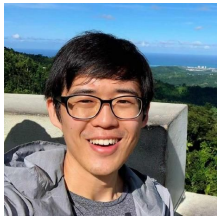
From Fundamentals to Recent Advances A Tutorial on Keyphrasification

Part 1.1 Introduction

Rui Meng, Debanjan Mahata, Florian Boudin

ECIR 2022

Presenters



Rui Meng

Research Scientist
Salesforce Research

Debanjan Mahata

Director of ML, Moody's Analytics

Adjunct faculty at IIIT-Delhi

Florian Boudin

Associate Professor
University of Nantes

We would like to thank all our colleagues, students and friends who have contributed to this tutorial.

What is this tutorial about?

- A walkthrough of the advancement of keyphrasification
  - A field of study since early 90s
  - Experiencing a renaissance by deep learning
- Comprehensive overview of methods for keyphrasification
  - Classic and SOTA models
  - Advanced topics: recent progress and applications
- Introducing PKE/DLKP, two toolkits dedicated to KP studies/applications
  - Hands-on practice with colab examples

All materials available at
https://keyphrasification.github.io/

Outline

Part I - Introduction and Classic Methods

Part II - Modern Neural Methods

Part III - Advanced Topics

All materials available at
https://keyphrasification.github.io/

Outline

Part I - Introduction and Classic Methods
- ➔ Definition, datasets and evaluation, classic methods

Part II - Modern Neural Methods
- ➔ Neural extraction and generation methods

Part III - Advanced Topics
- ➔ Keyphrase Generation for IR
- ➔ Data-efficient Domain Adaptation for Keyphrase Generation
- ➔ Learning Better Keyphrase Representations

All materials available at
https://keyphrasification.github.io/

Part I

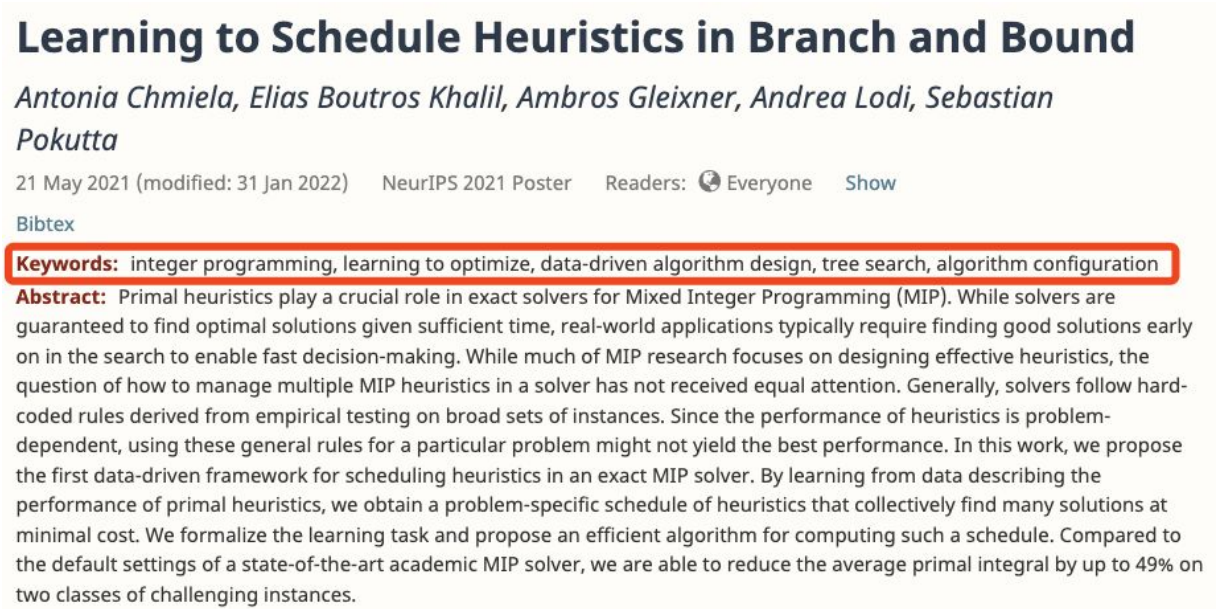
Introduction to Keyphrasification and Classic Methods

Part I - Outline

- **Introduction to keyphrasification (Rui)**
  - **Definitions and applications**
- Datasets and evaluations (Debanjan)
- Traditional methods for keyphrase extraction (Florian)
- Hands-on practice with PKE

What are keyphrases ?

- A set of phrases to capture the essence of a document



Learning to Schedule Heuristics in Branch and Bound

Antonia Chmiela, Elias Boutros Khalil, Ambros Gleixner, Andrea Lodi, Sebastian Pokutta

