European Network for Academic Integrity

Plagiarism Detection: Techniques, Tools and Policies

Tomáš Foltýnek

Mendel University in Brno, CZ :: University of Konstanz, DE

tomas.foltynek@academicintegrity.eu

Supported by the Erasmus+ Strategic Partnerships project 2016-1-CZ01-KA203-023949.

Introduction

European Network for Academic Integrity

- Literature review on Plagiarism detection
  - Overview of methods
  - Strengths, weaknesses
- Do PDS vendors learn from scientists?
- Plagiarism policies

Methodology

European Network for Academic Integrity

- Keyword-based automated search
- Google Scholar
  - restricted time 2013 – 2018
  - manual review of titles
  - manual review of abstracts
  - exclusion of papers in journals on the Beall's list
- Topically related conferences
  - Plagiarism across Europe and Beyond 2013, 2015, 2017
  - PAN competitions
- Sort to categories, identify research gaps

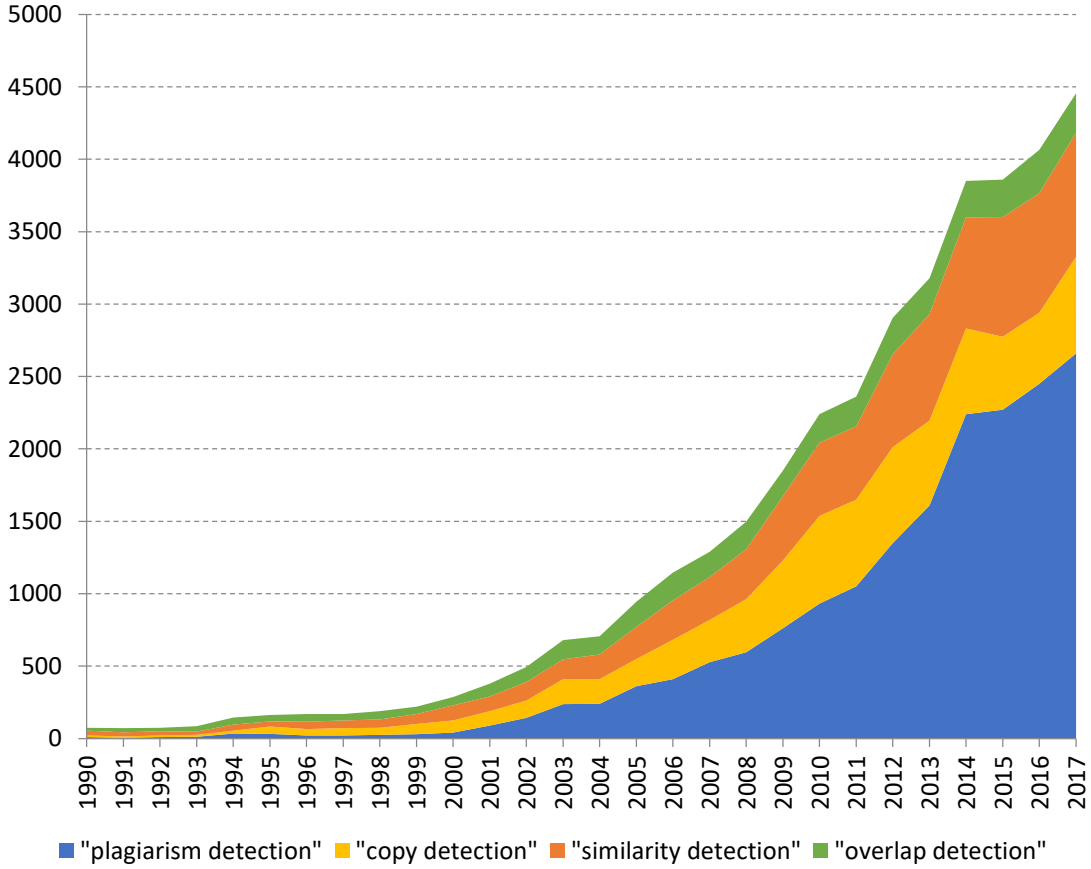
Typology of plagiarism

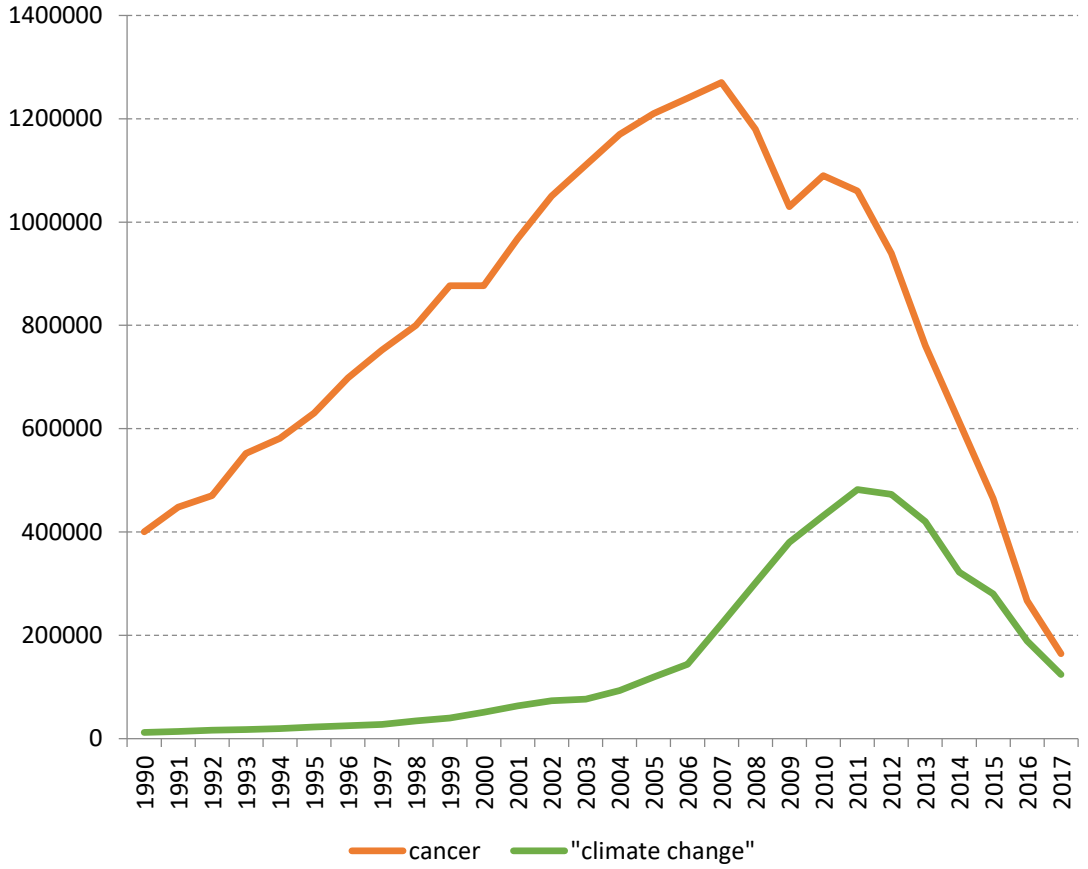
European Network for Academic Integrity

- Copy & Paste
  - Sources possibly mentioned (pawn sacrifice, cut & slide)
- Weak obfuscation
  - Technical disguise
  - Synonym substitution
- Translation
- Paraphrase
  - Mosaic, clause quilts
- Structural plagiarism
  - Boilerplate, template
  - Idea plagiarism
- Shake & Paste
  - Patch-writing, compilation, remix, mosaic, mash-up

Number of papers (Google Scholar)

