A New Corpus for the Evaluation of Arabic Intrinsic Plagiarism Detection

Imene Bensalem[1], Paolo Rosso[2], Salim Chikhi[1]

[1]Constantine 2 University, Algeria

[2]Universitat Politècnica de Valeència, Spain

Outline

- □ Arabic Text Reuse
- □ Plagiarism in Arab World
- □ Plagiarism Detection Approaches
  - □ Intrinsic Plagiarism Detection
  - □ Arabic Plagiarism Detection
- □ Building Plagiarism Detection Corpora
- □ InAra: the First Corpus for the Evaluation of Arabic Intrinsic Plagiarism Detection

Outline

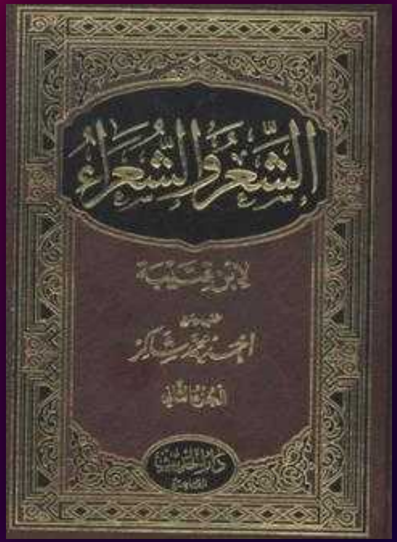
- □ Arabic Text Reuse
- □ Plagiarism in Arab World
- □ Plagiarism Detection Approaches
  - □ Intrinsic Plagiarism Detection
- □ Arabic Plagiarism Detection
- □ Building Plagiarism Detection Corpora
- □ InAra: the First Corpus for the Evaluation of Arabic Intrinsic Plagiarism Detection

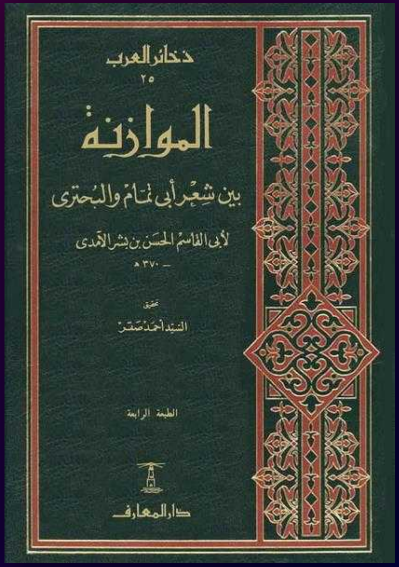
Outline

- □ Arabic Text Reuse
- □ Plagiarism in Arab World
- □ Plagiarism Detection Approaches
  - □ Intrinsic Plagiarism Detection
  - □ Arabic Plagiarism Detection
- □ Building Plagiarism Detection Corpora
- □ InAra: the First Corpus for the Evaluation of Arabic Intrinsic Plagiarism Detection

Arabic Text Reuse

- Text reuse studies were an active field in Arabic literature criticism in the middle ages (Al Manasrah 2009)
- Text reuse was spread and studied especially in the context of poetry

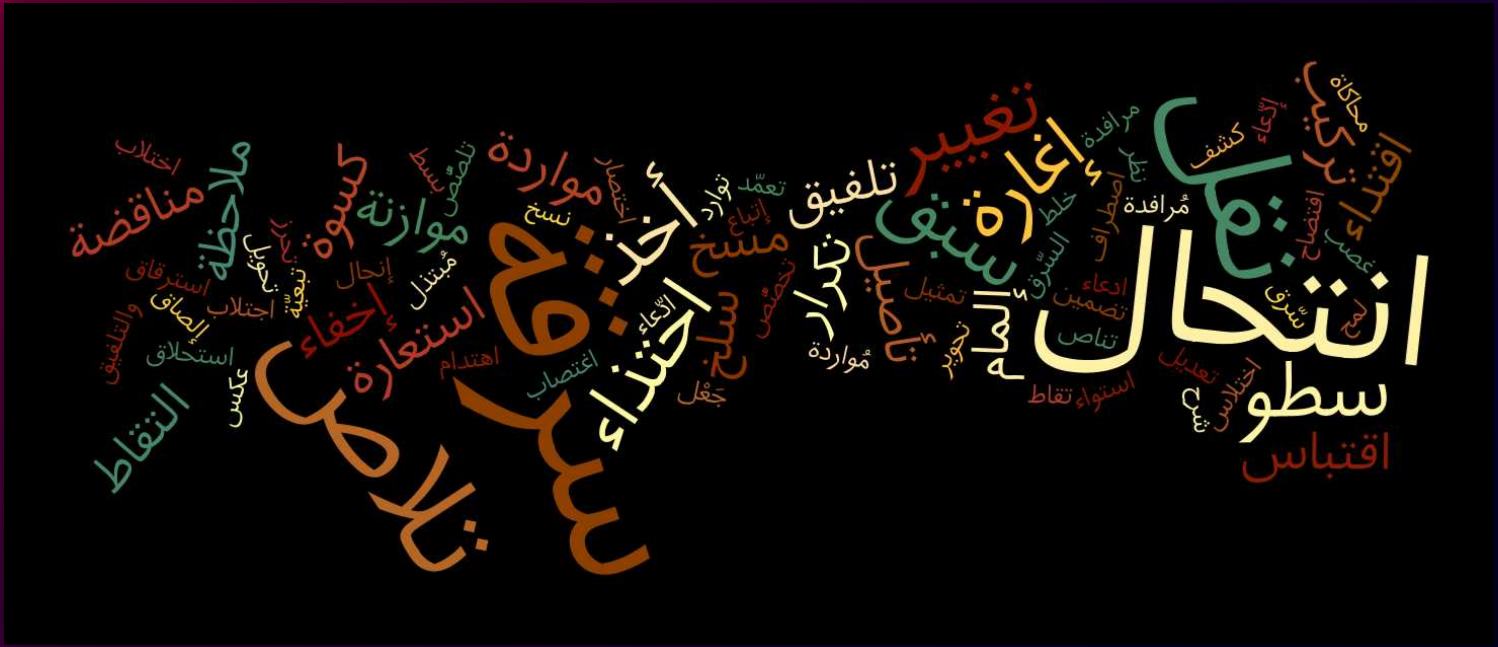




Medieval books on plagiarism in poetry

Arabic Text Reuse

- □ All kinds of text reuse in literature were called "literary theft" by the Arab medieval critics,
- □ BUT, "literary theft" was classified into many types and levels, some of which were acceptable or even considered an art that could be done only by the very prominent poets! (Tabana 1956)



Words used to express the different kinds of literary thefts

Arabic Text Reuse

Till now, the word “Plagiarism” does not have a unique term in Arabic.
Some of the used terms: literary theft , scientific theft, arrogation…

But there is a tendency of using the word انتحال /intihal/ that means arrogation of authorship (in poetry)

انتحال

Plagiarism in Arab World

Study (Al-Jundy 2013)*

- 18 magister candidates from different universities in Egypt were asked to write 3 essays : 2 in Arabic and 1 in English on topics related to library sciences. It was reported that:
- All of them plagiarized
- The percentage of plagiarism in each plagiarized essay was between 45%-76% , it reached 81% in essays written in English
- They forgot or do not know the importance of giving acknowledge in academic writing
- They plagiarized despite the warning on plagiarism because they did not know about the existence of any plagiarism detection software!

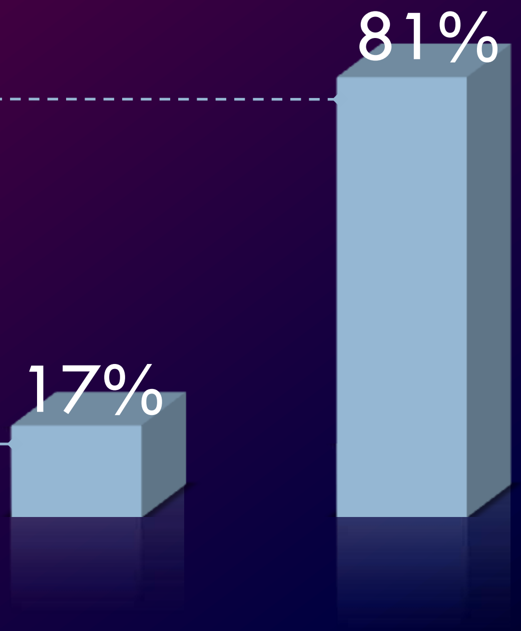
*برامج اكتشاف الانتحال فى البيئة الرقمية المتاحة عبر الويب: دراسة مسحية تطبيقية. البحث الكامل للملخص المطول المنشور بـ: المؤتمر العلمى العاشر لقسم المكتبات والوثائق والمعلومات كلية الآداب – جامعة القاهرة إتاحة المعرفة وحق الوصول إلى المعلومات فى المجتمع العربى المعاصر :التحديات والتطلعات. القاهرة 15-16 مايو2013 . لم ينشر بعد.
(Many thanks to the author for sending me this paper before its publication)

Plagiarism in Arab World

One of plagiarism reasons in Arab world

81% of the non Arab academic libraries PROVIDE on their websites information (tutorials , activities, policy…) on plagiarism

17% of Arab academic libraries DO NOT PROVIDE on their websites any information on plagiarism

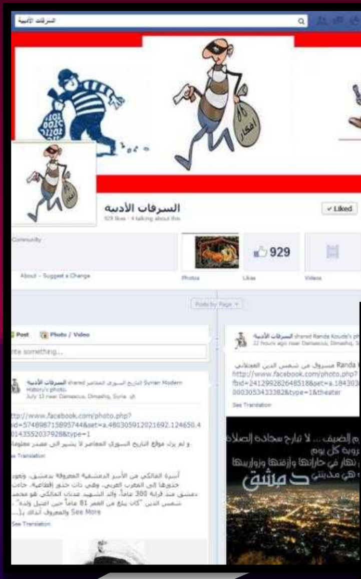


Source of statistics : (Saleh and Alsayed 2012)*

*عماد عيسى صالح، أماني محمد السيد. دور المكتبات الأكاديمية في منع السرقات العلمية واكتشافها: دراسة استكشافية لخدمات المكتبات وبرمجيات كشف الانتحال. المؤتمر الدولي للتعلم الالكتروني في الوطن العربي، القاهرة 9-11 يوليو 2012.

