IEDL

Model

- OLDA

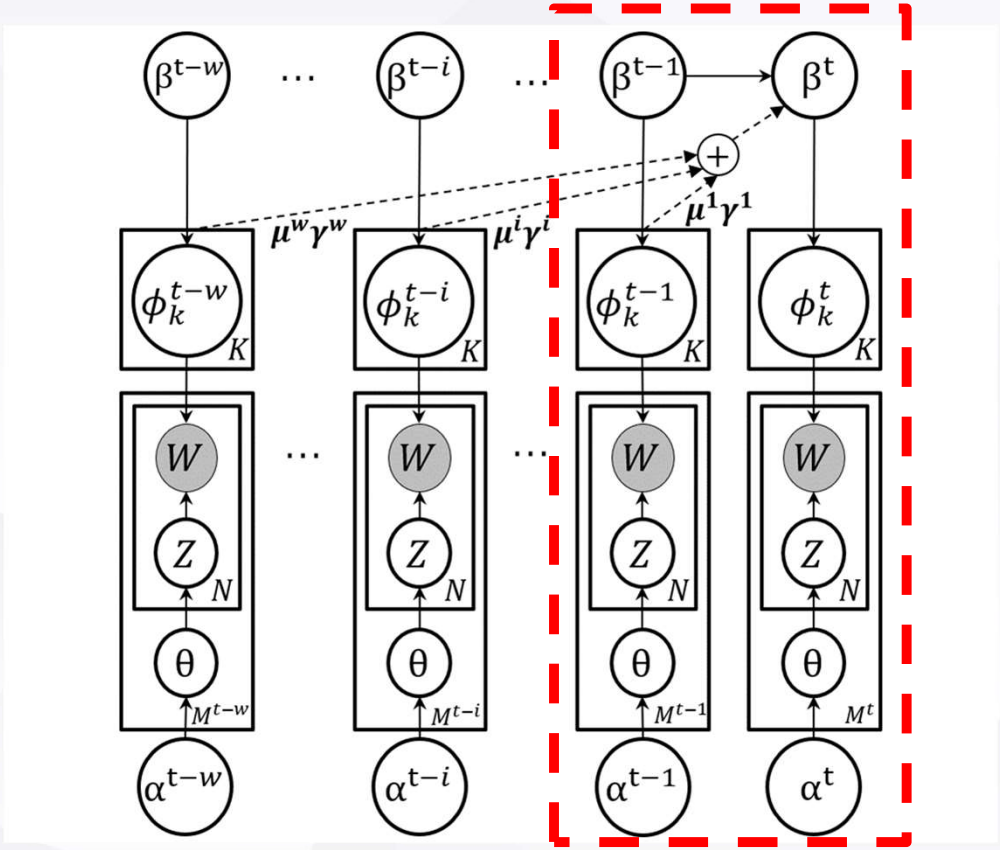


Illustration of IEDL

Model

$$\beta_k^t = \sum_{i=1}^{w} \mu^i \gamma_k^i \phi_k^{t-i}$$

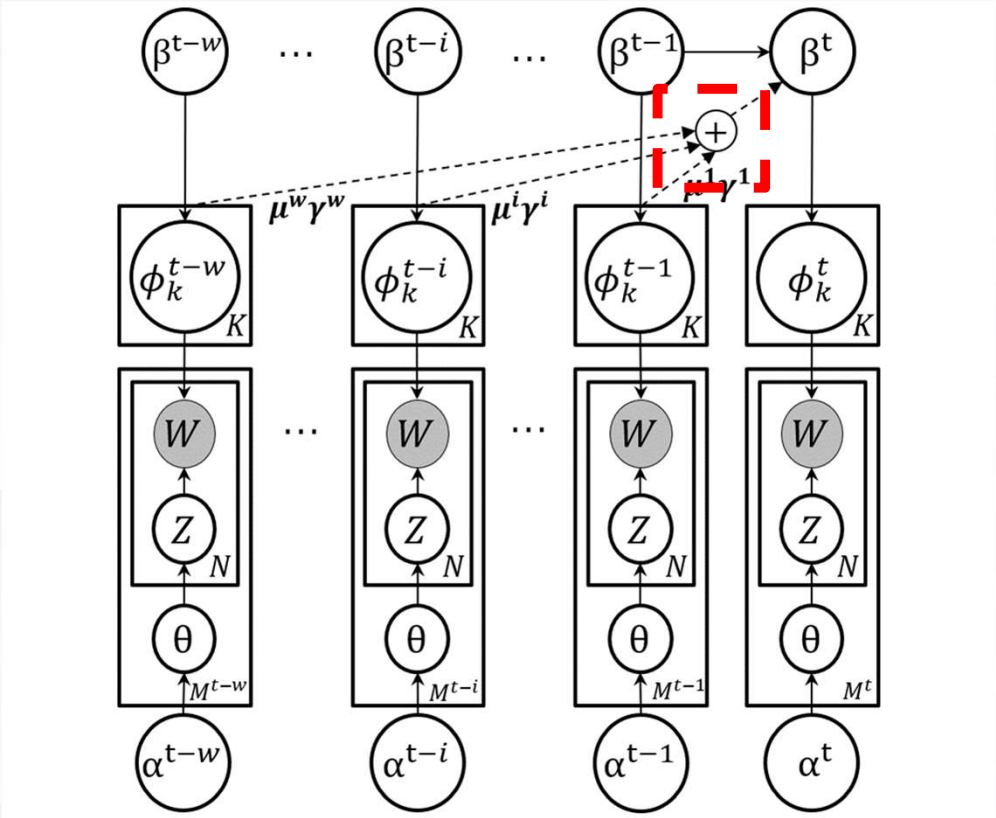


Illustration of IEDL

Model

$$\beta_k^t = \sum_{i=1}^{w} \mu^i \gamma_k^i \phi_k^{t-i}$$