21 May 2021 (modified: 31 Jan 2022) NeurIPS 2021 Poster Readers: Everyone Show

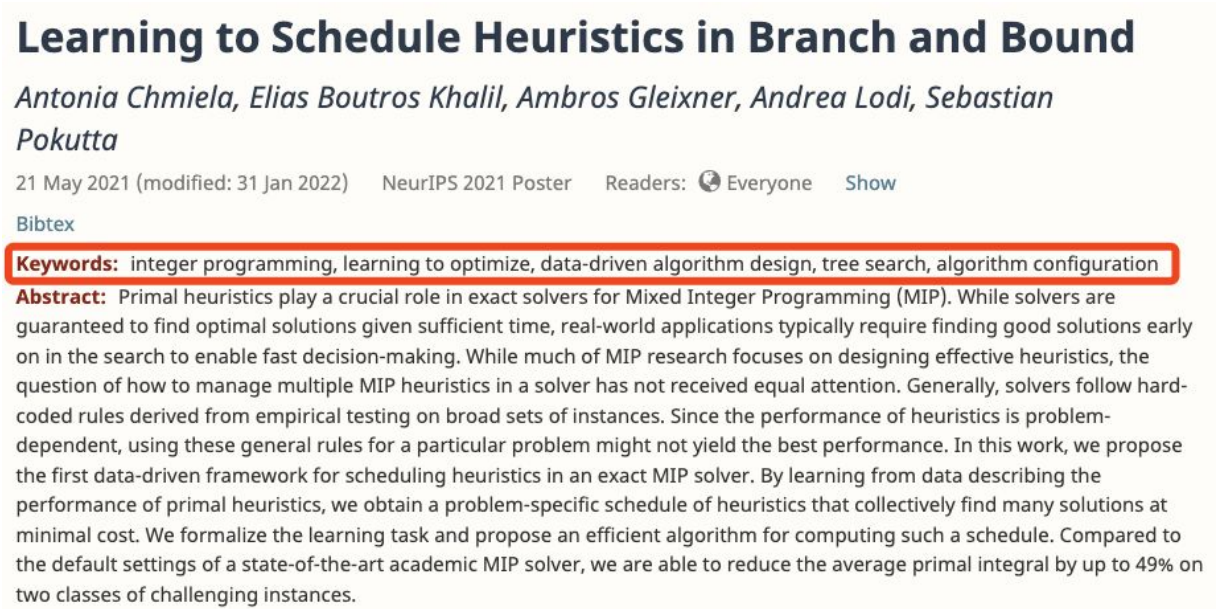
Bibtex

Keywords: integer programming, learning to optimize, data-driven algorithm design, tree search, algorithm configuration

Abstract: Primal heuristics play a crucial role in exact solvers for Mixed Integer Programming (MIP). While solvers are guaranteed to find optimal solutions given sufficient time, real-world applications typically require finding good solutions early on in the search to enable fast decision-making. While much of MIP research focuses on designing effective heuristics, the question of how to manage multiple MIP heuristics in a solver has not received equal attention. Generally, solvers follow hard-coded rules derived from empirical testing on broad sets of instances. Since the performance of heuristics is problem-dependent, using these general rules for a particular problem might not yield the best performance. In this work, we propose the first data-driven framework for scheduling heuristics in an exact MIP solver. By learning from data describing the performance of primal heuristics, we obtain a problem-specific schedule of heuristics that collectively find many solutions at minimal cost. We formalize the learning task and propose an efficient algorithm for computing such a schedule. Compared to the default settings of a state-of-the-art academic MIP solver, we are able to reduce the average primal integral by up to 49% on two classes of challenging instances.

What are keyphrases ?

- A set of phrases to capture the essence of a document
- A set of phrases that summarizes the content of data



Learning to Schedule Heuristics in Branch and Bound

Antonia Chmiela, Elias Boutros Khalil, Ambros Gleixner, Andrea Lodi, Sebastian Pokutta

21 May 2021 (modified: 31 Jan 2022) NeurIPS 2021 Poster Readers: Everyone Show

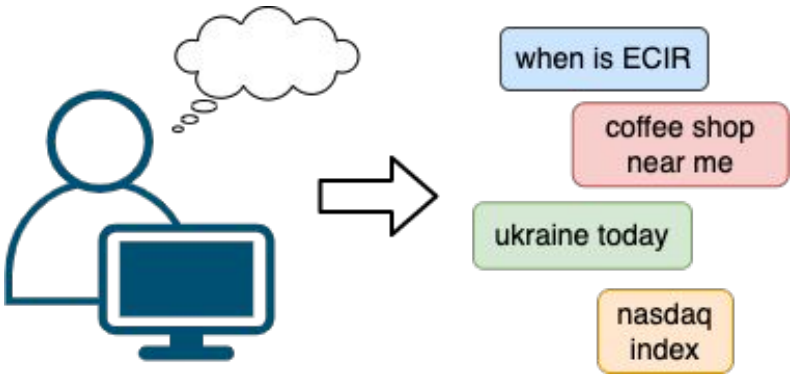
Bibtex

Keywords: integer programming, learning to optimize, data-driven algorithm design, tree search, algorithm configuration

Abstract: Primal heuristics play a crucial role in exact solvers for Mixed Integer Programming (MIP). While solvers are guaranteed to find optimal solutions given sufficient time, real-world applications typically require finding good solutions early on in the search to enable fast decision-making. While much of MIP research focuses on designing effective heuristics, the question of how to manage multiple MIP heuristics in a solver has not received equal attention. Generally, solvers follow hard-coded rules derived from empirical testing on broad sets of instances. Since the performance of heuristics is problem-dependent, using these general rules for a particular problem might not yield the best performance. In this work, we propose the first data-driven framework for scheduling heuristics in an exact MIP solver. By learning from data describing the performance of primal heuristics, we obtain a problem-specific schedule of heuristics that collectively find many solutions at minimal cost. We formalize the learning task and propose an efficient algorithm for computing such a schedule. Compared to the default settings of a state-of-the-art academic MIP solver, we are able to reduce the average primal integral by up to 49% on two classes of challenging instances.

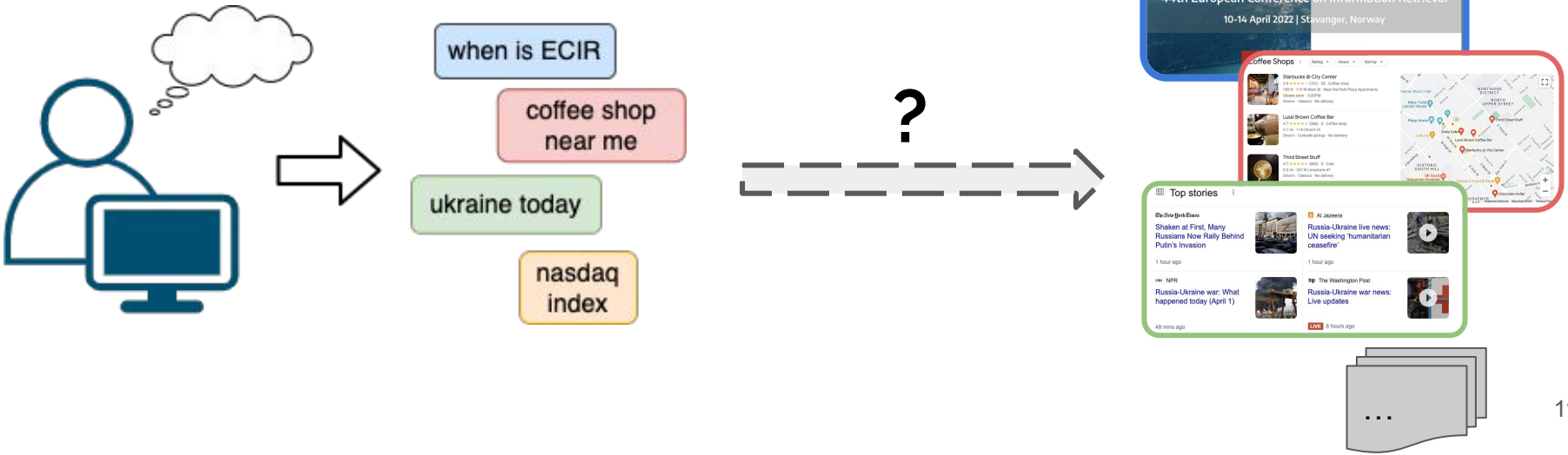
Why keyphrases?