To Follow Plagiarism Cases…

السرقات الأدبية@ Facebook



السرقات الأدبية

السرقات الأدبية

✓ Liked Message

Community

About – Suggest a Change

Photos 929 Likes Videos Notes

Post Photo / Video

السرقات الأدبية shared التاريخ السوري المعاصر Syrian Modern History's photo.
July 13 near Damascus, Dimashq, Syria

http://www.facebook.com/photo.php?fbid=574898715895744&set=a.480305912021692.124650.4014335520379288&type=1

و لم يرث موقع التاريخ السوري المعاصر لا يشير الى مصدر معلومات

See Translation

أسرة الفالكي من الأسر الدمشقية المعروفة بدمشق، وتعود جذورها الى المغرب العربي، وهي ذات جذور إقطاعية. جاءت دمشق منذ قرابة 300 عاماً. والد الشهيد عدنان الفالكي هو محمد شمس الدين "كان يبلغ من العمر 81 عاماً حين اغتيل ولده". والمعروف آنذاك بـ... See More

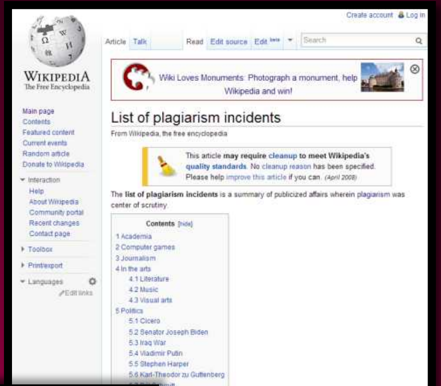
See Translation

السرقات الأدبية shared Randa Koudsi's photo.
22 hours ago near Damascus, Dimashq, Syria

Randa مسروق من شمس الدين العجلاني
http://www.facebook.com/photo.php?fbid=241299282648518&set=a.184303...000305343338&type=1&theater

See Translation

List of plagiarism incidents @ Wikipedia



Create account Log in

Article Talk Read Edit source Edit beta Search

Wikipedia The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Toolbox

Print/export

Languages

Edit links

Wiki Loves Monuments: Photograph a monument, help Wikipedia and win!

List of plagiarism incidents

From Wikipedia, the free encyclopedia

This article **may require cleanup to meet Wikipedia's quality standards**. No cleanup reason has been specified. Please help improve this article if you can. *(April 2008)*

The **list of plagiarism incidents** is a summary of publicized affairs wherein plagiarism was center of scrutiny.

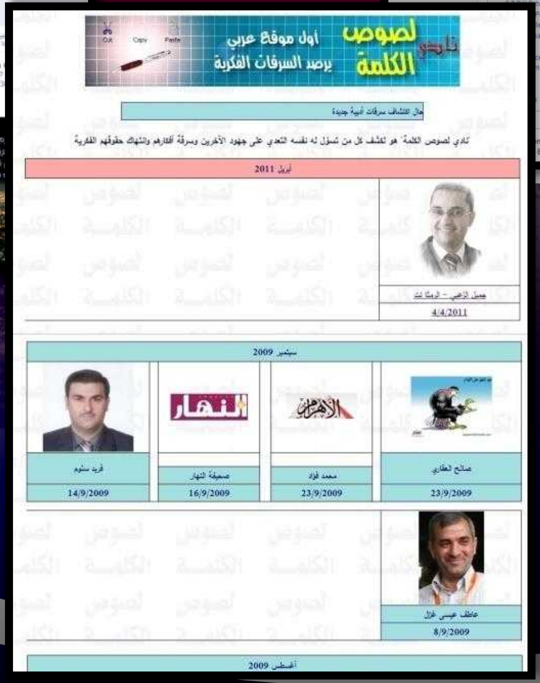
Contents [hide]

1 Academia
2 Computer games
3 Journalism
4 In the arts
4.1 Literature
4.2 Music
4.3 Visual arts
5 Politics
5.1 Cicero
5.2 Senator Joseph Biden
5.3 Iraq War
5.4 Vladimir Putin
5.5 Stephen Harper
5.6 Karl-Theodor zu Guttenberg
…mitt
…Ponta
…Semjén
…an Rowhani
…ces
…m H. Swanson

[edit source | edit beta]

…ntury Muslim scholar Al-Khatib al-Baghdadi considered the *Book of Animals* of …869) to be "little more than a plagiarism" of Aristotle's *Kitāb al-Hayawān*, a

http://bader59.com



حال اكتشاف سرقات أدبية جديدة

نادي لصوص الكلمة" هو لكشف كل من تسول له نفسه التعدي على جهود الآخرين وسرقة أفكارهم وانتهاك حقوقهم الفكرية

أبريل 2011

جميل الزعبي – الرمثا نت
4/4/2011

سبتمبر 2009

صالح العقاري
23/9/2009

محمد فؤاد
23/9/2009

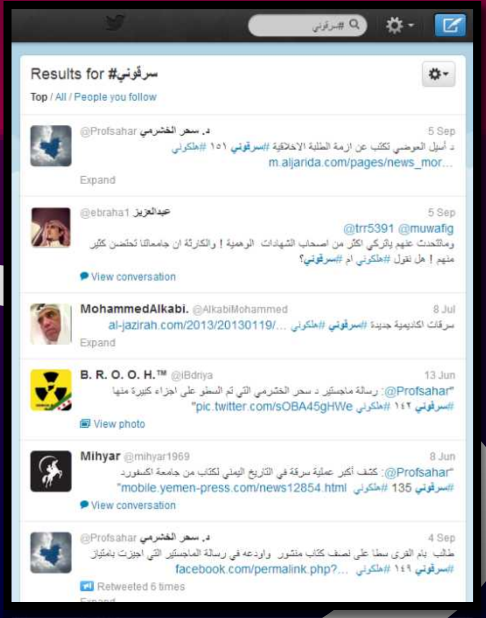
صحيفة النهار
16/9/2009

فريد سلوم
14/9/2009

عاطف عيسى غزال
8/9/2009

أغسطس 2009

#سرقوني@Tweeter



#سرقوني

Results for #سرقوني

Top / All / People you follow

د. سحر الخشرمي @Profsahar 5 Sep
د أسيل العوضي تكتب عن ازمة الطلبة الاخلاقية #سرقوني ١٥١ #هلكوني m.aljarida.com/pages/news_mor…
Expand

عبدالعزيز @ebraha1 5 Sep
@trr5391 @muwafig
وماتحدث عنهم باتركي اكثر من اصحاب الشهادات الوهمية ! والكارثة ان جامعاتنا تحتضن كثير منهم ! هل نقول #هلكوني ام #سرقوني؟
View conversation

MohammedAlkabi. @AlkabiMohammed 8 Jul
سرقات اكاديمية جديدة #سرقوني #هلكوني al-jazirah.com/2013/20130119/…
Expand

B. R. O. O. H.™ @iBdriya 13 Jun
"@Profsahar: رسالة ماجستير د سحر الخشرمي التي تم السطو على اجزاء كبيرة منها #سرقوني ١٤٢ #هلكوني pic.twitter.com/sOBA45gHWe"
View photo

Mihyar @mihyar1969 8 Jun
"@Profsahar: كشف أكبر عملية سرقة في التاريخ اليمني لكتاب من جامعة اكسفورد #سرقوني 135 #هلكوني mobile.yemen-press.com/news12854.html"
View conversation

د. سحر الخشرمي @Profsahar 4 Sep
طالب بام القرى سطا على نصف كتاب منشور واودعه في رسالة الماجستير التي اجيزت بامتياز #سرقوني ١٤٩ #هلكوني facebook.com/permalink.php?…
Retweeted 6 times

Outline

- Arabic Text Reuse
- Plagiarism in Arab World
- Plagiarism Detection Approaches
  - Intrinsic Plagiarism Detection
  - Arabic Plagiarism Detection
- Building Plagiarism Detection Corpora
- InAra: the First Corpus for the Evaluation Arabic Intrinsic Plagiarism Detection

Outline

- Arabic Text Reuse
- Plagiarism in Arab World
- Plagiarism Detection Approaches
  - Intrinsic Plagiarism Detection
- Arabic Plagiarism Detection
- Building Plagiarism Detection Corpora
- InAra: the First Corpus for the Evaluation Arabic Intrinsic Plagiarism Detection

Plagiarism Detection Approaches

Oh, I think I read this passage somewhere else !
Ah yes, I remembered where I read it!
External Approach

Plagiarism Detection Approaches

The writing style of this passage seems to be very different from the rest of the paper !

Intrinsic Approach

Plagiarism Detection Approaches

Plagiarism Detection

External

Signs of plagiarism
Similarities with other texts

Need a reference collection ?
YES

Related domain
Information retrieval

Intrinsic

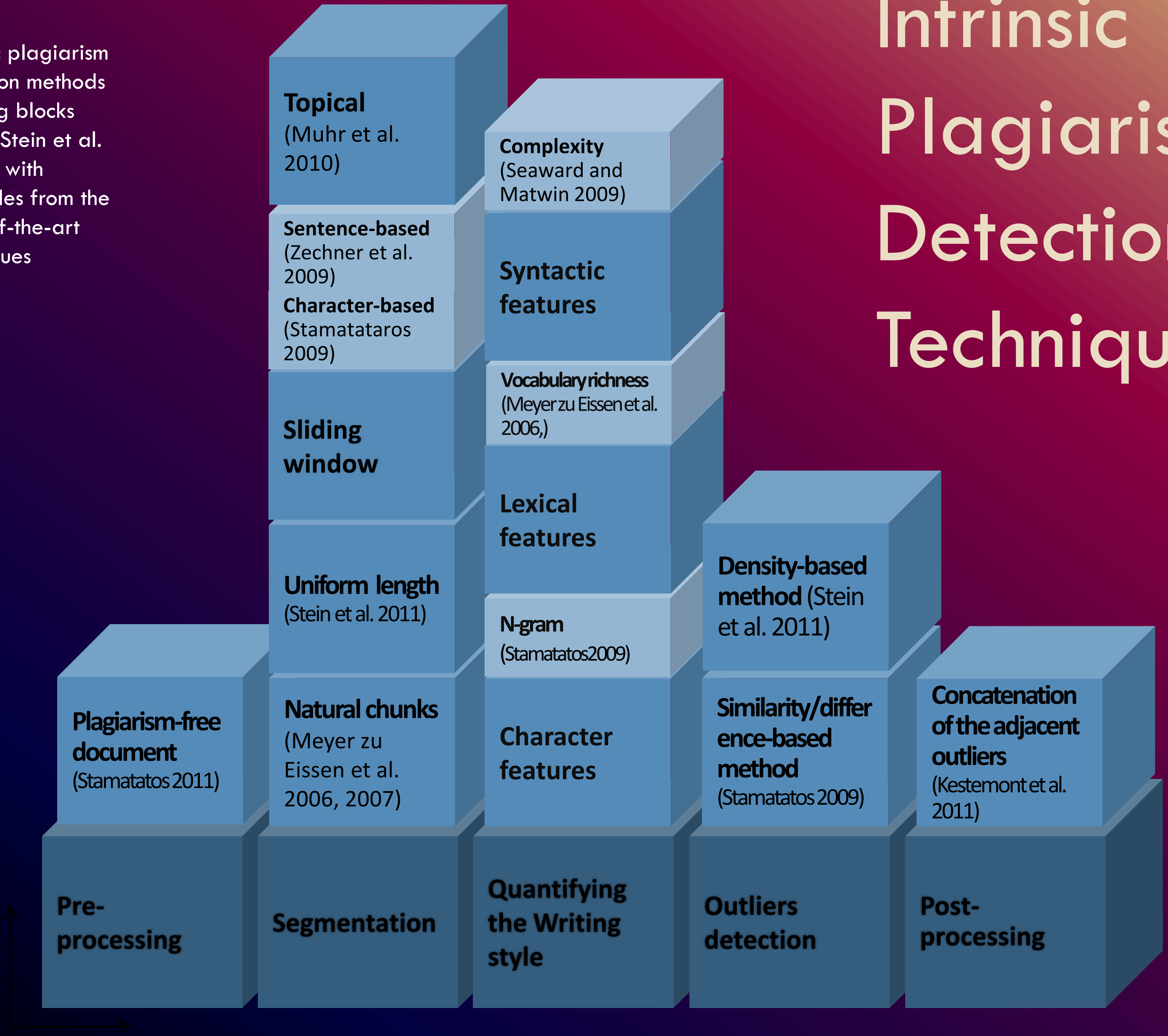
Signs of plagiarism
Inconsistencies in the writing style