European Network for Academic Integrity





Three Layers of Plagiarism Detection

European Network for Academic Integrity

Policies

IPPHEAE

COUNCIL OF EUROPE

CONSEIL DE L'EUROPE

Tools

turnitin®

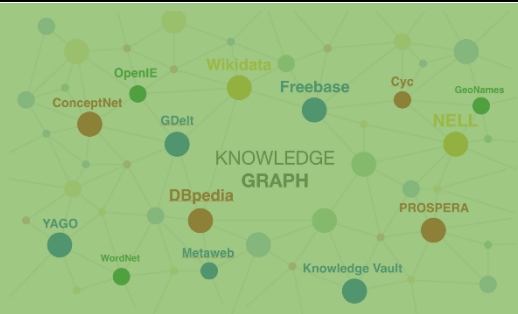
URKUND

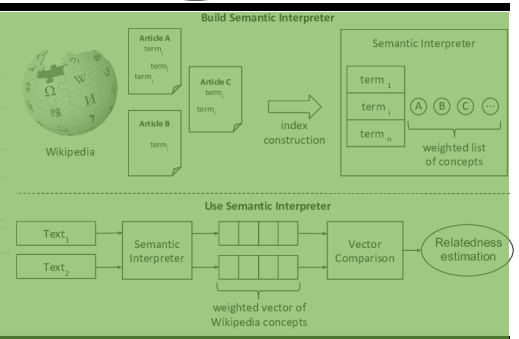
StrikePlagiarism.com

Techniques

$$A = U \begin{pmatrix} w_1 & & \\ & w_2 & \\ & & w_3 \end{pmatrix} V^T$$

$$\text{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|} = \frac{\sum_{i=1}^{N} u_i v_i}{\sqrt{\sum_{i=1}^{N} u_i^2} \sqrt{\sum_{i=1}^{N} v_i^2}}$$





Plagiarism Detection Tasks

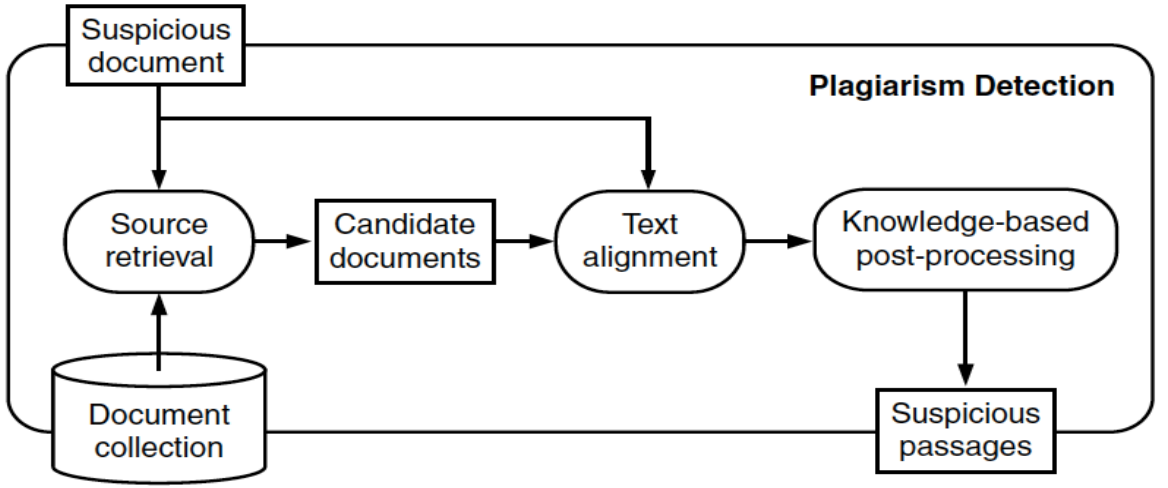
European Network for Academic Integrity

- What “Plagiarism detection” means?
- Traditional task
  - Document Level Detection
    - Given suspicious document D and set of source documents
    - Select source document(s) **similar** to D
- Up-to-date tasks
  - Candidate retrieval
  - Text alignment
  - Paraphrase identification
  - Intrinsic PD

Extrinsic Plagiarism Detection

European Network for Academic Integrity

- Formerly Document Level Detection
- Candidate Retrieval
  - Given a suspicious document D and a search engine / database
  - Retrieve all documents from which the text was reused in D
    - with minimum costs
- Text alignment
  - Given a suspicious document D and a set of candidate documents S
  - Identify passages
    - from documents in S
    - which were re-used in D
- Paraphrase Identification
  - Given two passages, decide whether they have same meaning



Source: Potthast et al. (2014)

Intrinsic Plagiarism Detection

European Network for Academic Integrity

- Given suspicious document
- Identify borders between parts with different style
- Group these parts according to authorship

- Stylometry
- Machine learning

- Related tasks
  - Author identification
  - Author profiling
- Applications: Marketing, law enforcement

Howe to Solve these Tasks?