- For humans, phrases are a natural and neat way to express their needs
  - More expressive/accurate than words
  - More efficient than sentences
    - Principle of least effort



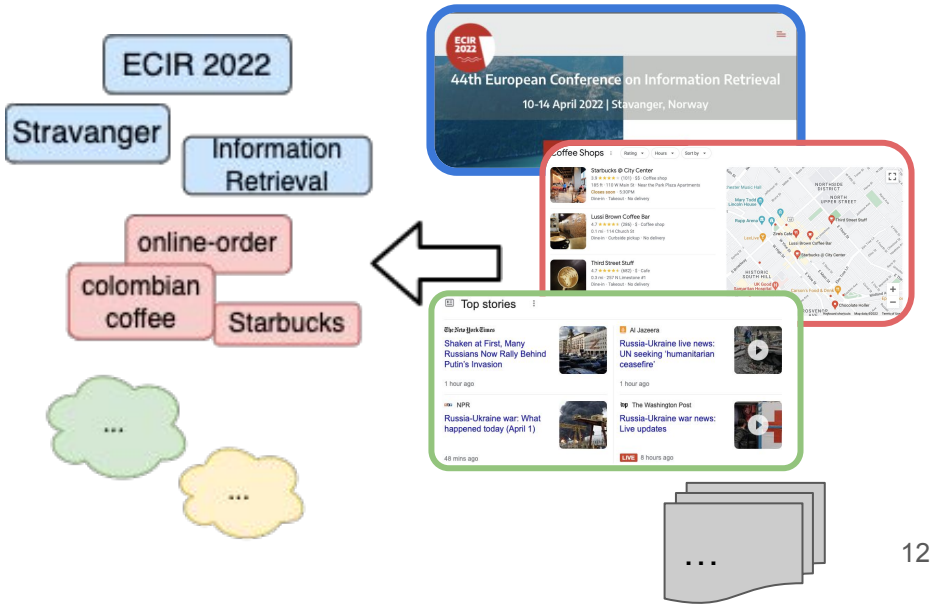
Why keyphrases?

- For humans, phrases are a natural and neat way to express their needs
  - More expressive/accurate than bag-of-words
  - More efficient than sentences
    - Principle of least effort



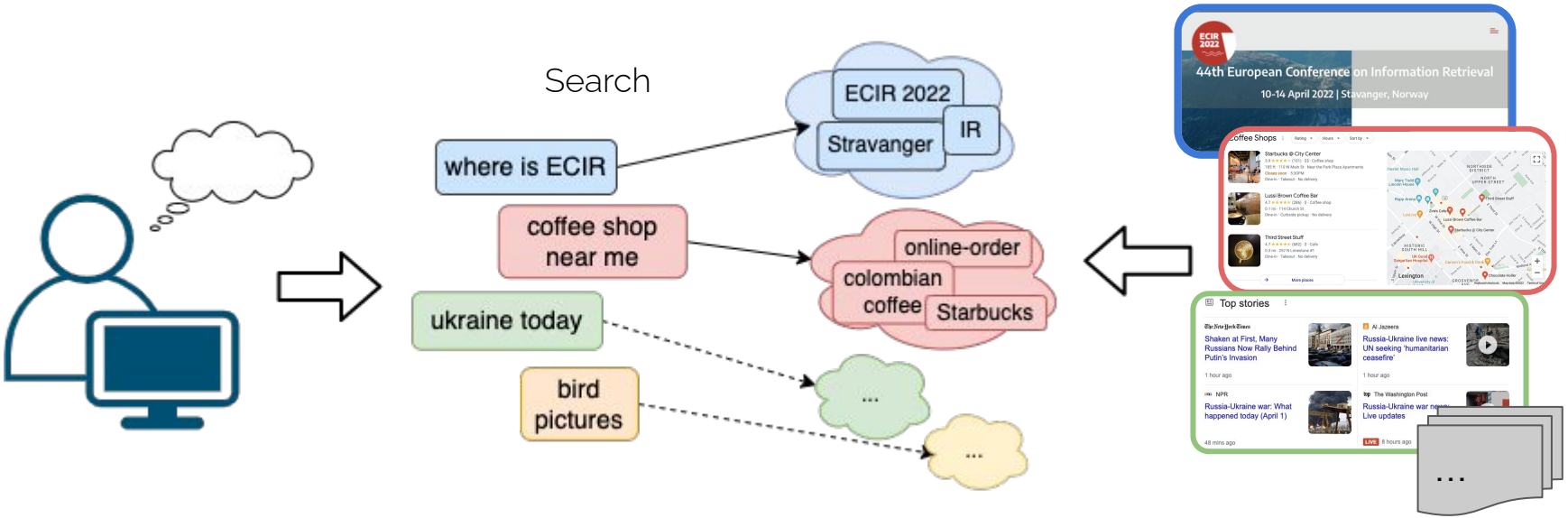
Why keyphrases?

