Digging in the Dirt:
Extracting Keyphrases from Texts with KD

Giovanni Moretti, Rachele Sprugnoli, Sara Tonelli
Digital Humanities Research Unit
Fondazione Bruno Kessler

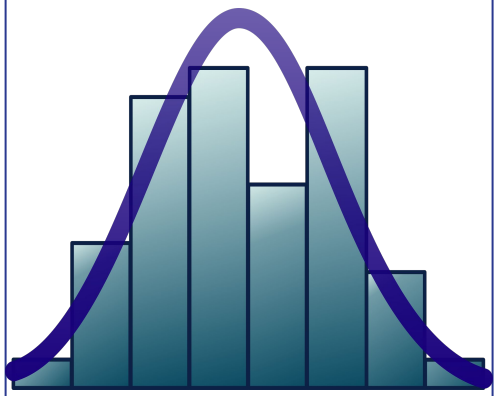
Website: http://dh.fbk.eu
Twitter: @DH_FBK

FONDAZIONE BRUNO KESSLER

DH DIGITAL HUMANITIES

KD = Keyphrase Digger

STATISTICAL MEASURES



LINGUISTIC INFORMATION

WEIGHTED LIST OF KEYPHRASES

rank	keyword synonyms	frequency	score
1	kd	33	43.07
2	types of texts	2	23.24
3	kx	17	22.36
4	multi-token expressions	4	14.04
5	level of customizability	2	12.64
6	keyphrase extraction	5	9.55
7	humanities scholars	2	8.31
8	document collections	4	7.08
9	configuration	4	6.87
10	set of key-concepts	2	6.63

NGRAMS capturing the main concepts of a document

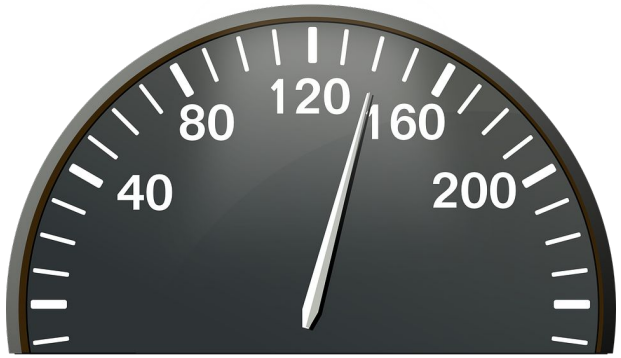
Reimplementation of KX (Pianta & Tonelli, 2010)

Challenge 1

Better extraction performances

Challenge 2

Higher processing speed



Challenge 3

More customizability

Challenge 1

Extraction Performances

- SEMEVAL 2010 dataset
  - scientific papers
  - best performing rule-based system

KX: 7th place → KD: 2nd place

- DE GASPERI corpus
  - historical documents

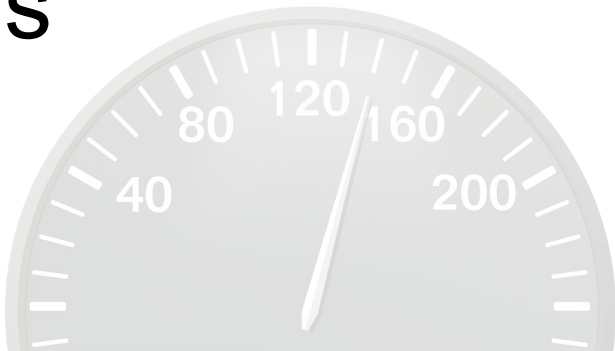
Precision over expert keyphrases: 42%

Challenge 2

Processing Speed

Same machine, same corpus (101K tokens), comparable settings

- KX: 3.4 minutes
- KD: 7 seconds



Challenge 3

Customizability

- Accepts different pre-processed texts
  e.g. TEXTPRO - TREETAGGER - STANFORD
- Uses custom configuration files
  e.g. POS-PATTERNS - STOPWORDS
- Many options available
  e.g. NUMBER of KEYPHRASES - NGRAM LENGTH
- Easily extendible to other languages

Applications and Availability

1. ALCIDE PLATFORM

   Analysis of Language and Content In a Digital Environment

2. ON-LINE DEMO

   http://celct.fbk.eu:8080/KD_KeyDigger/

3. SOFTWARE PACKAGE

   http://dh.fbk.eu/technologies/kd

SEE YOU AT THE POSTER SESSION!