European Network for Academic Integrity

- Typology of Plagiarism Detection Techniques
  - Lexical-based
  - Syntax-based
  - Semantics-based

- Applications
  - Recommender systems
  - Detection of similar documents
    - related news
    - duplicated work

Lexical-based Methods

European Network for Academic Integrity

- N-Grams
  - N consecutive characters/words
  - Former approach: Fingerprinting
- Vector Space Models
  - Bag-of-words approach
  - Each word is a dimension in a vector space
  - Values weighted (tf-idf)
  - Documents/passages treated as vectors
    - cosine similarity measure

Syntax-based Methods

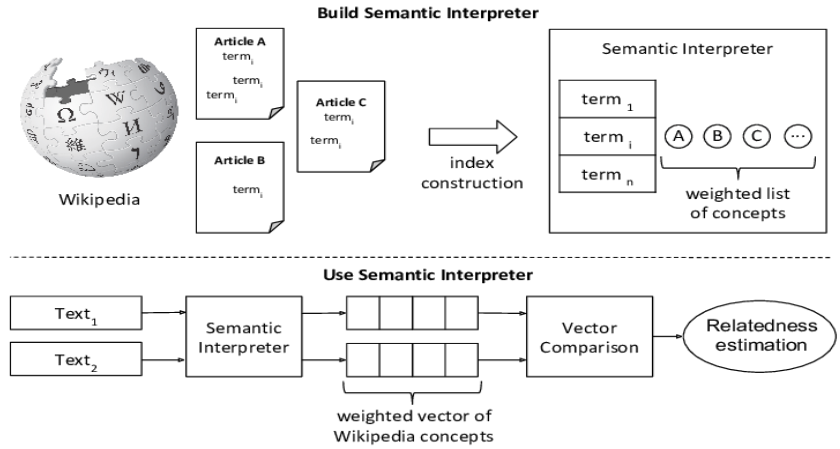
European Network for Academic Integrity

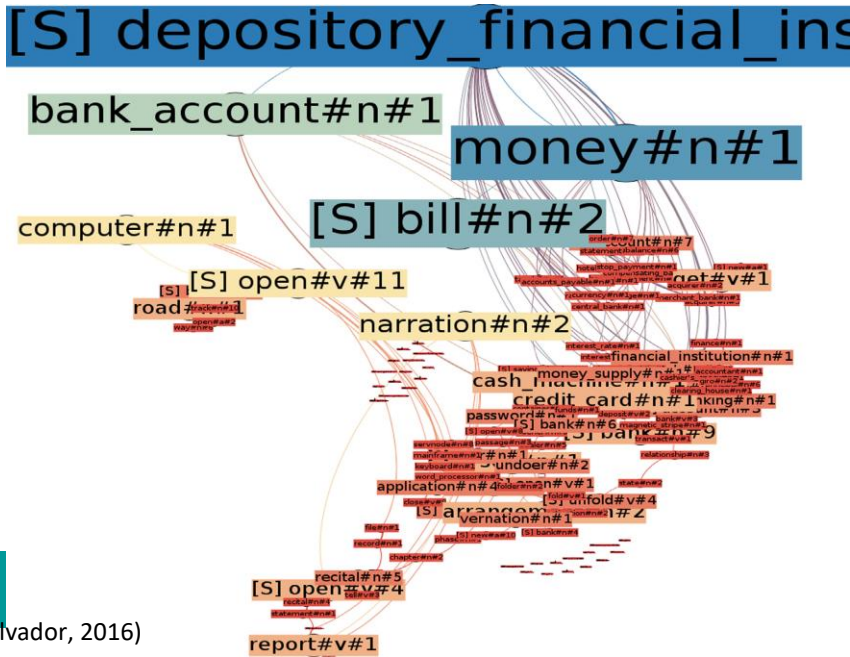
- Part-of-speech tagging
  - marks each word with its class (noun, verb,…)
  - pre-processing for semantic analysis
  - comparison within classes
- Syntactic graphs
  - sentence represented as a graph
  - syntactic relations marked
  - allows structural plagiarism detection

Semantics-based Methods

European Network for Academic Integrity

- Latent Semantic Analysis
  - Singular Value Decomposition
  - Grouping documents by meaning
- Explicit Semantic Analysis
  - Vector of concepts
- Knowledge Graph Analysis
  - Passage represented as a graph
  - Graph similarity metrics
- Bag of words – meaning problem
  - A loves B ≈ B loves A





Knowledge graph built from the sentence "I opened a new bank account" (Franco-Salvador, 2016)