- Data can often be described as understandable short symbols
  - Not vectors!
  - Tags, labels, categories, or keyphrases



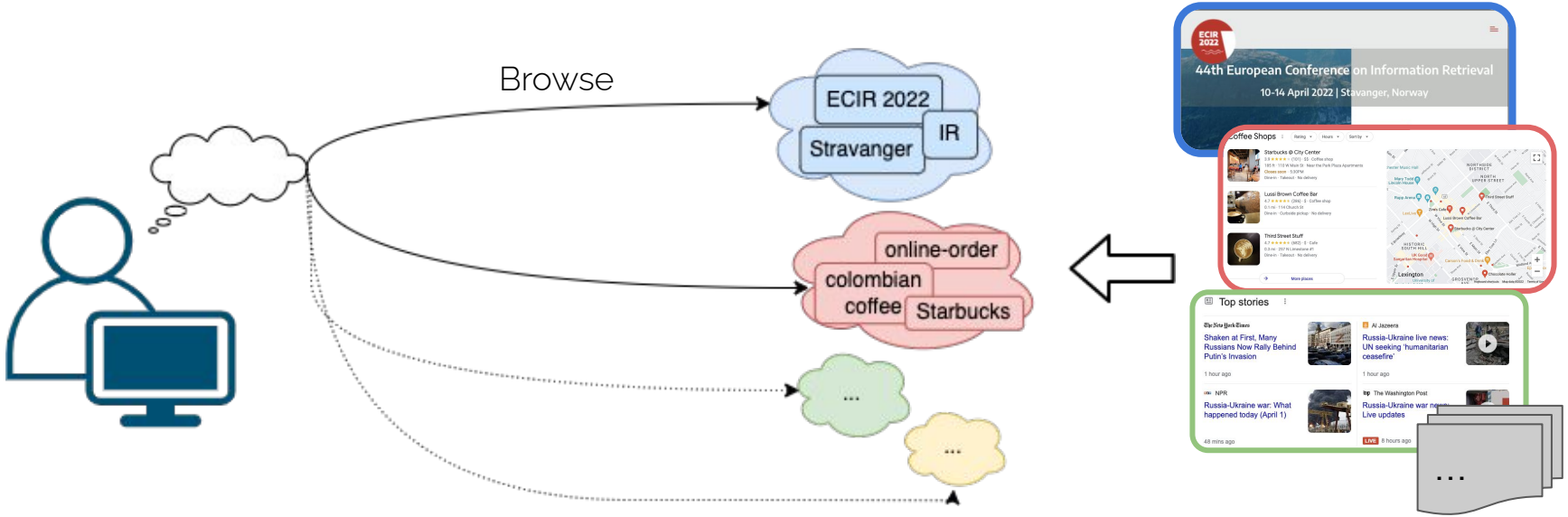
Why keyphrases?

- Keyphrase is a natural & efficient API connecting human and data
  - We have only limited time/capacity to process unlimited data
  - Especially for huge and obscure data



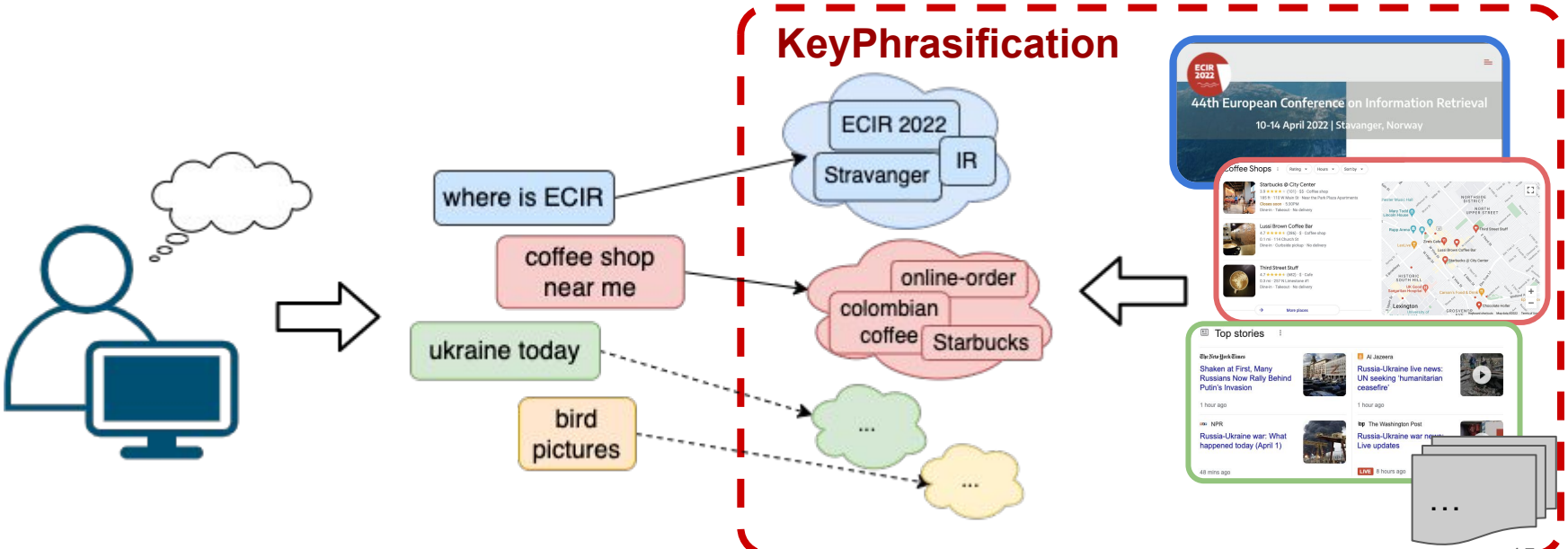
Why keyphrases?

- Keyphrase is a natural & efficient API connecting human and data
  - We have only limited time/capacity to process unlimited data
  - Especially for huge and obscure data



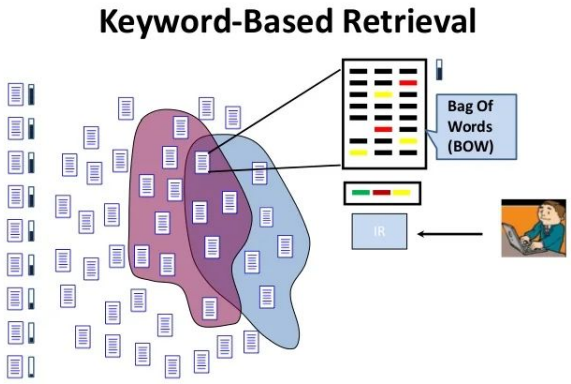
Why keyphrases?