$$\gamma_k^i = \frac{\exp(\phi_k^{t-i} \cdot \beta_k^{t-1})}{\sum_{j=1}^{w} \phi_k^{t-j} \cdot \beta_k^{t-1}},$$

$$\mu^i = \exp(-\lambda i),$$

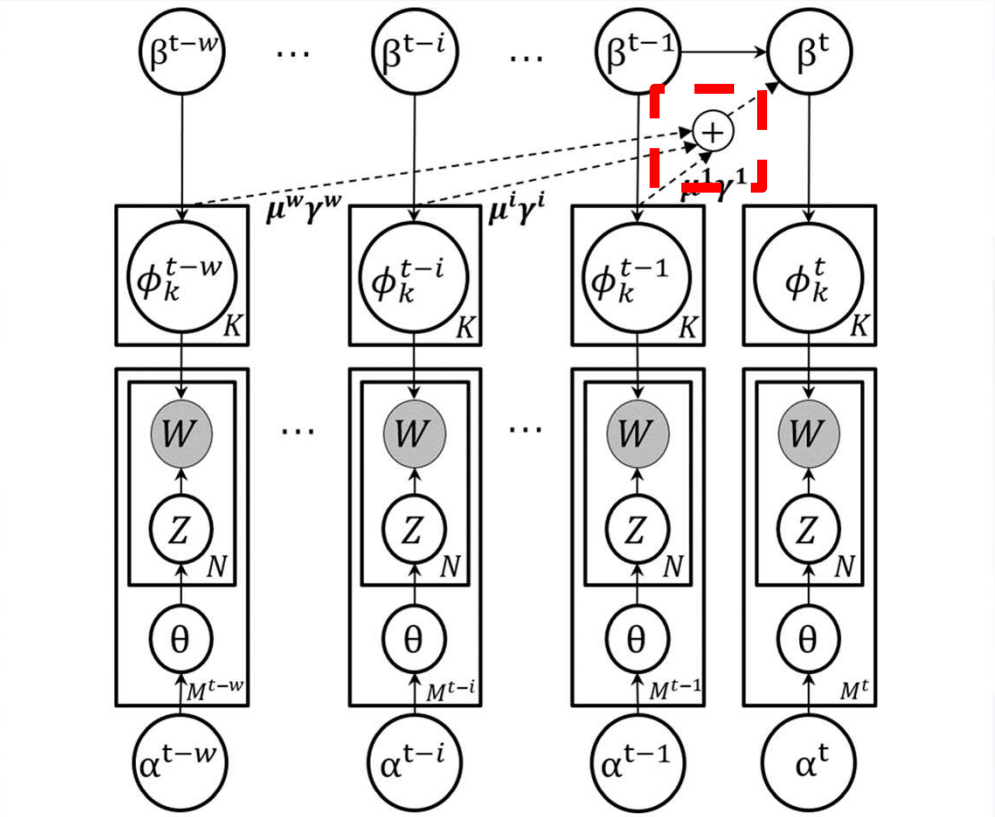


Illustration of IEDL

Model

Anomaly Detection:

$$D_{JS}(\phi_k^t || \phi_k^{t-1}) = \frac{1}{2} D_{KL}(\phi_k^t || M) + \frac{1}{2} D_{KL}(\phi_k^{t-1} || M)$$

$$M = \frac{1}{2}(\phi_k^t + \phi_k^{t-1})$$

$$D_{KL}(P||Q) = \sum_i P(i) log \frac{P(i)}{Q(i)}$$

Model

Automatic Topic Interpretation:

$$SCORE_{qua}(l) = \exp\left(-\frac{1}{\ln(v_l + 1)\ln(r_l + 1)} - \eta \times \frac{1}{\ln(h_l + 1)}\right)$$

- l is the post, and vl, rl, hl are the votes, views, and length of the post respectively.

03 Experiment

Dataset

- StackExchange: 7,067
- 507 labeled data
- Divided dataset in 2017 into 12 months

Month	Question No.	Month	Question No.
2017-01	294 questions	2017-07	358 questions
2017-02	226 questions	2017-08	458 questions
2017-03	288 questions	2017-09	358 questions
2017-04	306 questions	2017-10	374 questions
2017-05	272 questions	2017-11	350 questions
2017-06	228 questions	2017-12	378 questions
TOTAL	3,890 questions		

Classification

- We use the topic distribution of each post as features, and classify the 507 labeled posts by SVM.
- IEDL outperforms the baseline model by 5% for average precision.

Category	Method	Precision	Recall	F1
Image	IDEA	0.89	**0.73**	**0.80**
	IEDL	**1.00**	0.64	0.78
NLP	IDEA	0.68	0.76	0.72
	IEDL	**0.73**	**0.94**	**0.82**
Game-ai	IDEA	0.83	0.94	0.88
	IEDL	**0.83**	**0.97**	**0.90**
Self-driving	IDEA	0.94	0.89	0.91
	IEDL	**1.00**	**0.94**	**0.97**
Programming -language	IDEA	**0.92**	0.73	0.81
	IEDL	0.86	**0.86**	**0.86**
Reinforcement -learning	IDEA	0.86	**0.86**	**0.86**
	IEDL	**1.00**	0.62	0.76