Tools for Semantic Analysis

European Network for Academic Integrity

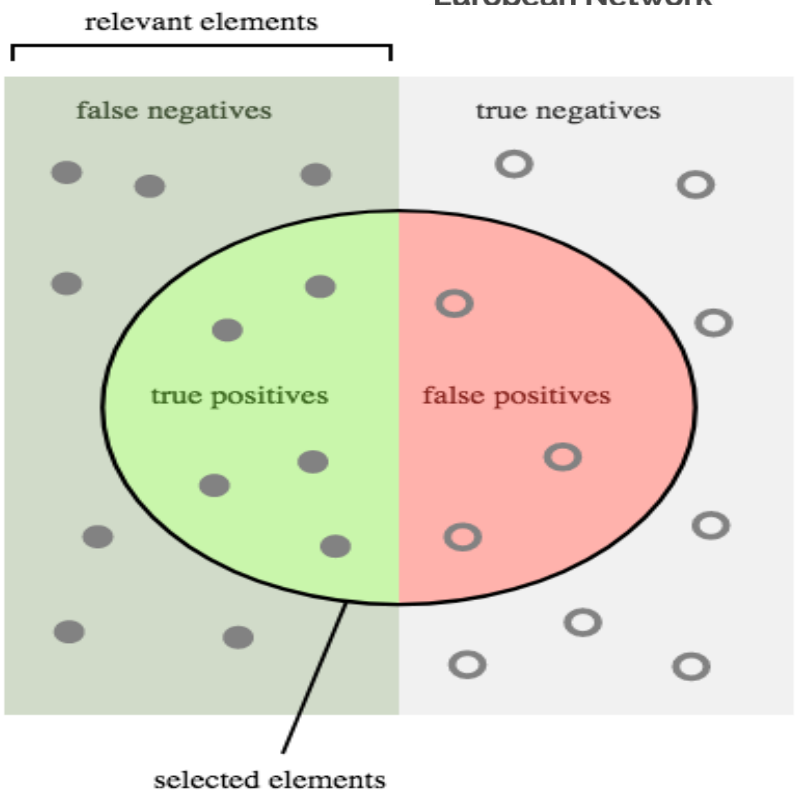
- WordNet (Princeton)
  - English dictionary + thesaurus
- EuroVoc (EU)
  - Multilingual thesaurus
- Wikipedia
- Wikidata
- BabelNet

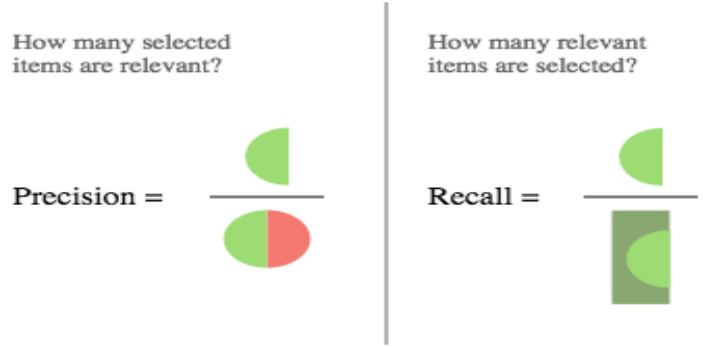
Evaluation of Techniques

- PAN corpuses, Microsoft Research Paraphrase Corpus, German Paraphrase Corpus, P4P…

- Recall
  - What portion of plagiarism was revealed?
  - R = TP/(TP+FN)
- Precision
  - What portion of revealed cases are plagiarism?
  - P = TP/(TP+FP)

- F-measure
  - Harmonic mean of Recall and Precision
  - F = 2*P*R/(P+R)