FONDAZIONE BRUNO KESSLER trentino italy **www.fbk.eu**

Digging in the Dirt: Extracting Keyphrases from Texts with KD

CENTER FOR INFORMATION TECHNOLOGY
DIGITAL HUMANITIES

FONDAZIONE BRUNO KESSLER

KD = KEYPHRASE DIGGER

Keyphrase Digger is a rule-based system for keyphrase extraction. It is a **Java** re-implementation of KX tool (Pianta and Tonelli, 2010) with a **new architecture** and **new features**.

It combines **statistical measures** with **linguistic information** given by PoS patterns to identify and extract weighted keyphrases from texts.

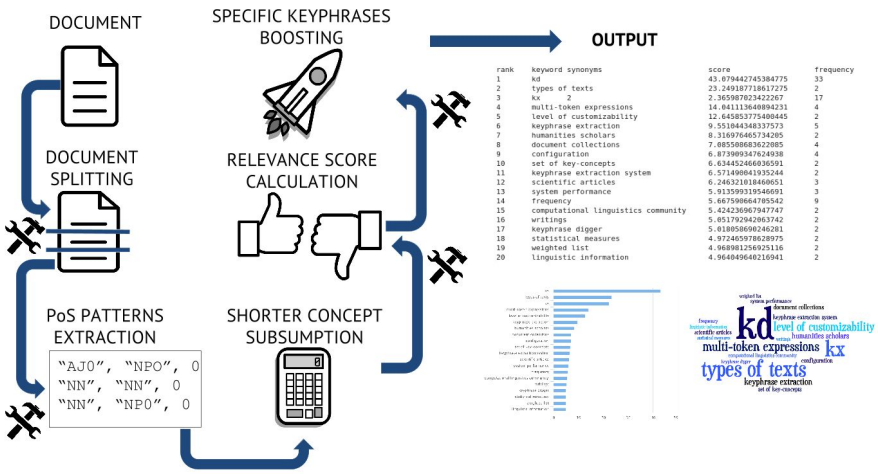
MAIN FEATURES:

- Extraction of multi-words
- Multilinguality (EN and IT)
- Easily extendible to other languages
- Higher customizability than KX
- High processing speed
- Clustering of keyphrases under the same lemma
- Various accepted formats and POS tagsets
- Boost of specific PoS patterns
- Integration of Apache Lucene Library

NEW!

KD is about 25 times faster than KX: very apt for web applications!

KD ARCHITECTURE



rank	keyword synonyms	score	frequency
1	kd	43.079442745384775	33
2	types of texts	23.249187718617275	2
3	kx 2	2.365987023422267	17
4	multi-token expressions	14.041113640894231	4
5	level of customizability	12.645853775400445	2
6	keyphrase extraction	9.551044348337573	5
7	humanities scholars	8.316976465734205	2
8	document collections	7.885508036222885	4
9	configuration	6.873909347624938	4
10	set of key-concepts	6.634452466036591	2
11	keyphrase extraction system	6.571490841935244	2
12	scientific articles	6.246321018460651	3
13	system performance	5.913599319546691	3
14	frequency	5.667590664705542	9
15	computational linguistics community	5.424236967947747	2
16	writings	5.051729420063742	2
17	keyphrase digger	5.018058690246281	2
18	statistical measures	4.972465978628975	2
19	weighted list	4.968981256925116	2
20	linguistic information	4.964049640216941	2

KD AVAILABILITY

ONLINE DEMO
http://celct.fbk.eu:8080/KD_KeyDigger/

JAVA RUNNABLE JAR and LIBRARY
https://dh.fbk.eu/technologies/kd

http://dh.fbk.eu
https://twitter.com/DH_FBK

moretti@fbk.eu
sprugnoli@fbk.eu
satonelli@fbk.eu

DH DIGITAL HUMANITIES