

Introduction to text mining training course

Pham Quang Nhat Minh

FPT Technology Research Institute (FTRI)
minhpqn@fpt.edu.vn

November 17, 2016

Table of contents

- 1 Course information
- 2 Project Description
- 3 Literature Review
- 4 Proposed Solution
- 5 Project Plan
- 6 Task Assignment
- 7 QA & Discussion

Table of Contents

- 1 Course information
- 2 Project Description
- 3 Literature Review
- 4 Proposed Solution
- 5 Project Plan
- 6 Task Assignment
- 7 QA & Discussion

- Instructor: Pham Quang Nhat Minh
 - Emails: minhpn2@fe.edu.vn, minhpham0902@gmail.com
 - Ph.D. in Information Science (research field: Natural Language Processing)
 - Now researcher at FPT Technology Research Institute (FTRI)
 - **Research Interests**
 - Factoid/Non-factoid Question Answering
 - Semantic Relation Extraction
 - Textual Entailment Recognition (RTE)
 - Contradiction Detection
 - Homepage:
http://ftri.fpt.edu.vn/?page_id=1826

- Short text mining training course (8-day course)
- Learning outcome:
 - Understand and practice some natural language processing techniques for English language.
 - Can use topic modeling methods (LDA) for text mining
 - Get experiences in doing a real text mining project
 - Can do good data visualization
 - Apply Web programming techniques to build a web-based demo for the course's project
 - *Assume that you have knowledge and experiences about that*
 - Apply security knowledge to simulate attacking/defending your website.
 - *Assume that you have knowledge and experience about that*

- Course project: **Mining hot tech topics/trends from collections of tech articles.**
 - Focus on **Big Data** and **Internet of Things (IoT)** tech articles
- Lectures in the course will support/toward the goal of the course project.
- Course's github repository: `https://github.com/minhpqn/Text-Mining-Training-Course`
 - Lecture slides
 - Resources (tools, data, etc)
 - Example source codes
 - References

- **Day 1**

- Introduce the training course
- Describe the project's requirements, task list
- Assign/Discuss tasks for internship students

- **Day 2**

- Using topic modeling for mining topics/trends in raw text corpora
- Recommended readings for topic modeling & applications for mining topics
- Topic modeling tools

- **Day 3**

- Collecting data for the project (tech news articles about big data and IoT)
- Using Scrapy (Python package) to crawl data on the internet
- Introduce some data resources for crawling raw data
- How to processing crawled text data

- **Day 4**

- Using nltk for processing crawled text data
- Transform text data to the data in the format of LDA tools

- **Day 5**

- Run LDA to train topic models
- Observe & analyze output

- **Day 6**

- Data visualization & Make a simple website

- **Day 7**

- Review the product

- **Day 8**

- Project demo & defense

Table of Contents

- 1 Course information
- 2 Project Description
- 3 Literature Review
- 4 Proposed Solution
- 5 Project Plan
- 6 Task Assignment
- 7 QA & Discussion

- Objectives:
 - To detect research/technology trends by mining a large collection of articles.
 - Visualize mining results to show:
 - Distribution of topics in the corpus, strength of topics/trends
 - Change of topics/trends overtime
 - **Key phrases** in the corpus
 - Build a web-based demo for the course project
 - Automatically crawl data data given some input keywords about a technology (e.g., Big Data, IoT, etc)
 - Mine topics/**key phrases** and visualize the result

- Keep track the development of ideas/technologies is important in strategic decision making.
 - Which technology should we invest?
- It requires huge efforts to read large amount of tech articles
- Some trends may be obvious but others may be more subtle.
- Good data visualization make us easily to grasp trends/changes of topics overtime.

We need an automatic tool to do the job!

- What data sources from which we crawl the data?
 - Tech news (Techcrunch, Techmeme, etc) (HTML documents)
 - OR/AND Paper articles (in pdf format)
 - Abstracts of scientific articles (more available)
- How do we crawl data?
- How do we store the data?
- How do we process crawled data?
 - Clean text data
 - Transform to a corpus of the required format
- How do we mine topics/key phrases from the text corpus?
- How do we visualize mined result?

Table of Contents

- 1 Course information
- 2 Project Description
- 3 Literature Review**
- 4 Proposed Solution
- 5 Project Plan
- 6 Task Assignment
- 7 QA & Discussion

- David Hall et al. 2008. Studying the history of ideas using topic models. EMNLP 2008.
 - Apply LDA (Blei et al., 2003) on ACL Reference Corpus of scientific papers related to NLP and computational linguistics.
 - To address the change of topics over years, proposed **empirical probability** to calculate the probability that a document d written in year y was about topic z .

- David Hall et al. 2008. Studying the history of ideas using topic models. EMNLP 2008.
 - Apply LDA (Blei et al., 2003) on ACL Reference Corpus of scientific papers related to NLP and computational linguistics.

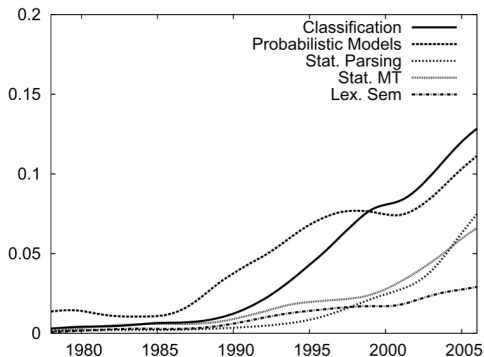


Figure: Topics in the ACL Anthology that show a strong recent increase in strength.

