TEXT TECHNOLOGIES FOR DATA SCIENCE

Course Upgrade

Walid Magdy

Abstract

The course is an upgrade to the "TEXT TECHNOLOGIES FOR DATA SCIENCE" course, where some content to be updated new content to be added.

The course will continue to focus on text technologies, especially information retrieval, as before, however, it will expand to include new technologies such as information filtering, text classification, and applications on domains such as social media. In addition, more emphasis on practical work would be added to allow students experience these technologies in practice.

1 Main Upgrade Items

- 14 lectures → 18 lectures
- 2 assignments → 2 individual assignments + 1 group course project
- 70% exam + 30% coursework → 60% exam + 40% coursework
- 10 credits → 20 credits

2 Course General Information

- Course Level: 11
- Course Points: 20
- Subject area: Informatics
- Programme Collections: Computer Science, Artificial Intelligence, Computational Linguistics.

3 Course Upgrade Details

3.1 Motivation for Upgrade

The current version of Text Technologies for data Science (TTS) course focuses on text information retrieval (IR), especially on the aspect of web search. It mainly teaches the main technologies behind web search engines such as Google, Bing, and Yahoo. The current content is nicely aligned with this purpose. However recent trends in IR and text technologies in general started to include additional technologies and applications of increasing importance beside the web search. Technologies such as information filtering and text classification are grabbing much attention in the recent 6-8 years, especially for application related to social media platforms, such as Facebook and Twitter, that was introduced to the market only 10 years ago. This new trends in IR is not currently covered by the course, which creates a large gap between the learning objectives of the course and the market requirements for text technologies. As an illustration, Figure 1 shows the number of papers published in ACM SIGIR and Springer ECIR, that are considered the main two venues for research in IR, on social media applications. As it is clearly shown, there was a large burst in the number of publications on social media since 2011, which is totally neglected in the current content of the course.

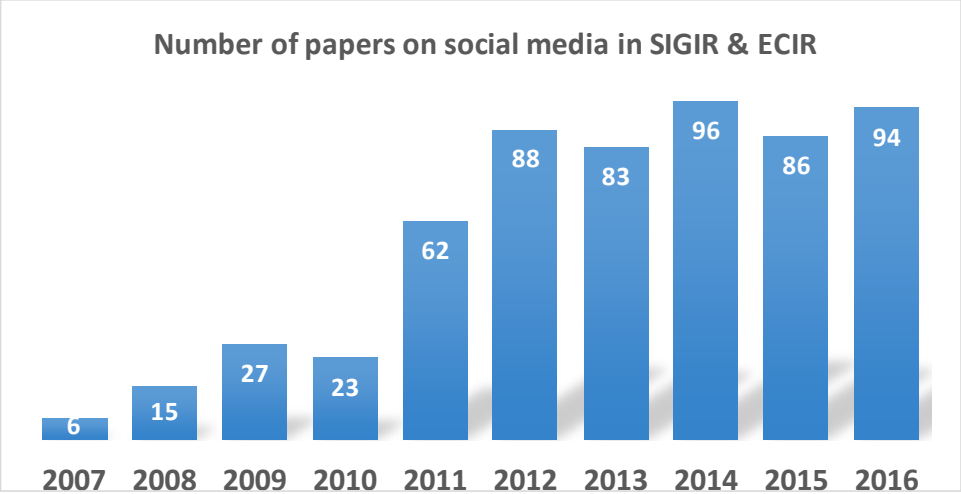

Figure 1: Number of papers published in SIGIR and ECIR on social media between 2007 and 2016

3.2 Suggested Changes

Since the original content of the TTS course still relevant as it teaches the main fundamentals of IR and an important application such as web search, it is suggested to include additional content to the course without the need to remove any of the old content. The suggested new content would cover the following:

1. Information filtering
2. Text classification
3. Social media applications

The added new content would be in two forms

1. Additional lectures on the new content
2. Additional coursework in the form of final course project

The new content will focus on more practical coursework to allow students to develop some applications using the learnt content in the course, which should better qualify the students for the market with additional practical experience.

3.3 Detailed list of Learning Objectives:

On completion of this course, the student should be able to:

1. Describe the main algorithms for processing, storing and retrieving text.
2. Show familiarity with theoretical aspects of IR, including the major retrieval models.
3. Discuss the range of issues involved in building a real search engine
4. Evaluate the effectiveness of a retrieval algorithm
5. Build text classification systems with different types of text features
6. Work in group to develop social media applications using text processing techniques

Objectives 1-4 are not changed from current version of the course. Objectives 5 and 6 are newly added to the upgraded version.