Topic Coherence

OLDA	IDEA	IDEA+Quality Score	IEDL
0.133	0.166	0.217	**0.222**

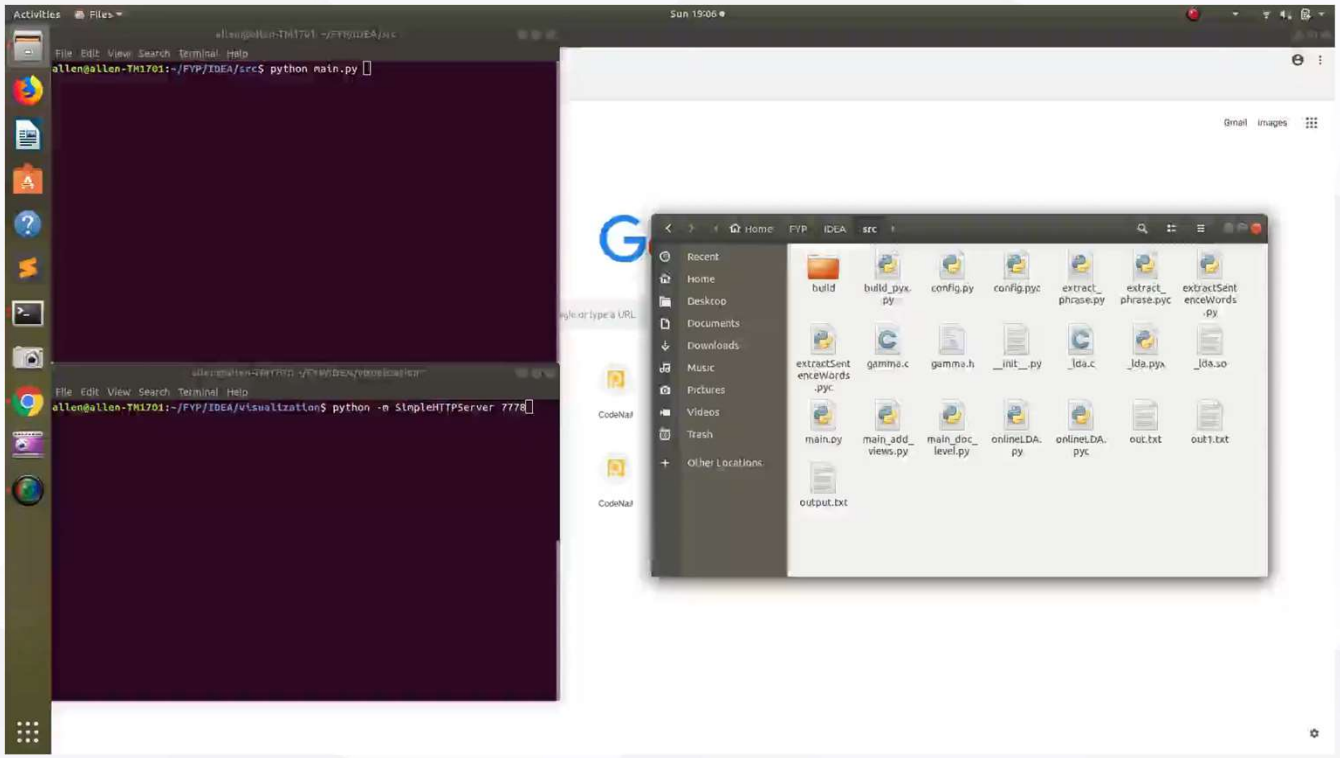
- IEDL improves the topic coherence greatly!

Visualization

- https://github.com/AllenFenglei/IEDL

- ./run normal
- ./run test
- ./run views

- http://appsrv.cse.cuhk.edu.hk/~fljin7/fyp_term2/index.html



Case Study

How will morality questions be settled in the domain on self-driving cars?

Ask Question

asked 1 year, 6 months ago
viewed 158 times
active 1 year, 6 months ago

2

For example...

1) If a dog is crossing the road, I'd expect the car to try to avoid it. But what if this leads to .00001% more risk for the driver? What is the 'risk cut-off'?

2) What if a cockroach is crossing the road? Will the car have a list of animals okay to run over?

3) What if a kid is crossing the street and avoiding it would kill the driver?

These questions seem to not really have an answer, yet self driving cars are almost ready. What are they doing about all of this?

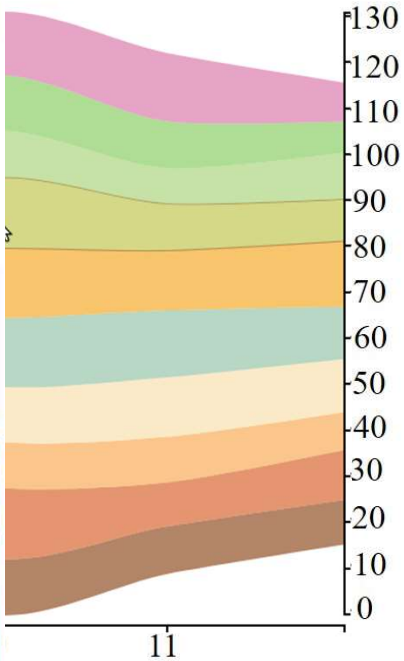
ai-design self-driving

share improve this question

asked Oct 12 '17 at 21:46

Featured on Meta

Unicorn Meta Zoo #2: What is the role of moderators?



Emerging Topics: self drivi
Sentence: 1: i also agree_th
2: in_my_opinion the_botto
3: however i can also imagi
4: although_there be potenti
5: for_example_if a human_

Tesla plans to unveil its all-electric semi truck on October 26th

Month later than Musk originally announced

By Zac Estrada | @zacestrada | Sep 13, 2017, 7:55pm EDT

Acknowledgement

- My deeper gratitude goes to my supervisor Michael and PhD mentor Cuiyun Gao
- Submitted to ACL workshop & ICML workshop

THANK YOU