Accuracy of Extrinsic Plagiarism Detection

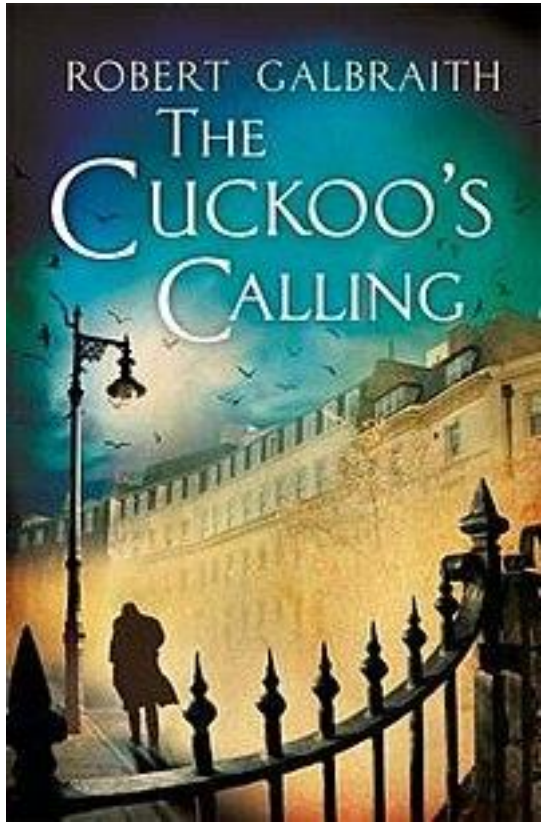
European Network for Academic Integrity

- Copy-paste detection ≈ 100 %
- Synonym replacement detection ≈ 90 %
- Paraphrase identification ≈ 80 %
- Summary identification ≈ 75%
- Cross-language plagiarism detection ≈ 70 %
- Structure, idea,… ???

Author Identification & Profiling

European Network for Academic Integrity

- Up-to-date accuracy
  - Style change identification ≈ 60 %
  - Grouping by authorship ≈ 60 %
  - Native language prediction ≈ 65 – 85 %
  - Gender identification ≈ 80 %
  - Age-group identification ≈ 50 – 55 %
- Robert Galbraith: *The Cuckoo's Calling* (2013)
- Real author: J.K. Rowling
- Related task: Author obfuscation



Three Layers of Plagiarism Detection

European Network for Academic Integrity

Policies

PPHEAE

COUNCIL OF EUROPE

CONSEIL DE L'EUROPE

Tools

turnitin®

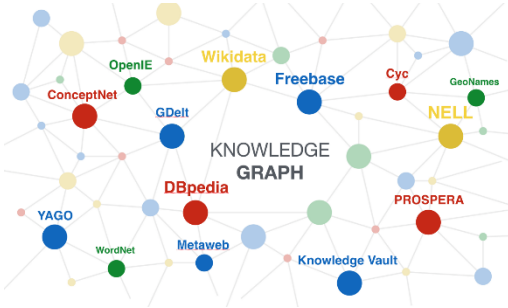
URKUND

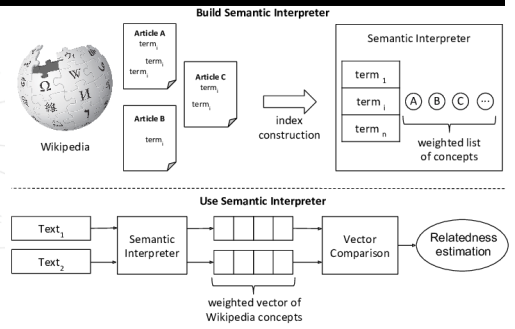
StrikePlagiarism.com

Techniques

$$\begin{pmatrix} & & \\ & A & \\ & & \end{pmatrix} = \begin{pmatrix} & & \\ & U & \\ & & \end{pmatrix} \begin{pmatrix} w_1 & & \\ & w_2 & \\ & & w_3 \end{pmatrix} \begin{pmatrix} & & \\ & V^T & \\ & & \end{pmatrix}$$

$$\text{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|} = \frac{\sum_{i=1}^{N} u_i v_i}{\sqrt{\sum_{i=1}^{N} u_i^2} \sqrt{\sum_{i=1}^{N} v_i^2}}$$





Layer 2: Plagiarism Detection Tools

European Network for Academic Integrity

- Which one is the best?
  - For given language
  - For given type of plagiarism
  - Considering other criteria
- Prof. Debora Weber-Wulff Testing
  - 2004, 2007, 2008, 2010, 2013
- Less sound tests
  - Vani&Gupta, 2016; Chowdhury&Bhattacharyya, 2016

Technical Disguise Test

European Network for Academic Integrity

System	Indexes IJEI	Homoglyphs	White characters	Synonyms	Text as image
Turnitin	YES	YES	NO	YES	NO
Urkund	NO	YES	NO (PLAINTEXT)		NO
Theses.cz	NO	NO (PLAINTEXT)	NO (PLAINTEXT)	NO	NO
Slovenia	NO	YES	YES	YES	YES
Slovakia	NO	NO (PLAINTEXT)	NO (PLAINTEXT)		NO

ENAI Standardized Testing

European Network for Academic Integrity

- Research gap: No standardized testing in place

- Develop sound methodology
- Prepare good testing set
  - Various European languages
    - cross-language plagiarism
  - Various plagiarism types
  - Various types of sources
    - Wikipedia, Internet, OA papers, paid papers
- More evaluation criteria
  - Results, speed, user friendliness,…