3.4 SYLLABUS

1. Introduction: search applications, search tasks, user's information need
2. Definitions: documents, queries, bag-of-words trick
3. Laws of text: Zipf, Heaps, clumpting, index size. (practical)
4. Vector space: term weighting, similarity functions. (practical)
5. Vocabulary mismatch: tokenization, stemming, synonyms. (practical)
6. Indexing: inverted lists, compression, query execution. (practical)
7. Web crawling: XML feeds, crawling, expected age. (practical)
8. Content Extraction: XML tags, DOM, Finn's method.
9. Locality Sensitive Hashing: duplicates, Simhash. (practical)
10. Evaluation: recall, precision, F1, MAP, nDCG, query logs.
11. Web search: PageRank, hubs and authorities, link spam. (practical)
12. Probabilistic model: probability ranking principle, BM25. (practical)
13. Relevance models: exchangeability, cross-language search.
14. Language models for IR: query likelihood, smoothing.
15. Machine learning in IR: PA, SVM, SMO algorithms, LeToR. (practical)
16. Social media search, nature, challenges, tasks. (practical)
17. Information filtering, topic drift
18. Text classification. (2 practical's)

Lectures 2-15 are the same of the current course. Lecture 1 is added to give a general introduction to the topic of text technologies and IR applications in general for motivating students to the topic by understanding the real-life applications of the technology. Lectures 16-18 are newly added to the course to give other applications of text technologies beside standard search technologies, namely: information filtering, text classification, and social media applications. Lectures that will have practical hours are shown above.

3.5 Readings

- Text books: "Introduction to Information Retrieval", C.D. Manning, P. Raghavan and H. Schutze
- "Search Engines: Information Retrieval in Practice", W. Bruce Croft, Donald Metzler, Trevor Strohman
- Readings: "Machine Learning in Automated Text Categorization". F Sebastiani "The Zipf Mystery",
- YouTube video: https://www.youtube.com/watch?v=fCn8zs912OE
- "Information Retrieval", C.J. van Rijsbergen
- "Recommended Reading for IR Research Students", A. Moffat, J. Zobel, D. Hawking

Readings for the course are not highly changed from the current version. However, the main text book will be Manning et al. instead of Croft et al. In addition, YouTube videos from YouTube channels that simplify science to ordinary people would be recommended to students to learn about science related to the course in a more fun way.

3.6 Coursework

One of the main changes in the upgraded version of the course will be in the coursework. Coursework will jump from 25 hours to 70 hours in a step to get the students experience what they learn in practical assignments.

Assignments of the course would be as follow:

1. Assignment 1:
   Coursework hours: 15
   Award (mark): 10%
   Time: After Lecture 6 (week 3-4)
   Delivery: within 3 weeks
   Nature: Individual
   Objective: Each student would work on building an inverted index for a small size collection to emphasise on understanding one of the main fundamental of IR, which is indexing.
2. Assignment 2:
   Coursework hours: 15
   Award (mark): 10%
   Time: After Lecture 12 (week 6-7)
   Delivery: within 3 weeks
   Nature: Individual
   Objective: Students will be asked to implement some of the studied retrieval models to compare their effectiveness on text retrieval.
3. Course project:
   Coursework hours: 40
   Award (mark): 20%
   Time: After Lecture 16 (week 8-9)
   Delivery: before the exams
   Nature: groups of 2-5 students
   Objective: Students will form groups of 2-5 to work on implementing a retrieval or classification system of their choice using the technologies they learnt during the course. A report and presentation to be delivered for each project. Marks will be on three criteria:
   1. Project soundness and excellence.
   2. Teamwork and organization
   3. Contribution of each individual (varies for each individual)

The objective of the final course project is to put students in a work environment similar to that they will experience after their graduation.

4 Allocations

4.1 Hours Allocation

Learning Activity	Current	Proposed	Change
Lectures	14	18	+4
Supervised Practical Hours	8	12	+4
Coursework	25	70	+45
Directed and Independent Learning	53	100	+47
Total	100	200	+100

4.2 Award

Assessment	Current	Proposed	Change	Assessment to
Written Exam	70%	60%	+10%	Understanding to fundamentals of text technologies and IR
Coursework	30%	40%	-10%	Depth of understanding to basics when applied to practical problems