Need a reference collection ?
NO

Related domain
Authorship verification

Intrinsic Plagiarism Detection Techniques

Intrinsic plagiarism detection methods building blocks (after (Stein et al. 2011)) with examples from the state-of-the-art techniques



Pre-processing

Plagiarism-free document (Stamatatos 2011)

Segmentation

Natural chunks (Meyer zu Eissen et al. 2006, 2007)

Uniform length (Stein et al. 2011)

Sliding window

Sentence-based (Zechner et al. 2009)

Character-based (Stamatataros 2009)

Topical (Muhr et al. 2010)

Quantifying the Writing style

Character features

N-gram (Stamatatos2009)

Lexical features

Vocabulary richness (Meyer zu Eissen et al. 2006,)

Syntactic features

Complexity (Seaward and Matwin 2009)

Outliers detection

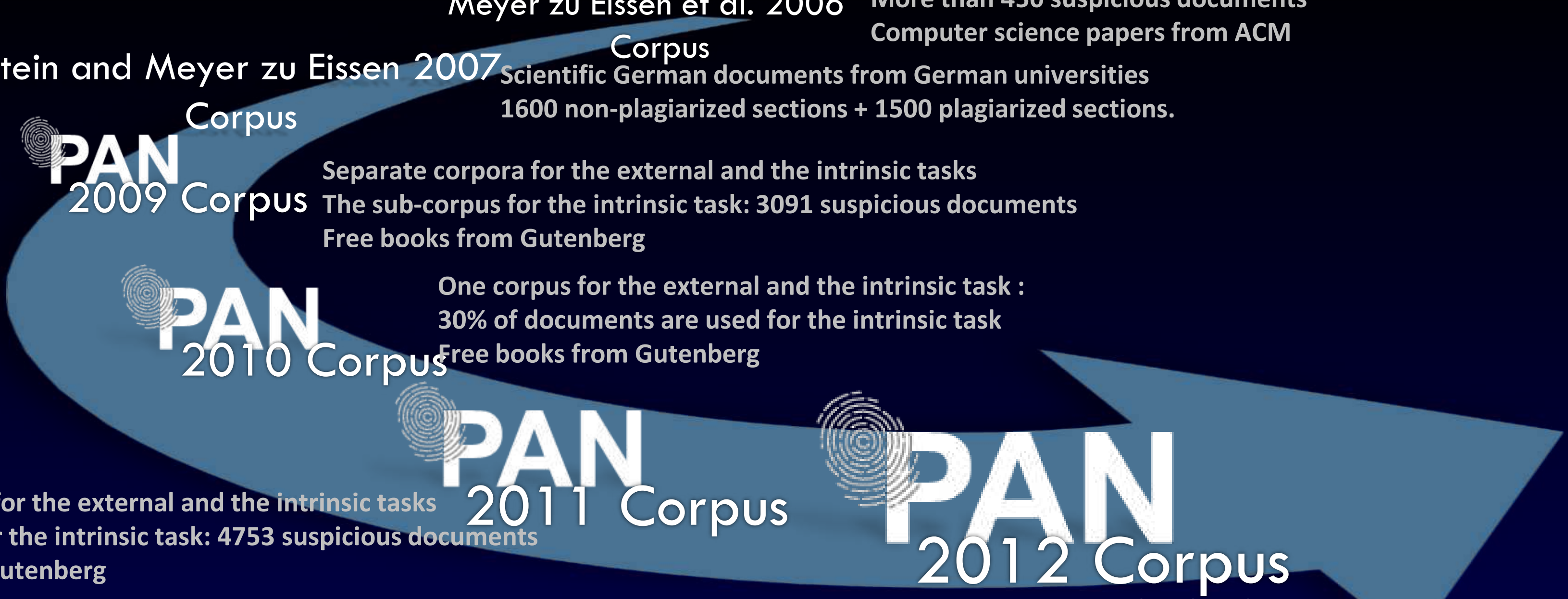
Similarity/difference-based method (Stamatatos 2009)

Density-based method (Stein et al. 2011)

Post-processing

Concatenation of the adjacent outliers (Kestemont et al. 2011)

Intrinsic Plagiarism Detection Corpora



Meyer zu Eissen et al. 2006 Corpus

More than 450 suspicious documents
Computer science papers from ACM

Stein and Meyer zu Eissen 2007 Corpus

Scientific German documents from German universities
1600 non-plagiarized sections + 1500 plagiarized sections.

PAN 2009 Corpus

Separate corpora for the external and the intrinsic tasks
The sub-corpus for the intrinsic task: 3091 suspicious documents
Free books from Gutenberg

PAN 2010 Corpus

One corpus for the external and the intrinsic task :
30% of documents are used for the intrinsic task
Free books from Gutenberg

PAN 2011 Corpus

Separate corpora for the external and the intrinsic tasks
The sub-corpus for the intrinsic task: 4753 suspicious documents
Free books from Gutenberg

PAN 2012 Corpus

Authorship clustering corpus
Less than 10 documents composed of segments from many authors

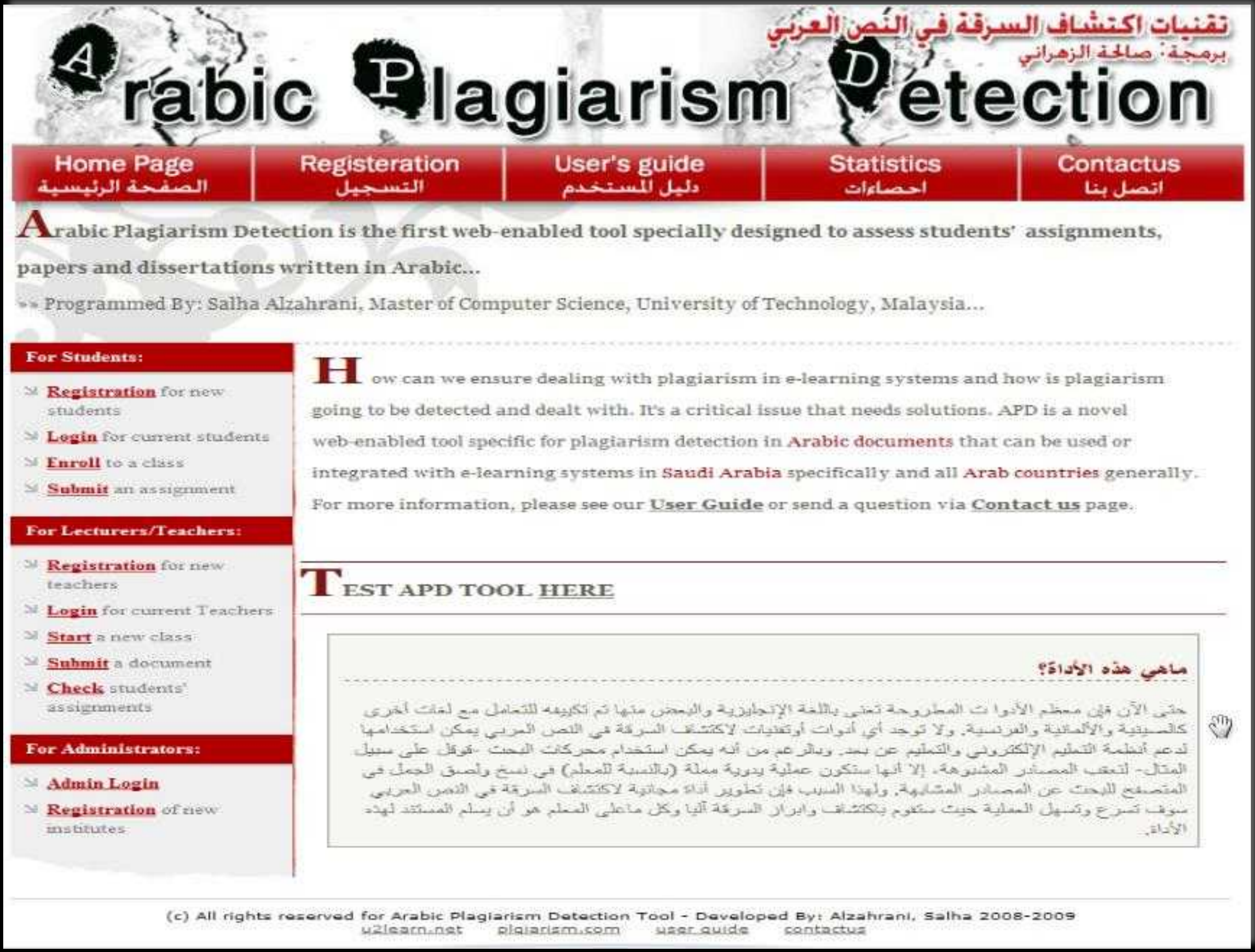
Arabic Plagiarism Detection

- □ Few works evaluated on different corpora
- □ All of which are in external plagiarism detection

Stop-word removal
Relative Frequency Model
fuzzy information retrieval
Synonyms' replacement
Longest Common Substring
lemmatization
Babylon translator
Wikipedia
Winnowing
Heuristic Algorithm
document tree structure
n-gram
lexical chain
Arabic Wordnet
Fingerprinting
Tokenization
Stemming

Cloud of some keywords from Arabic plagiarism detection papers

Arabic Plagiarism Detection Software



Arabic Plagiarism Detection is the first web-enabled tool specially designed to assess students' assignments, papers and dissertations written in Arabic...

» Programmed By: Salha Alzahrani, Master of Computer Science, University of Technology, Malaysia...