Three Layers of Plagiarism Detection

European Network for Academic Integrity

Policies

PPHEAE

COUNCIL OF EUROPE

CONSEIL DE L'EUROPE

Tools

turnitin®

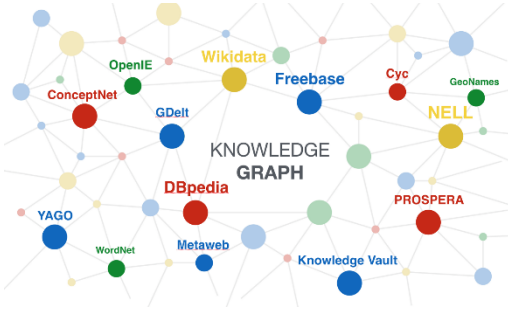
URKUND

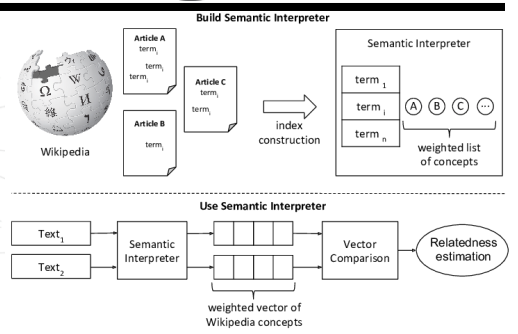
StrikePlagiarism.com

Techniques

$$\begin{pmatrix} & & \\ & A & \\ & & \end{pmatrix} = \begin{pmatrix} & & \\ & U & \\ & & \end{pmatrix} \begin{pmatrix} w_1 & & \\ & w_2 & \\ & & w_3 \end{pmatrix} \begin{pmatrix} & & \\ & V^T & \\ & & \end{pmatrix}$$

$$\text{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|} = \frac{\sum_{i=1}^{N} u_i v_i}{\sqrt{\sum_{i=1}^{N} u_i^2} \sqrt{\sum_{i=1}^{N} v_i^2}}$$





Layer 3: Policies for Plagiarism

European Network for Academic Integrity

- Projects exploring policies
  - IPPHEAE project (2010-2013)
  - SEEPPAI project (2016-2017)
  - Follow-up (CoE) in 2018-2019
- Great books
  - Debora Weber-Wulff (2014): False Feathers
  - Tracey Bretag (2016): Handbook for Academic Integrity

Systemic Effort Works

European Network for Academic Integrity

Curtis & Vardanega, 2016

- 10-years study (2004-2014)
  - Introduced course on academic writing
  - Started using Turnitin
  - Introduced criterion and standards-based assessment
  - Implemented educational changes
- All forms of plagiarism DECREASED
  - except recycling and ghostwriting

Owens & White, 2013

- 5-years study (2007-2011)
  - PD SW as both deterrent and formative tool
  - Educational about academic writing and authorship
- Self-reported plagiarism DROPPED significantly
- Nature of assessment has a central role

Academic Integrity Maturity Model

European Network for Academic Integrity

- **Transparency** in academic integrity and quality assurance
- **Policies**: fair, effective and consistent
- **Sanctions**: standard range
- **Software tools**
- **Prevention** strategies and measures
- **Communication** about policies and procedures
- **Knowledge** and understanding about academic integrity
- **Training** for students and teachers
- **Research** and innovation in academic integrity

(IPPHEAE, 2013)

Institutional Policy

European Network for Academic Integrity

- Problem identification
  - Students? Teachers? Scientists? Management?
  - Plagiarism? Exam cheating? Contract cheating?
  - Inconsistent approach of teachers?
  - Risk of a scandal?
- Propose set of measures
  - Inspiration from others
- Consider differences
  - Culture: focus on motivation or penalties?
  - Society: positive/negative examples
  - Field of study: type of assignments, type of assessment
  - Institution…

General Findings

European Network for Academic Integrity

- One-fits-all strategy ineffective
- Multilayered, evidence-based, longitudinal strategy

- Central role of **assessment**
- Address both **educational approaches**
  - for “good” students
- and **deterrence strategies**
  - for “bad” students
  - students who cheat tend to cheat repeatedly
- Research for gathering **evidence** and impact evaluation