- Keyphrase is a natural & efficient API connecting human and data
  - We have only limited time/capacity to process unlimited data
  - Especially for huge and obscure data



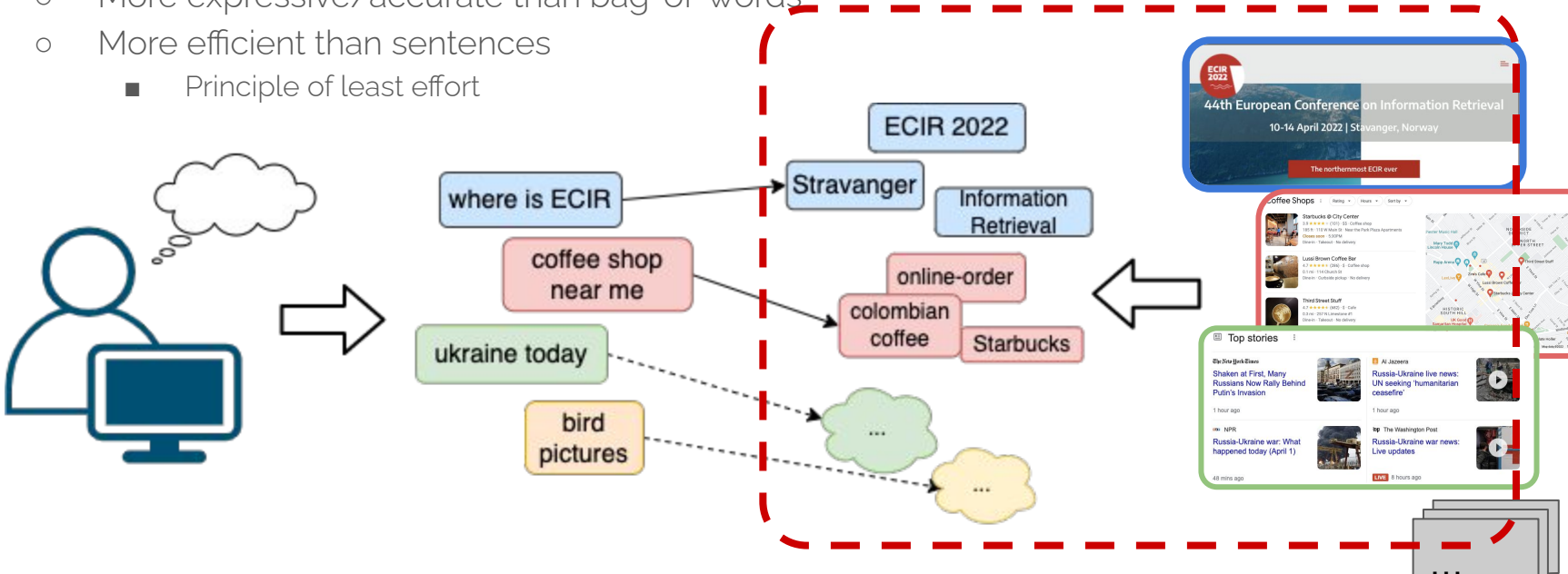
Why keyphrases?

- Keyphrases in real-world scenarios
  - Search
    - Use keyphrases (queries) to locate relevant information
  - Browsing
    - Explore data, navigated by different keyphrases (tags)



Why keyphrases?

- Why key is important?
  - We have only limited time/capacity to process unlimited data
- Why phrase is important?
  - More expressive/accurate than bag-of-words
  - More efficient than sentences
    - Principle of least effort



What is Keyphrasification?

- The general task/process for summarizing data with keyphrases
  - **Key-***

    adj. indicating crucial importance

    -> keyness: importance to a document
  - **Phrasify**

    v. to utter or express in a phrase (phraseness)

    -> phraseness: how phrase-like a short text is
- It can be instantiated in different ways
  - **Keyphrase extraction/generation (texts)**
  - Automatic tagging/labeling (fixed-size target phrase vocab)
  - Object detection/image classification (vision)
  - Etc.

Keyphrase Extraction & Generation

- **Keyphrase Extraction**
  - Extract important phrases from text
  - KPE can only predict phrases that appear in the text
- **Keyphrase Generation**
  - Predict keyphrases by text generation
  - Output may contain low-quality phrases

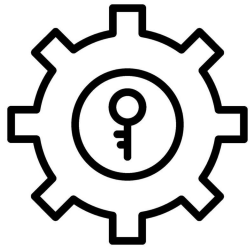
Keyphrasification

Domain-specific keyphrase extraction (1999)

Abstract

Keyphrases are an important means of document summarization, clustering, and topic search. Only a small minority of documents have author-assigned keyphrases, and manually assigning keyphrases to existing documents is very laborious. Therefore it is highly desirable to automate the keyphrase extraction process. This paper shows that a simple procedure for keyphrase extraction based on the naive Bayes learning scheme performs comparably to the state of the art. It goes on to explain how this procedure's performance can be boosted by automatically tailoring the extraction process to the particular document collection at hand. Results on a large collection of technical reports in computer science show that the quality of the extracted keyphrases improves significantly when domain-specific information is exploited.