For Students:

- Registration for new students
- Login for current students
- Enroll to a class
- Submit an assignment

For Lecturers/Teachers:

- Registration for new teachers
- Login for current Teachers
- Start a new class
- Submit a document
- Check students' assignments

For Administrators:

- Admin Login
- Registration of new institutes

How can we ensure dealing with plagiarism in e-learning systems and how is plagiarism going to be detected and dealt with. It's a critical issue that needs solutions. APD is a novel web-enabled tool specific for plagiarism detection in Arabic documents that can be used or integrated with e-learning systems in Saudi Arabia specifically and all Arab countries generally. For more information, please see our User Guide or send a question via Contact us page.

TEST APD TOOL HERE

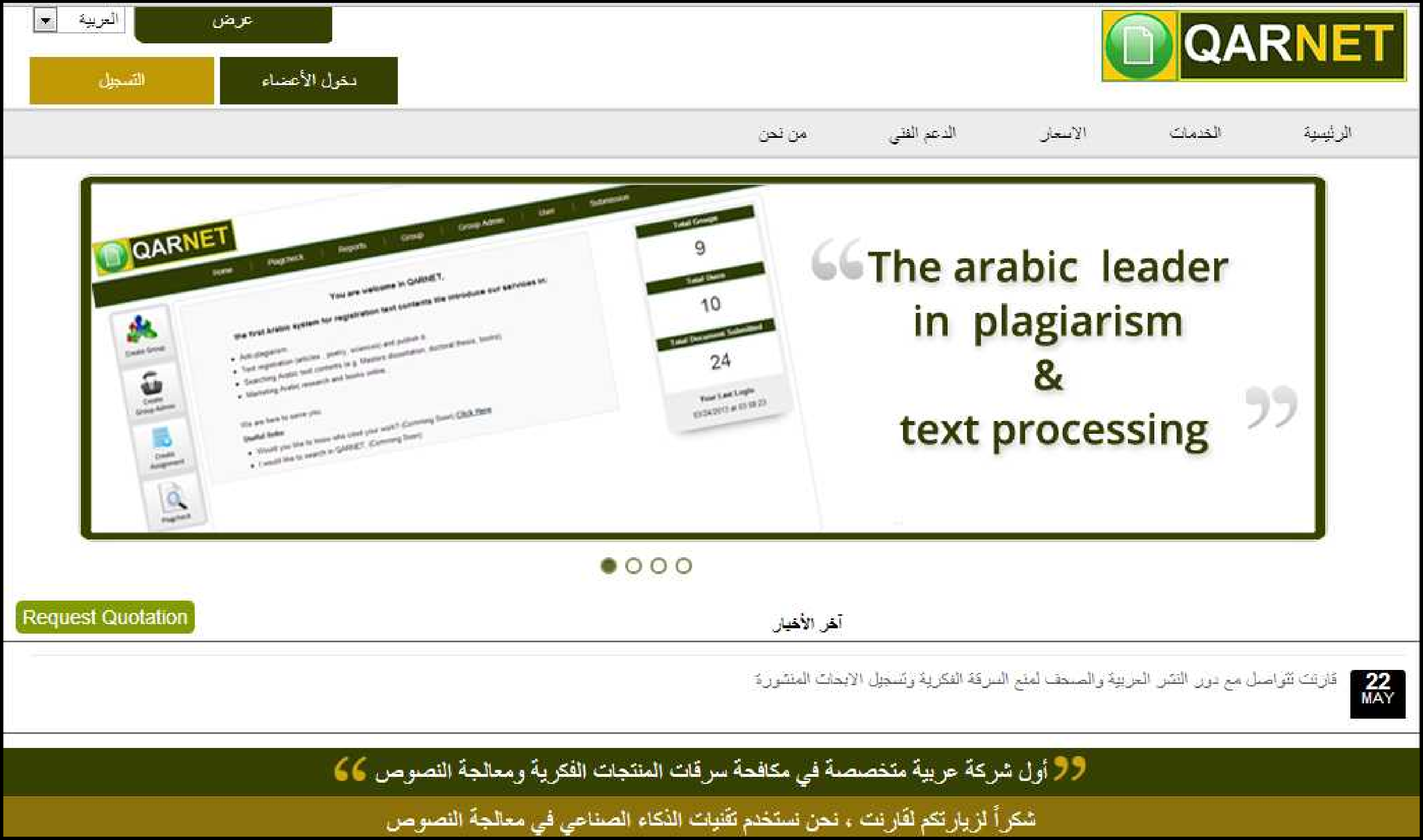
ماهي هذه الأداة؟

حتى الآن فإن معظم الأدوا ت المطروحة تعنى باللغة الإنجليزية والبعض منها تم تكييفه للتعامل مع لغات أخرى كالصينية والألمانية والفرنسية. ولا توجد أي أدوات أوتقنيات لاكتشاف السرقة في النص العربي يمكن استخدامها لدعم أنظمة التعليم الإلكتروني والتعليم عن بعد. وبالرغم من أنه يمكن استخدام محركات البحث -قوقل على سبيل المثال- لتعقب المصادر المشبوهة، إلا أنها ستكون عملية يدوية مملة (بالنسبة للمعلم) في نسخ ولصق الجمل في المتصفح للبحث عن المصادر المشابهة. ولهذا السبب فإن تطوير أداة مجانية لاكتشاف السرقة في النص العربي سوف تسرع وتسهل العملية حيث ستقوم باكتشاف وابراز السرقة آليا وكل ماعلى المعلم هو أن يسلم المستند لهذه الأداة.

(c) All rights reserved for Arabic Plagiarism Detection Tool - Developed By: Alzahrani, Salha 2008-2009
u2learn.net plgiarism.com user guide contactus

Source: Alzahrany papers

Arabic Plagiarism Detection Software



You are welcome in QARNET.

the first Arabic system for registration text contents we introduce our services in:

- Anti-plagiarism
- Text registration (articles , poetry, scientific) and publish it
- Searching Arabic text contents (e.g. Masters dissertation, doctoral thesis, books)
- Marketing Arabic research and books online

We are here to serve you

Useful links

- Would you like to know who cited your work? (Comming Soon) Click Here
- I would like to search in QARNET. (Comming Soon)

Your Last Login 03/24/2013 at 03:08:23

“The arabic leader in plagiarism & text processing”

آخر الأخبار

قارنت تتواصل مع دور النشر العربية والصحف لمنع السرقة الفكرية وتسجيل الابحاث المنشورة

”أول شركة عربية متخصصة في مكافحة سرقات المنتجات الفكرية ومعالجة النصوص“

شكراً لزيارتكم لقارنت ، نحن نستخدم تقنيات الذكاء الصناعي في معالجة النصوص

http://qarnet.com

Intrinsic Plagiarism Detection

State and Challenges

In general

- It is still in its infancy
- It is a reduction of the authorship verification task (one-class classification problem) BUT
  - The class is not well defined (text from which the author writing style is extracted is noisy with the plagiarized chunks)
  - Verify the authorship of short segment of text

In Arabic

- Very few works on Arabic Stylometry and Authorship Analysis (the source of techniques)
- May need religion quotation detection : Arabic text may contain quotation from religious text

Outline

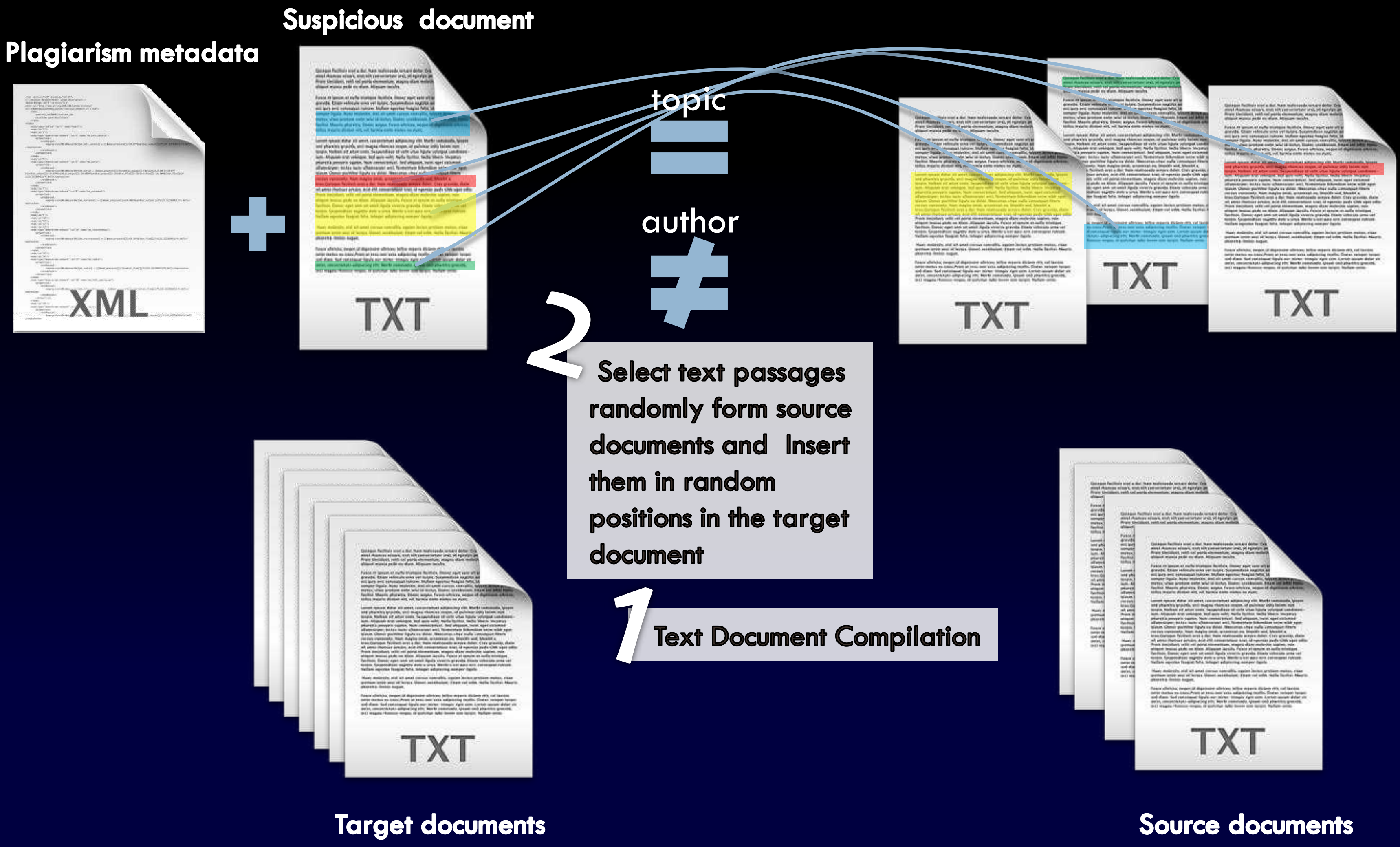
- □ Arabic Text Reuse
- □ Plagiarism in Arab World
- □ Plagiarism Detection Approaches
  - □ Intrinsic Plagiarism Detection
- □ Arabic Plagiarism Detection
- □ Building Plagiarism Detection Corpora
- □ InAra: the First Corpus for the Evaluation of Arabic Intrinsic Plagiarism Detection