- Anton Barua et al. 2014. What are developers talking about? An analysis of topics and trends in Stack Overflow. Empirical Softw. Engg. 19, 3 (June 2014), 619-654.
- Paul, Michael J. and Roxana Girju. "Topic Modeling of Research Fields: An Interdisciplinary Perspective." RANLP (2009).

- Anton Barua et al. 2014. What are developers talking about? An analysis of topics and trends in Stack Overflow. Empirical Softw. Engg. 19, 3 (June 2014), 619-654.
 - Use latent Dirichlet allocation (LDA), to automatically discover the main topics present in developer discussions.
 - Analyse topics, their relationship and trends over time.

- Gollapalli, Sujatha Das and Xiaoli Li. “EMNLP versus ACL: Analyzing NLP research over time.” EMNLP (2015).
 - Compare trends, topics in two NLP conferences
 - Apply LDA on keyphrases extracted from scientific documents
 - Apply a probabilistic distance metric to calculate the difference of papers in two conferences.

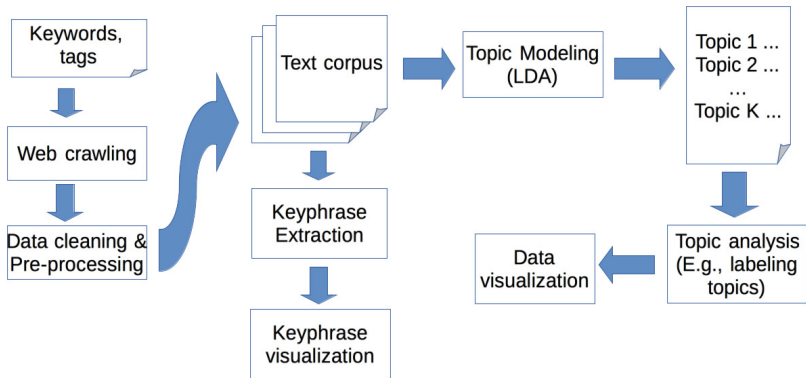
- **Keyphrase extraction**

- Kazi Saidul Hasan and Vincent Ng. 2014. *Automatic keyphrase extraction: A survey of the state of the art*. ACL 2014.
- **ExpandRank**: Xiaojun Wan and Jianguo Xiao. 2008. *Single document keyphrase extraction using neighborhood knowledge*. In AAAI.
- Mihalcea, Rada and Paul Tarau. “TextRank: Bringing Order Into Texts.” (2004).
 - Represent text units (words, sentences, documents,...) as vertexes in a graph.
 - Adapt the Pagerank algorithm for ranking text.
 - Can apply for both undirected and directed graphs
 - Allow using weights of edges in graphs

Table of Contents

- 1 Course information
- 2 Project Description
- 3 Literature Review
- 4 Proposed Solution**
- 5 Project Plan
- 6 Task Assignment
- 7 QA & Discussion

System Architecture



- Apply Latent Dirichlet allocation
 - D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993–1022, January 2003.
- For keyphrase extraction, we apply **TextRank** algorithm
 - Mihalcea, Rada and Paul Tarau. “TextRank: Bringing Order Into Texts.” (2004).

Table of Contents

- 1 Course information
- 2 Project Description
- 3 Literature Review
- 4 Proposed Solution
- 5 Project Plan**
- 6 Task Assignment
- 7 QA & Discussion

#No	Description
1	Choose data sources and crawl text data
2	Clean & process crawled text data
3	Use topic modeling to mine topics in the text corpus
4	Important keyword extraction
5	Interpret mined topics/trends
6	Data visualization
7	Make a website for project demo
8	Prepare presentation for project demo/defence

- Data sources
 - Tech news articles such as Techcrunch website with tags/topics big data, Internet of Things.
 - E.g., <https://techcrunch.com/tag/big-data/>
 - <https://techcrunch.com/tag/iot/>
 - <https://techcrunch.com/topic/subject/internet-of-things>
 - Paper articles
 - International Conference on the Internet of Things:
<http://dblp2.uni-trier.de/db/conf/iot/> (pdf format)
 - Journal of Big data:
<http://journalofbigdata.springeropen.com/> (Open access)
 - Big data analytics:
<http://bdataanalytics.biomedcentral.com/>

- Crawling text data: use **Scrapy** (in Python) for web crawling
- Cleaning and process crawled text data: use `nltk` toolkit
 - `nltk.org`
- Mining topics from text corpus
 - **gensim**:
<https://radimrehurek.com/gensim/index.html>
- Keyphrase extraction
 - Use **TextRank** algorithm
- Data visualization: use `matplotlib`, `seaborn`
- Make website for the project demo
 - Use LAMP: Linux, Apache, MySQL, PHP
 - You can use your favorite web technologies

Table of Contents

- 1 Course information
- 2 Project Description
- 3 Literature Review
- 4 Proposed Solution
- 5 Project Plan
- 6 Task Assignment**
- 7 QA & Discussion

- I have made a tentative assignment on:
`http://bit.ly/2fTotat`

- Install Python & libraries for scientific computing
 - Recommended package: Anaconda
`https://www.continuum.io/downloads`
- Install `nltk` data
 - Large size (≈ 2.5 GB)
 - You can copy from my laptop.
- Install Git tool
- Clone repository of the course
`git clone`
`https://github.com/minhpqn/Text-Mining-Training-Course`
- Get update from repository (I will update more lecture slides and resources)
 - `git pull`

Table of Contents

- 1 Course information
- 2 Project Description
- 3 Literature Review
- 4 Proposed Solution
- 5 Project Plan
- 6 Task Assignment
- 7 QA & Discussion**

Any Questions?



- Practice with nltk toolkit
- Topic modeling (LDA)
- Practice using topic modeling tools for mining topics from text corpora