Input



Output

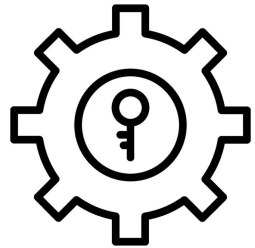
Keyphrases

domain-specific keyphrase extraction · large collection · document summarization · topic search · computer science · show · keyphrase extraction process · naive bayes · extraction process · scheme performs · technical report · keyphrase extraction · important mean · small minority · simple procedure · author-assigned keyphrases · domain-specific information · extracted keyphrases · particular document collection

Source: https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.148.3598

Keyphrasification

Input
Long text



Output
Keyphrases

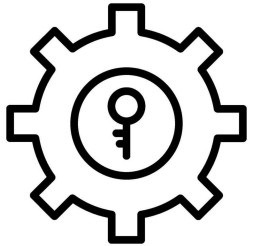
Keyphrasification

Muslim Women in Hijab Break Barriers: 'Take the Good With the Bad'

When Ginella Massa, a Toronto-based TV reporter, recently accepted a request to host an evening newscast, she was not planning or expecting to make history for wearing a hijab. She was just covering for a colleague who wanted to go to a hockey game. And that's how Ms. Massa, who works at CityNews in Toronto, became the first Canadian woman to host a newscast from a large media company while wearing the head scarf. [...] This new trend of inclusion occurs amid a more sinister one, as reported hate crimes against Muslims are on the rise in the United States and Canada. The F.B.I. says that a surge in hate crimes against Muslims has led to an overall increase in hate crimes in the United States; Muslims have borne the brunt of the increase with 257 recorded attacks. [...] In Canada, where Ms. Massa has lived since she was a year old, the number of reported hate crimes has dropped slightly overall, but the number of recorded attacks against Muslims has grown: 99 attacks were reported in 2014, according to an analysis by the news site Global News of data from Statistics Canada, a government agency. [...]

keywords: US; Islam; Fashion; Muslim Veiling; Women and Girls; (News media, journalism); Hate crime; Canada

Input



Output

Gallina, Y., Boudin, F., & Daille, B. (2019). KPTimes: A large-scale dataset for keyphrase generation on news documents. *arXiv preprint arXiv:1911.12559*.

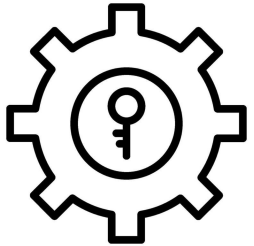
Present Keyphrases (Extractable)

Muslim Women in Hijab Break Barriers: 'Take the Good With the Bad'

When Ginella Massa, a Toronto-based TV reporter, recently accepted a request to host an evening newscast, she was not planning or expecting to make history for wearing a hijab. She was just covering for a colleague who wanted to go to a hockey game. And that's how Ms. Massa, who works at CityNews in Toronto, became the first Canadian woman to host a newscast from a large media company while wearing the head scarf. [...] This new trend of inclusion occurs amid a more sinister one, as reported hate crimes against Muslims are on the rise in the United States and Canada. The F.B.I. says that a surge in hate crimes against Muslims has led to an overall increase in hate crimes in the United States; Muslims have borne the brunt of the increase with 257 recorded attacks. [...] In Canada, where Ms. Massa has lived since she was a year old, the number of reported hate crimes has dropped slightly overall, but the number of recorded attacks against Muslims has grown: 99 attacks were reported in 2014, according to an analysis by the news site Global News of data from Statistics Canada, a government agency. [...]

keywords: US; Islam; Fashion; Muslim; Veiling; Women and Girls; (News media journalism); Hate crime; Canada

Input



Output

Keyphrases (or part of) appearing in the document are colored

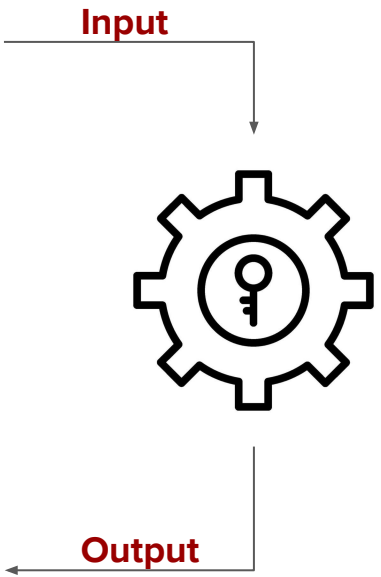
Gallina, Y., Boudin, F., & Daille, B. (2019). KPTimes: A large-scale dataset for keyphrase generation on news documents. *arXiv preprint arXiv:1911.12559*.

Absent Keyphrases (Not Extractable)

Muslim Women in Hijab Break Barriers: 'Take the Good With the Bad'

When Ginella Massa, a Toronto-based TV reporter, recently accepted a request to host an evening newscast, she was not planning or expecting to make history for wearing a hijab. She was just covering for a colleague who wanted to go to a hockey game. And that's how Ms. Massa, who works at CityNews in Toronto, became the first Canadian woman to host a newscast from a large media company while wearing the head scarf. [...] This new trend of inclusion occurs amid a more sinister one, as reported hate crimes against Muslims are on the rise in the United States and Canada. The F.B.I. says that a surge in hate crimes against Muslims has led to an overall increase in hate crimes in the United States; Muslims have borne the brunt of the increase with 257 recorded attacks. [...] In Canada, where Ms. Massa has lived since she was a year old, the number of reported hate crimes has dropped slightly overall, but the number of recorded attacks against Muslims has grown: 99 attacks were reported in 2014, according to an analysis by the news site Global News of data from Statistics Canada, a government agency. [...]

keywords: US; Islam; Fashion; Muslim Veiling; Women and Girls; (News media, journalism); Hate crime; Canada



Keyphrases (or part of) appearing in the document are colored

Gallina, Y., Boudin, F., & Daille, B. (2019). KPTimes: A large-scale dataset for keyphrase generation on news documents. *arXiv preprint arXiv:1911.12559*.