Outline

- □ Arabic Text Reuse
- □ Plagiarism in Arab World
- □ Plagiarism Detection Approaches
  - □ Intrinsic Plagiarism Detection
  - □ Arabic Plagiarism Detection
- □ Building Plagiarism Detection Corpora
- □ InAra: the First Corpus for the Evaluation of Arabic Intrinsic Plagiarism Detection

Building Plagiarism Detection Corpora

Automatic Approach



Plagiarism metadata

Suspicious document

topic

author

2 Select text passages randomly form source documents and Insert them in random positions in the target document

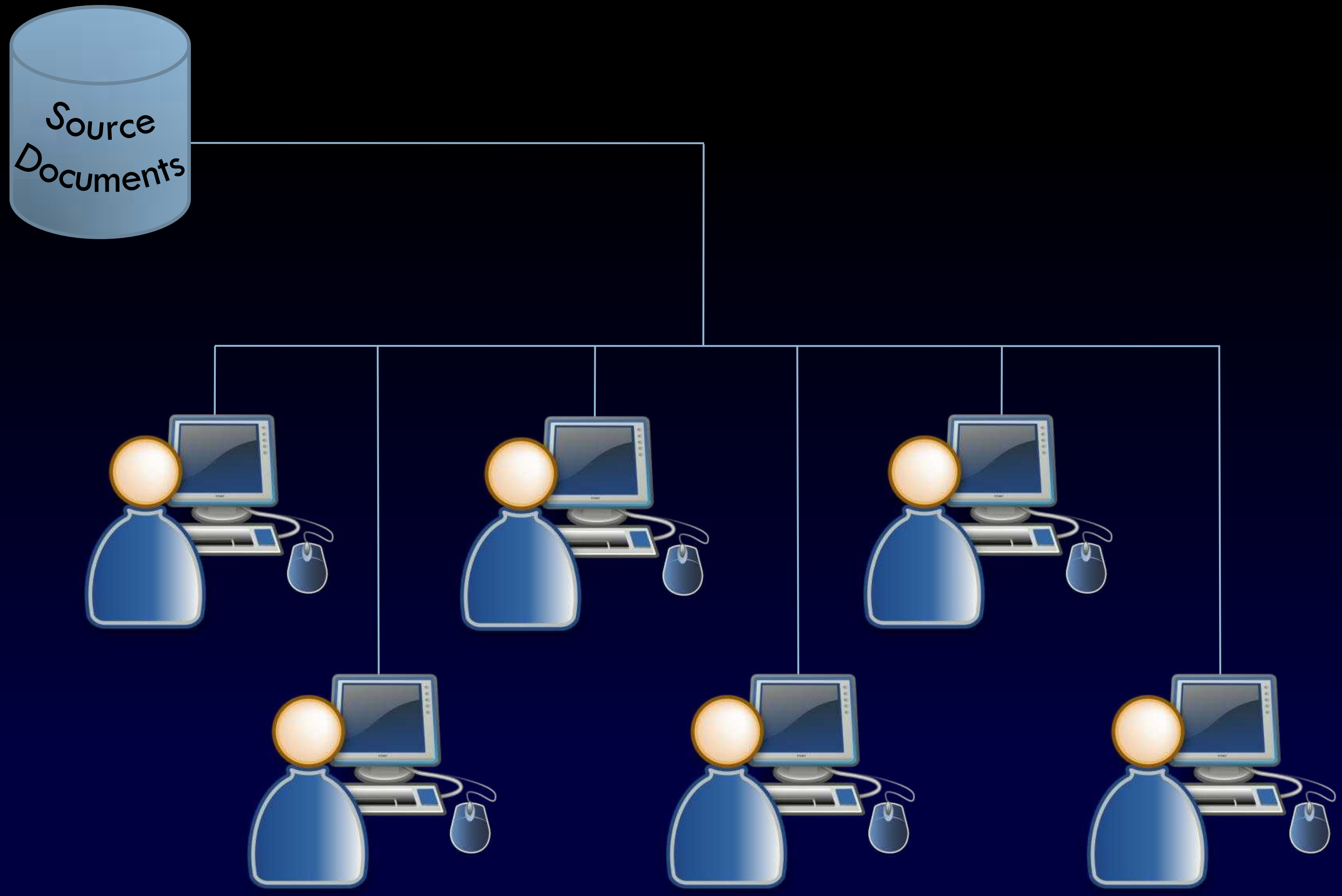
1 Text Document Compilation

Target documents

Source documents

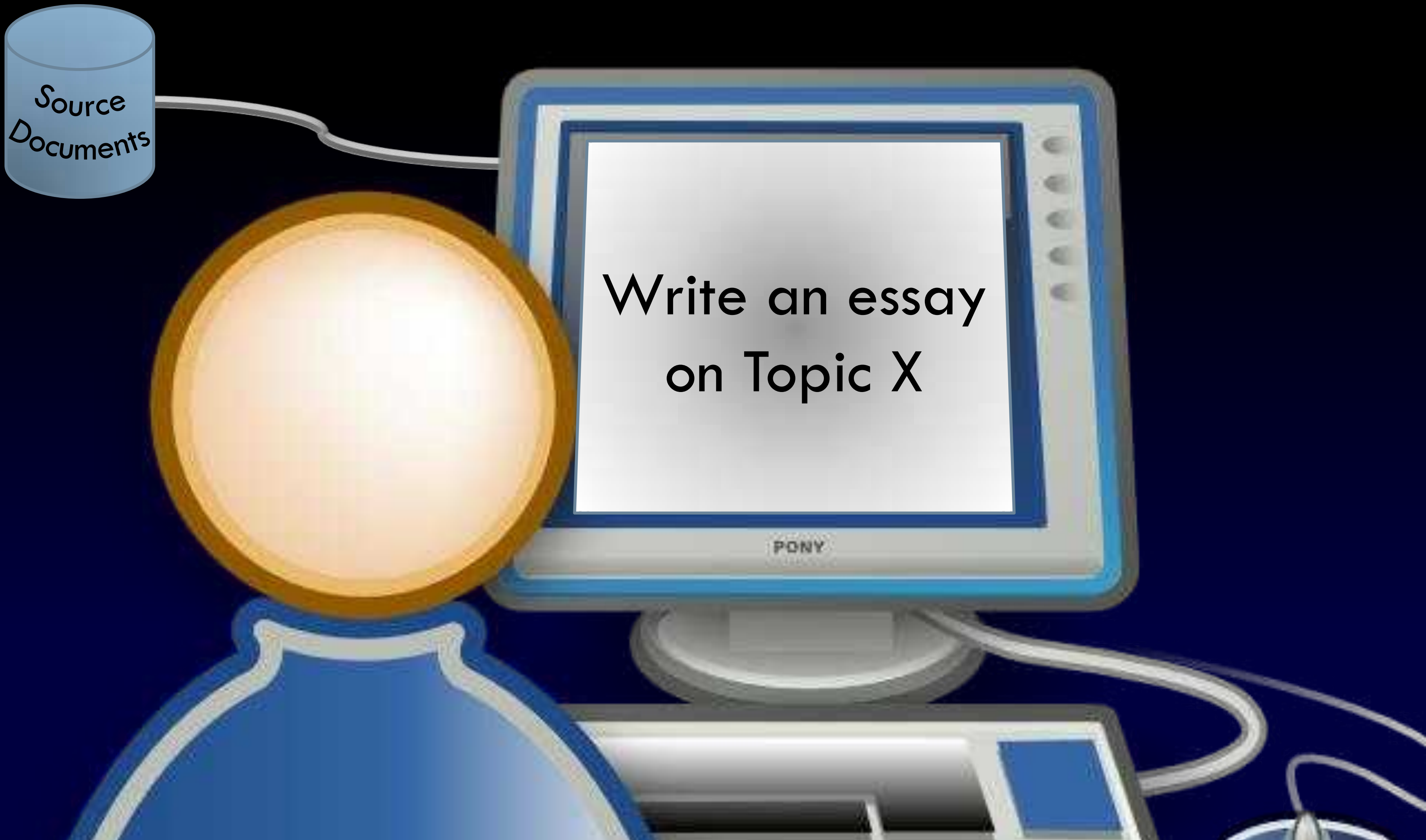
Building Plagiarism Detection Corpora

Manual Approach
(Crowdsoursing)