ENAI Tools Coming Soon

European Network for Academic Integrity

- Glossary of terms
  - already published
- Guidelines for academic integrity
  - almost ready
- Self-evaluation tools
  - this autumn
- Manual for cross-sector cooperation
  - this summer
- Collection of educational materials
  - this summer
- Announcements in ENAI Newsletter

Sources

- Chowdhury, H. A., & Bhattacharyya, D. K. (2016). Plagiarism: Taxonomy, Tools and Detection Techniques. In 19th National Convention on Knowledge, Library and Information Networking (NACLIN 2016). Retrieved from http://arxiv.org/abs/1801.06323
- Curtis, G. J., & Clare, J. (2017). How Prevalent is Contract Cheating and to What Extent are Students Repeat Offenders? Journal of Academic Ethics, 15(2), 115–124. https://doi.org/10.1007/s10805-017-9278-x
- Curtis, G. J., & Vardanega, L. (2016). Is plagiarism changing over time? A 10-year time-lag study with three points of measurement. Higher Education Research & Development, 35(6), 1167–1179. https://doi.org/10.1080/07294360.2016.1161602
- Franco-Salvador, M., Rosso, P., & Montes-y-Gómez, M. (2016). A systematic study of knowledge graph analysis for cross-language plagiarism detection. Information Processing & Management, 52(4), 550–570. https://doi.org/https://doi.org/10.1016/j.ipm.2015.12.004
- Meuschke, N., & Gipp, B. (2013). State-of-the-art in detecting academic plagiarism. International Journal for Educational Integrity, 9(1), 50–71. Retrieved from http://www.ojs.unisa.edu.au/index.php/IJEI/article/view/847
- Meuschke, N., Siebeck, N., Schubotz, M., & Gipp, B. (2017). Analyzing Semantic Concept Patterns to Detect Academic Plagiarism. In Proceedings of the 6th International Workshop on Mining Scientific Publications (pp. 46–53). New York, NY, USA: ACM. https://doi.org/10.1145/3127526.3127535
- Owens, C., & White, F. A. (2013). A 5-year systematic strategy to reduce plagiarism among first-year psychology university students. Australian Journal of Psychology, 65(1), 14–21. https://doi.org/10.1111/ajpy.12005
- Potthast, M., Hagen, M., Beyer, A., Busse, M., Tippmann, M., Rosso, P., & Stein, B. (2014). Overview of the 6th International Competition on Plagiarism Detection. In L. Cappellato, N. Ferro, M. Halvey, & W. Kraaij (Eds.), Working Notes Papers of the CLEF 2014 Evaluation Labs. CLEF and CEUR-WS.org. Retrieved from http://www.clef-initiative.eu/publication/working-notes
- Potthast, M., Rangel, F., Tschuggnall, M., Stamatatos, E., Rosso, P., & Stein, B. (2017). Overview of PAN'17: Author identification, author profiling, and author obfuscation. In Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-319-65813-1_25
- Rosso, P., Rangel, F., Potthast, M., Stamatatos, E., Tschuggnall, M., & Stein, B. (2016). Overview of PAN'16: New challenges for authorship analysis: Cross-genre profiling, clustering, diarization, and obfuscation. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/978-3-319-44564-9_28
- Vani, K., & Gupta, D. (2016). Study on Extrinsic Text Plagiarism Detection Techniques and Tools. Journal of Engineering Science & Technology Review, 9(5). Retrieved from http://jestr.org/downloads/Volume9Issue5/fulltext2952016.pdf
- Weber-Wulff, D. (2014). False feathers: A perspective on Academic Plagiarism. Berlin: Springer. ISBN 978-3-642-39960-2
- https://en.wikipedia.org/wiki/The_Cuckoo%27s_Calling
- https://www.scientificamerican.com/article/how-a-computer-program-helped-show-jk-rowling-write-a-cuckoos-calling/
- https://cdn-images-1.medium.com/max/1600/1*A-VaSSx2cqNHrI19_t4cfA.png
- https://c1.staticflickr.com/4/3738/12221292503_c3eff8db68_b.jpg
- https://www.teqsa.gov.au/sites/g/files/net2046/f/figure-17-unisa-business-school-academic-integrity.png
- https://upload.wikimedia.org/wikipedia/commons/2/26/Precisionrecall.svg
- http://www.alglib.net/matrixops/general/i/svd1.gif

Thank you for your attention

European Network for Academic Integrity

Tomáš Foltýnek

Mendel University in Brno, Czechia

University of Konstanz, Germany

tomas.foltynek@academicintegrity.eu