Applications

- Because keyphrases distill the important information from documents, they are useful for many applications in NLP and IR
  - Information retrieval/indexing (Jones and Staveley, 1999; Gutwin et al., 1999)
    - Interactive information retrieval
    - Query expansion
  - Document Analysis
    - Summarization (Zha, 2002; Wan et al., 2007)
    - Text categorization/Clustering (Hulth and Megyesi, 2006)
    - Sentiment analysis (Berend, 2011)
  - Reading Enhancement

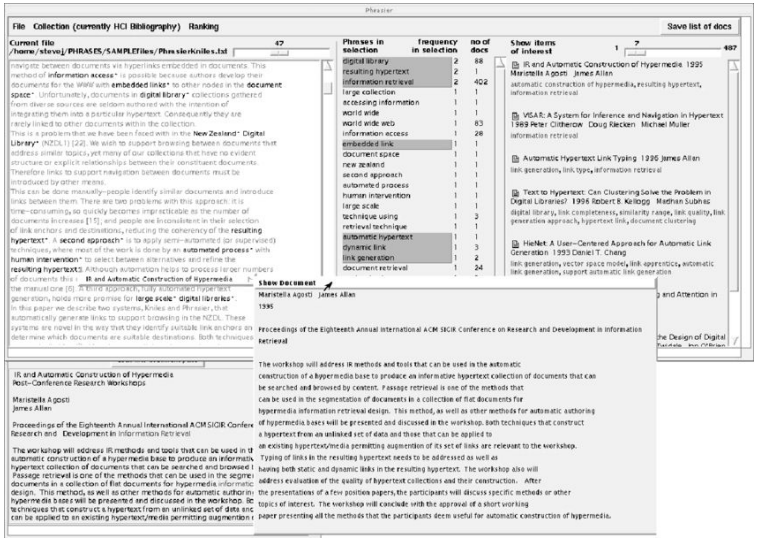
(Gutwin et al., 1999) C. Gutwin, G. Paynter, I. Witten, C. Nevill-Manning and E. Frank, Improving browsing in digital libraries with keyphrase indexes, Decision Support Systems.
(Jones and Staveley, 1999) S. Jones and M. S. Staveley, Phrasier: a system for interactive document retrieval using keyphrases, SIGIR.
(Zha, 2002) H. Zha, Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering, SIGIR.
(Wan et al., 2007) X. Wan, J. Yang, J. Xiao, Towards an Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction, ACL.
(Hulth and Megyesi, 2006) A. Hulth and B. Megyesi, A study on automatically extracted keywords in text categorization, ACL.
(Berend, 2011) G. Berend, Opinion Expression Mining by Exploiting Keyphrase Extraction, IJCNLP.
(Collins and Beel, 2019) A. Collins and J. Beel, Document embeddings vs. keyphrases vs. terms for recommender systems: A large-scale online evaluation, JCDL.

Applications

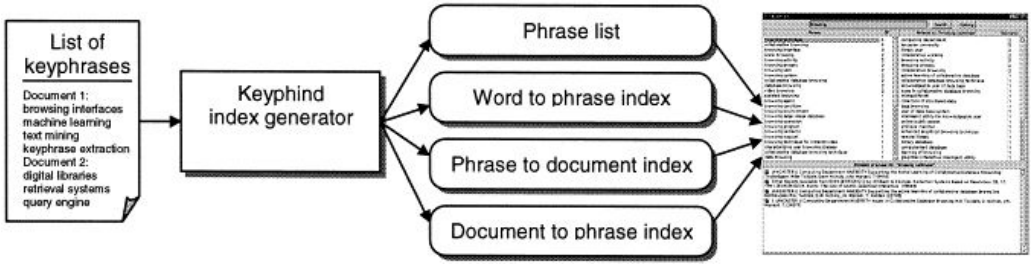
- Because **keyphrases distill the important information** from documents, they are useful for many applications in NLP and IR
  - Document retrieval (Jones and Staveley, 1999; Gutwin et al., 1999)
  - Summarization (Zha, 2002; Wan et al., 2007)
  - Text categorization (Hulth and Megyesi, 2006)
  - Sentiment analysis (Berend, 2011)
  - Paper recommendation (Collins and Beel, 2019)

(Gutwin et al., 1999) C. Gutwin, G. Paynter, I. Witten, C. Nevill-Manning and E. Frank, Improving browsing in digital libraries with keyphrase indexes, Decision Support Systems.
(Jones and Staveley, 1999) S. Jones and M. S. Staveley, Phrasier: a system for interactive document retrieval using keyphrases, SIGIR.
(Zha, 2002) H. Zha, Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering, SIGIR.
(Wan et al., 2007) X. Wan, J. Yang, J. Xiao, Towards an Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction, ACL.
(Hulth and Megyesi, 2006) A. Hulth and B. Megyesi, A study on automatically extracted keywords in text categorization, ACL.
(Berend, 2011) G. Berend, Opinion Expression Mining by Exploiting Keyphrase Extraction, IJCNLP.
(Collins and Beel, 2019) A. Collins and J. Beel, Document embeddings vs. keyphrases vs. terms for recommender systems: A large-scale online evaluation, JCDL.

Applications - Interactive Document Retrieval



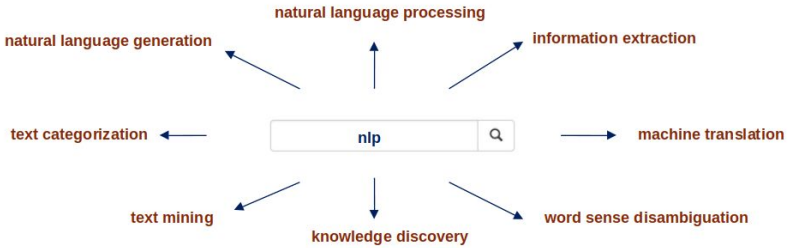
Jones, S., & Staveley, M. S. (1999, August). Phrasier: a system for interactive document retrieval using keyphrases. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 160-167).



Gutwin, C., Paynter, G., Witten, I., Nevill-Manning, C., & Frank, E. (1999). Improving browsing in digital libraries with keyphrase indexes. *Decision Support Systems*, *27*(1-2), 81-104.