Building Plagiarism Detection Corpora

Manual Approach

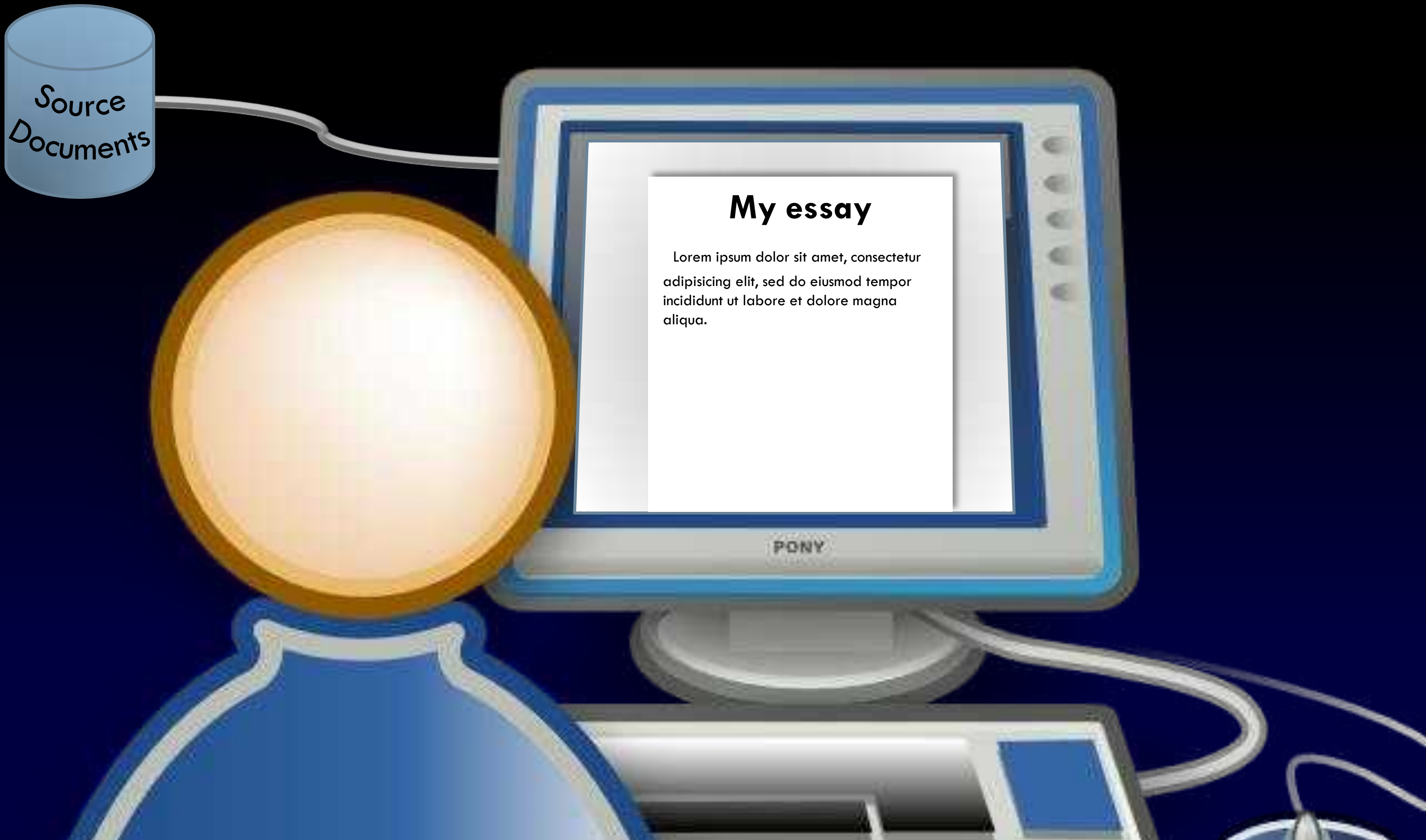


Source Documents

Write an essay on Topic X

Building Plagiarism Detection Corpora

Manual Approach

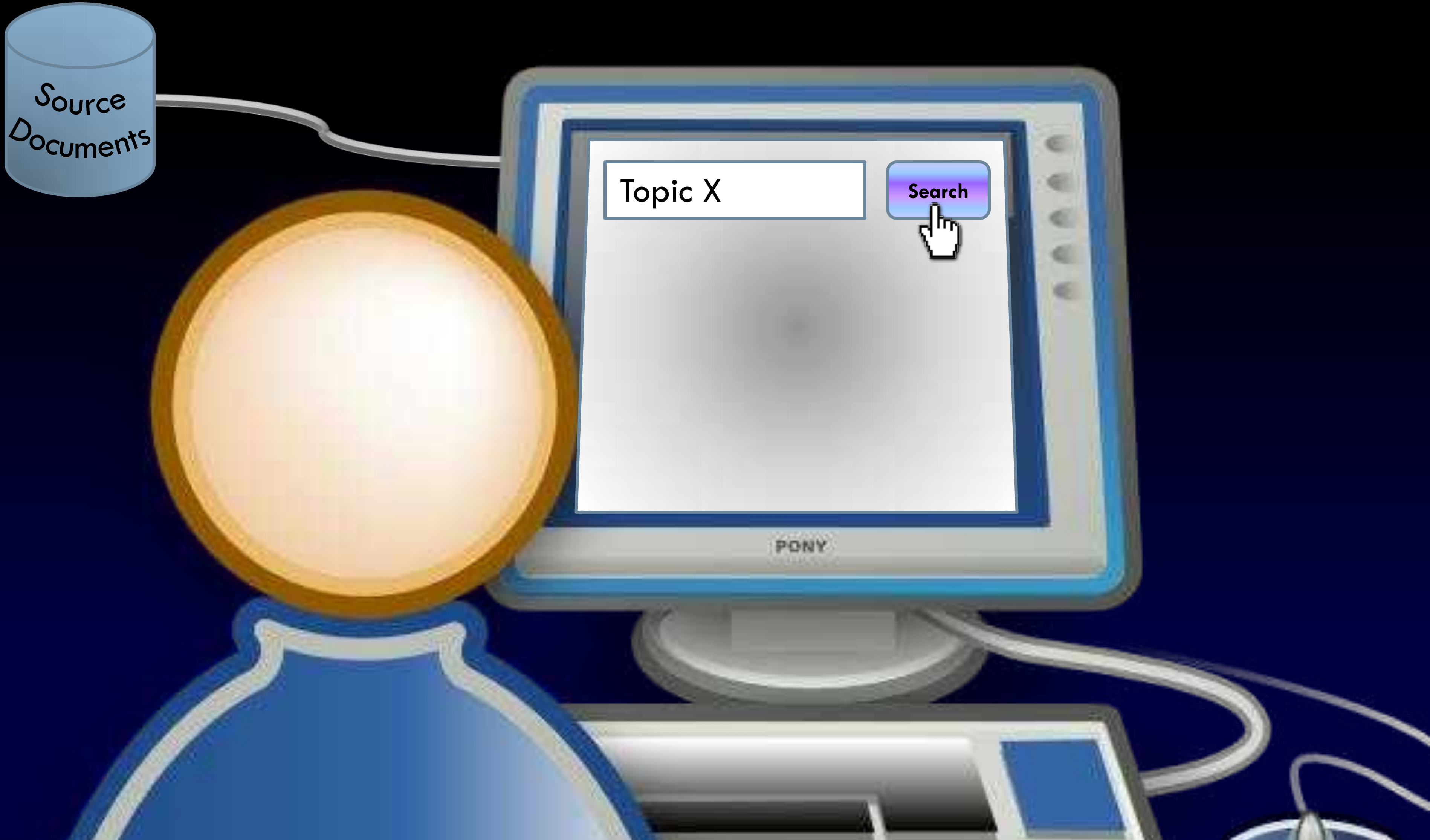


My essay

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

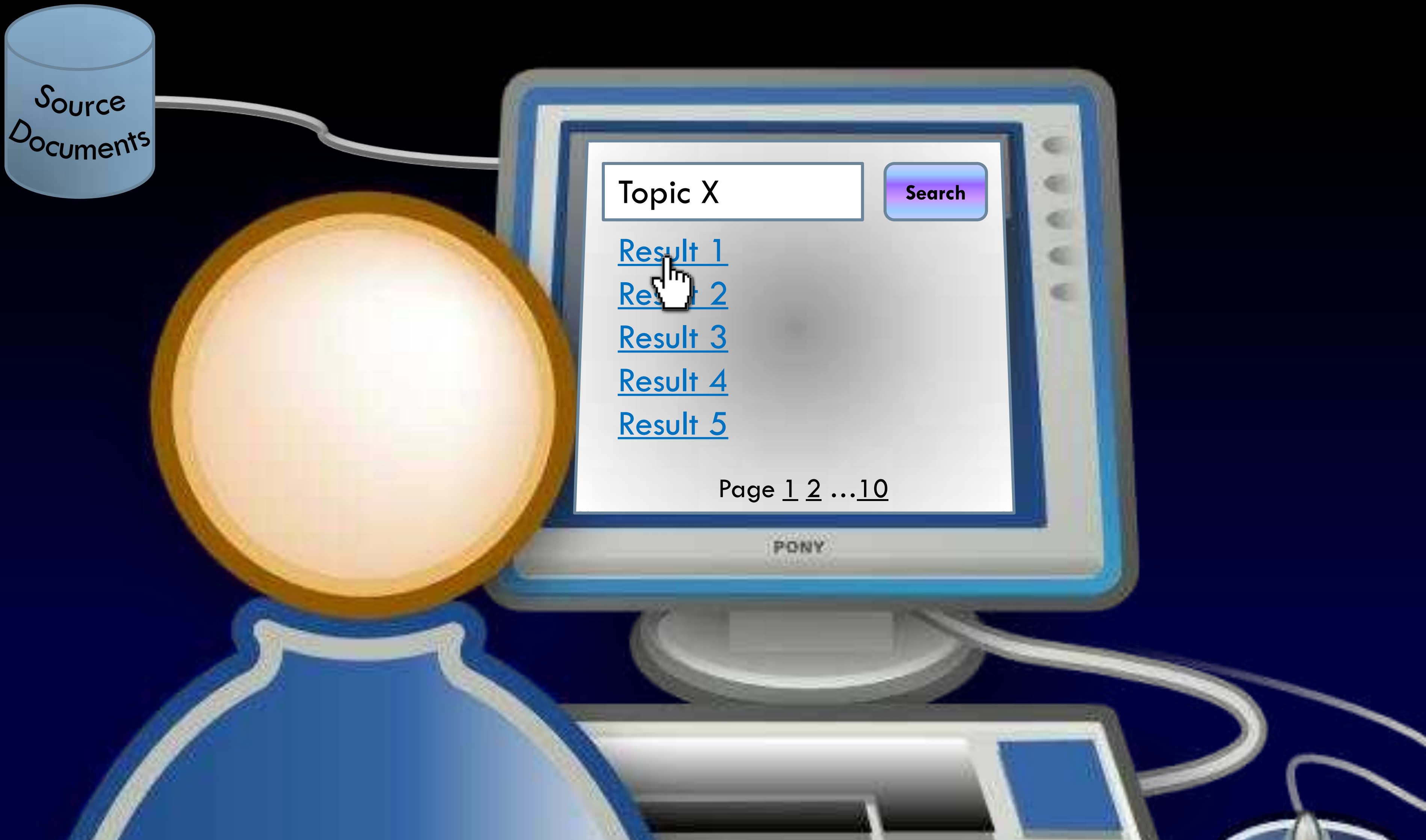
Building Plagiarism Detection Corpora

Manual Approach



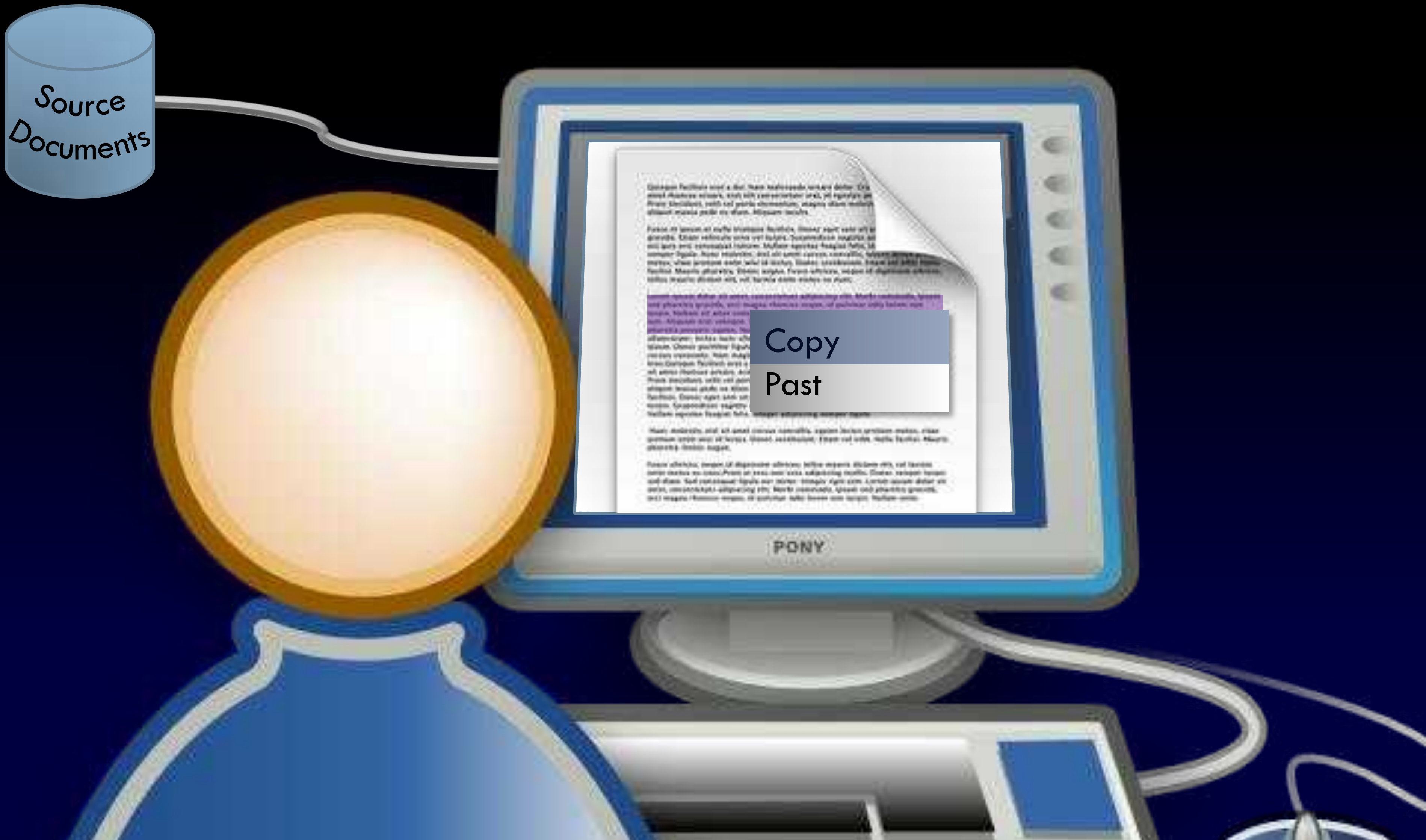
Building Plagiarism Detection Corpora

Manual Approach



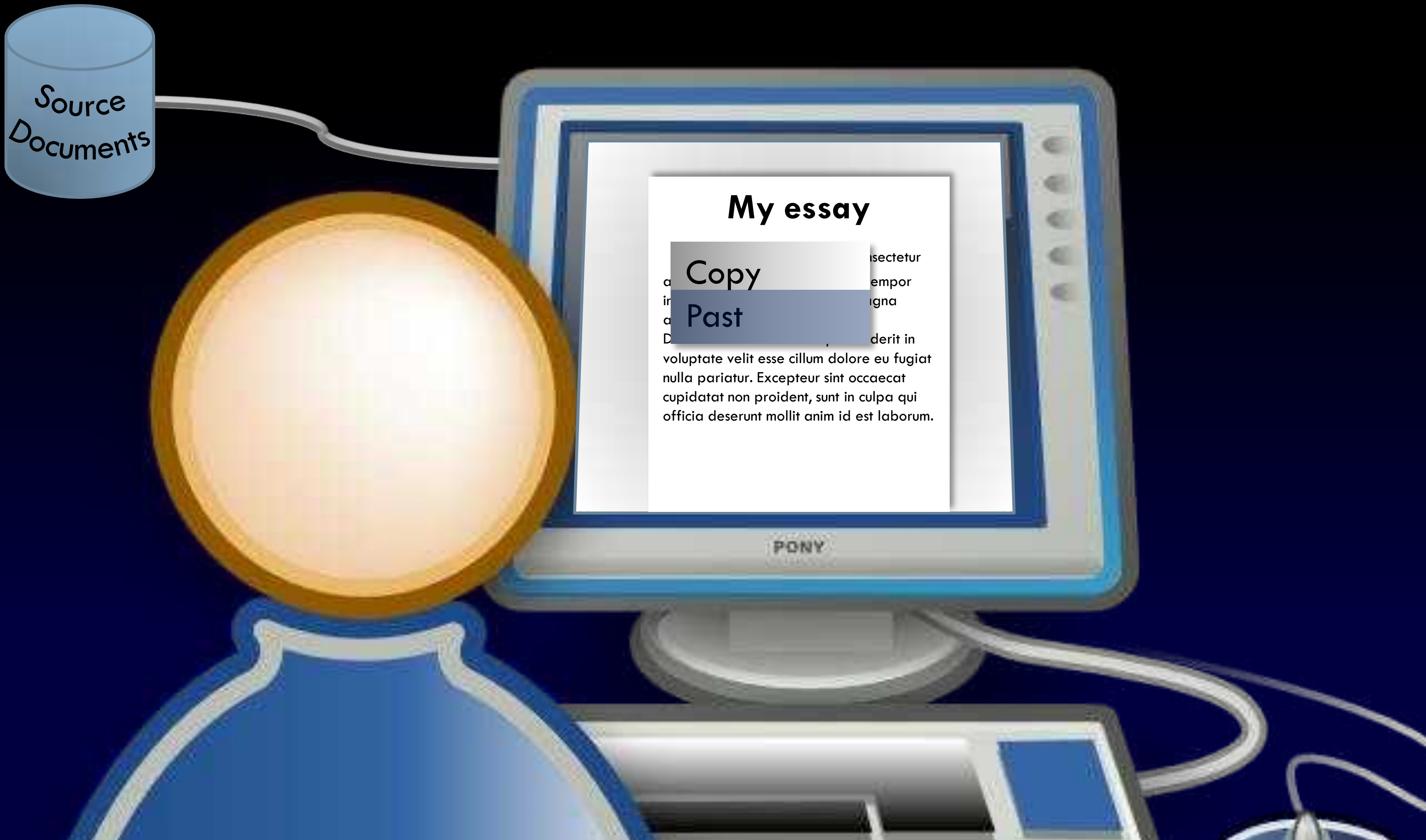
Building Plagiarism Detection Corpora

Manual Approach



Building Plagiarism Detection Corpora

Manual Approach



My essay

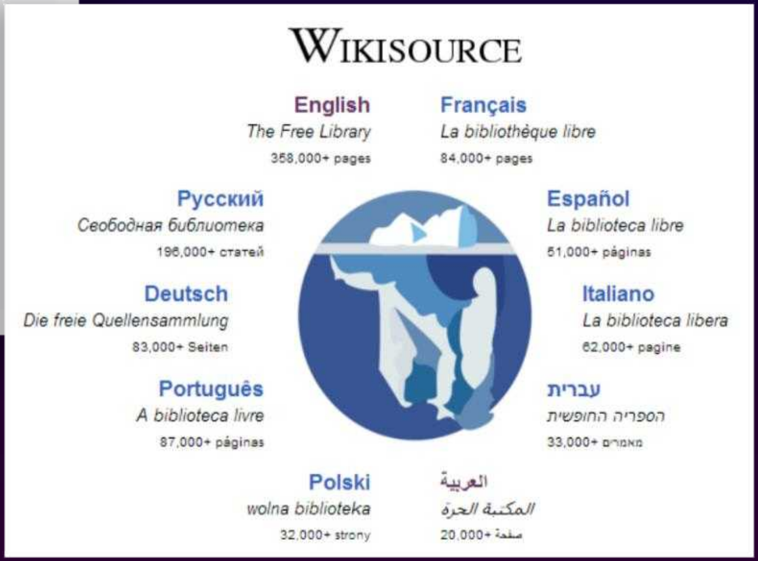
sectetur
a empor
ir gna
a
D derit in
voluptate velit esse cillum dolore eu fugiat
nulla pariatur. Excepteur sint occaecat
cupidatat non proident, sunt in culpa qui
officia deserunt mollit anim id est laborum.

Building Plagiarism Detection Corpora
Comparison of Approaches

	Automatic approach	Manual approach
Document length variety	✓	✗
Real plagiarism scenario ?	✗	✓
Copyright issue on suspicious documents	✗	✓
Cost (Material and human resource)	✓	✗

InAra: the First Corpus for the Evaluation Arabic Intrinsic Plagiarism Detection

- □ Built using the automatic approach.
- □ Target and source document were collected from Wikisource.
- □ Copyright-free documents.
- □ Documents tagged with topics and author name.



Criteria of Target Documents

Criteria to not complicate the detection task further more

- □ Written by one author only: ~~Wikipedia~~
- □ Should not include much of text reuse or many quotations: ~~News stories~~, ~~religion books with many quotations~~

Criteria to obtain reliable writing style model

- □ Should not be too short: ~~News stories~~
- □ Should be well edited: ~~texts without punctuation~~

Problem Encountered when Building the Corpus

- Limited copyright-free source of text
- Poor quality of the online free texts: not well punctuated and edited
- A high percentage from the free text are books on religious topics => contain many quotations => not very suited for a corpus to evaluate the intrinsic approach