Applications - Query Expansion

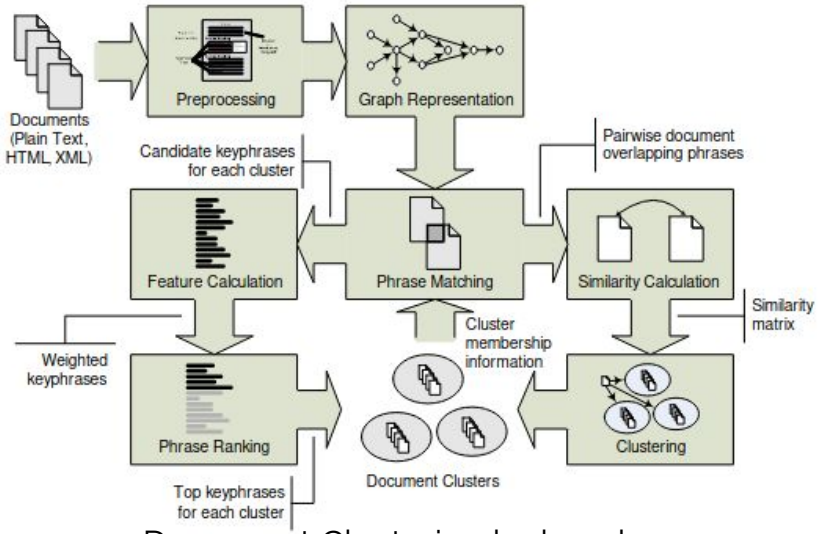
```
<keyphrases id="350">
    <keyphrase weight="0.39282" category="2" >
        computer screen </keyphrase>
    <keyphrase weight="0.38114" category="1" >
        occupational health </keyphrase>
    <keyphrase weight="0.38566" category="1" >
        workplace disorders </keyphrase>
    <keyphrase weight="0.38432" category="1" >
        physical injury</keyphrase>
    <keyphrase weight="0.38427" category="1" >
        computer terminal </keyphrase>
    <keyphrase weight="0.38320" category="2" >
        workers computer </keyphrase>
   <keyphrase weight="0.38293" category="4">
        report </keyphrase>
    <keyphrase weight="0.38174" category="1" >
        terminals activity </keyphrase>
</keyphrases>
```



Song, I. Y., Allen, R. B., Obradovic, Z., & Song, M. (2006, June). Keyphrase extraction-based query expansion in digital libraries. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'06)* (pp. 202-209). IEEE.

Applications - Document Analysis

Visualization by
Keyphrase WordCloud



Document Clustering by keyphrases

(Chuang et al., 2012). Without the clutter of unimportant words" Descriptive keyphrases for text visualization. TOCHI.
(Hammouda et al., 2005). Corephrase: Keyphrase extraction for document clustering. MLDM..

Applications - Document Understanding

Taliban attacks German consulate in northern Afghan city of Mazar-i-Sharif with truck bomb

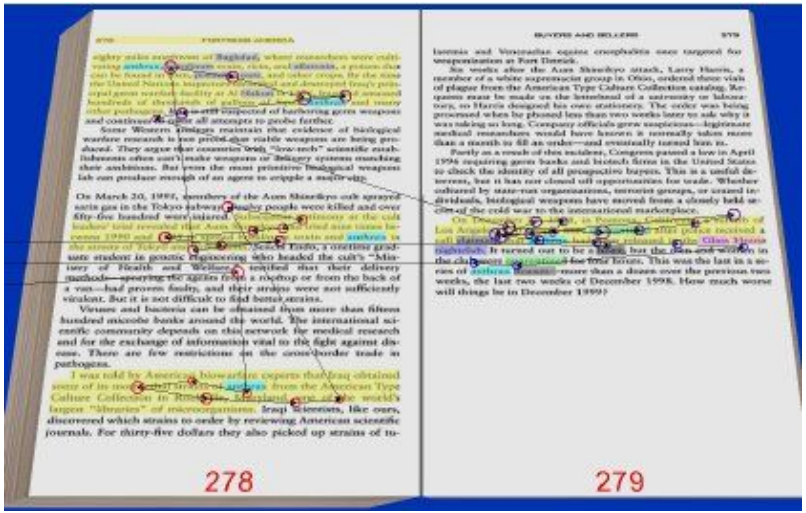
The death toll from a powerful Taliban truck bombing at the German consulate in Afghanistan's Mazar-i-Sharif city rose to at least six Friday, with more than 100 others wounded in a major militant assault.

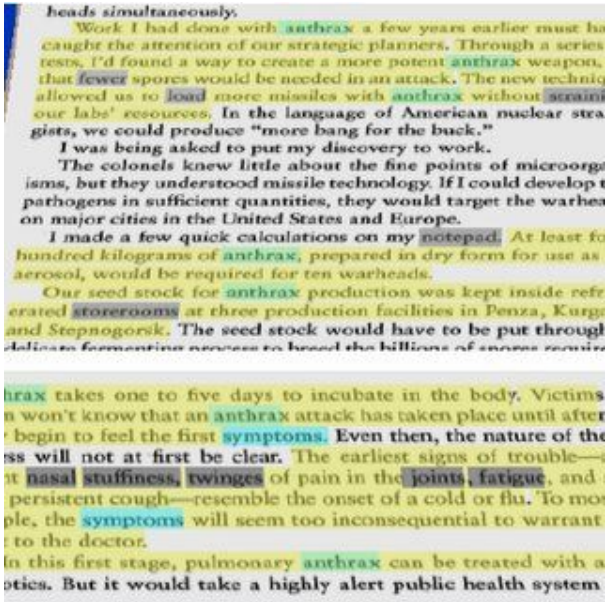
The Taliban said the bombing late Thursday, which tore a massive crater in the road and overturned cars, was a "**revenge attack**" for US air strikes this month in the volatile province of Kunduz that left 32 civilians dead. […]

Who did What When Where Why and How

Hamborg, F., Breitinger, C., & Gipp, B. (2019). Giveme5w1h: A universal system for extracting main events from news articles. *arXiv preprint arXiv:1909.02766*.

Applications - Reading Enhancement





Chi, Ed H., Michelle Gumbrecht, and Lichan Hong. "Visual foraging of highlighted text: An eye-tracking study." *International Conference on Human-Computer Interaction*. Springer, Berlin, Heidelberg, 2007.

Keyphrase Application at Bloomberg

https://ecir2022.org/

The Impact and Importance of Keyphrases in Building NLP Products

Mayank Kulkarni,
Bloomberg, USA

Part I - Outline

- Introduction to keyphrasification
  - Definitions and applications
- **Datasets and evaluations (Debanjan)**
- Traditional methods for keyphrase extraction (Florian)
- Hands-on practice with PKE

All materials available at
https://keyphrasification.github.io/