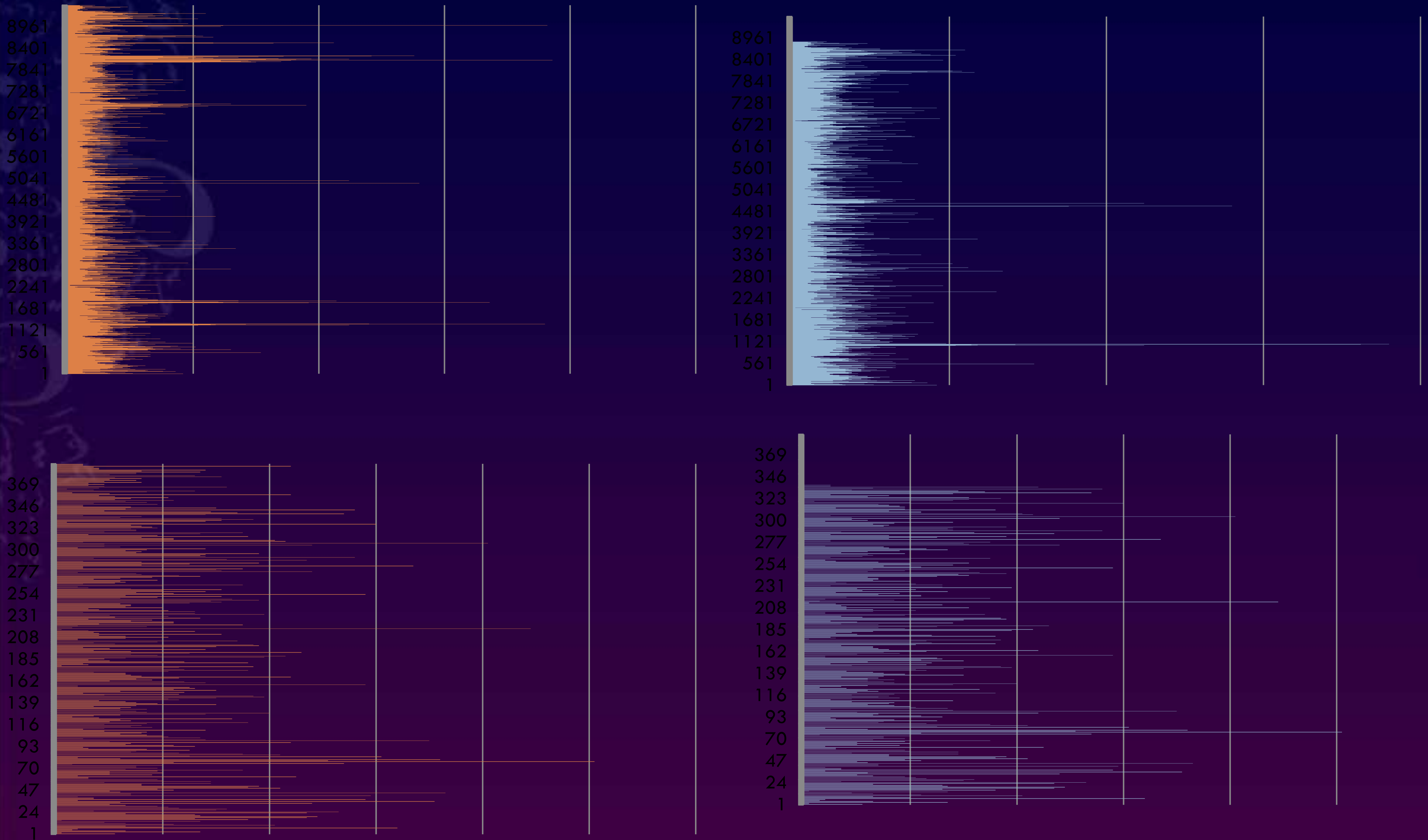
Why a Corpus in Arabic ?

- □ The ability to test language-dependent methods i.e. that take into account Arabic peculiarities (e.g. diacritics) or based on processing like parsing, stemming…etc

- □ Writing style discriminators in English are not necessarily discriminator in Arabic, e.g. average word length (Bensalem et al. 2012) (The next slide for other experiment on the sentence length)

- □ The importance of corpora to foster research

Visualization of Sentence length of some documents before and after the insertion of plagiarism (Sentence length does not seem to be a good discriminator in Arabic text

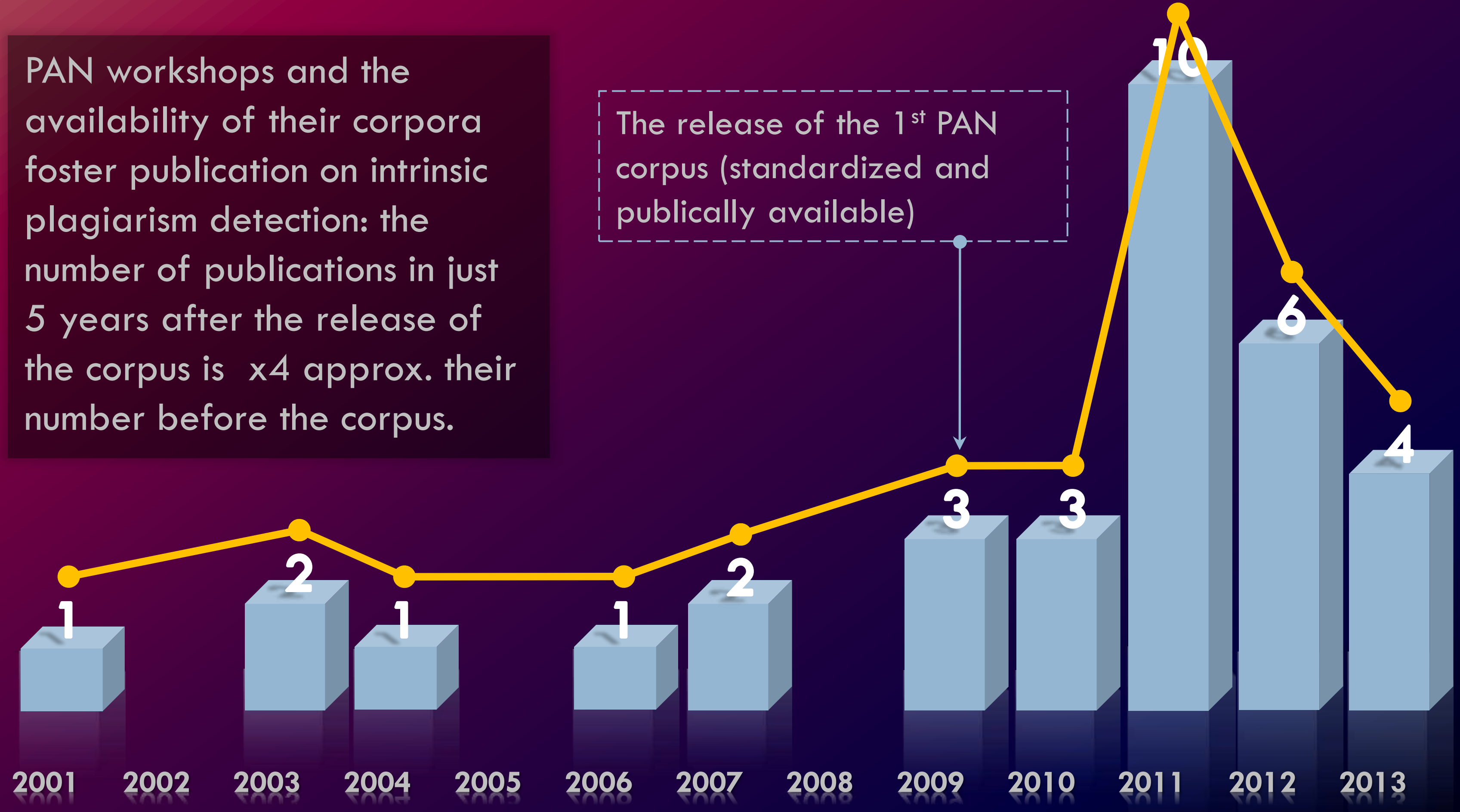


Why a Corpus in Arabic ?

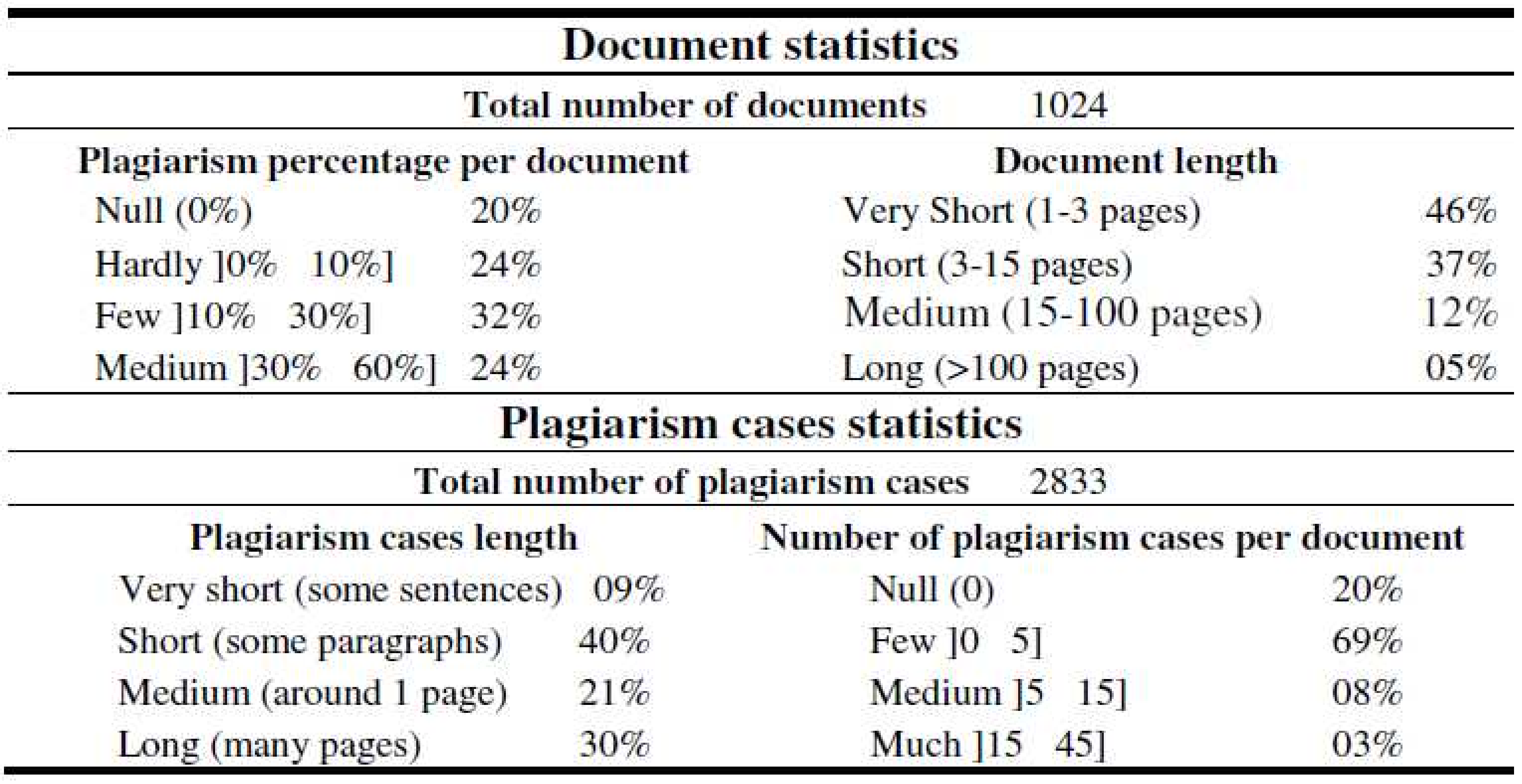
- □ The ability to test language-dependent methods i.e. that take into account Arabic peculiarities (e.g. diacritics) or based on processing like parsing, stemming…etc
- □ Writing style discriminators in English are not necessarily discriminator in Arabic, e.g. average word length (Bensalem et al. 2012)
- □ The importance of the corpus to foster research (The next slide)

Number of Publications on Intrinsic Plagiarism Detection from 2001 to 2013

PAN workshops and the availability of their corpora foster publication on intrinsic plagiarism detection: the number of publications in just 5 years after the release of the corpus is x4 approx. their number before the corpus.

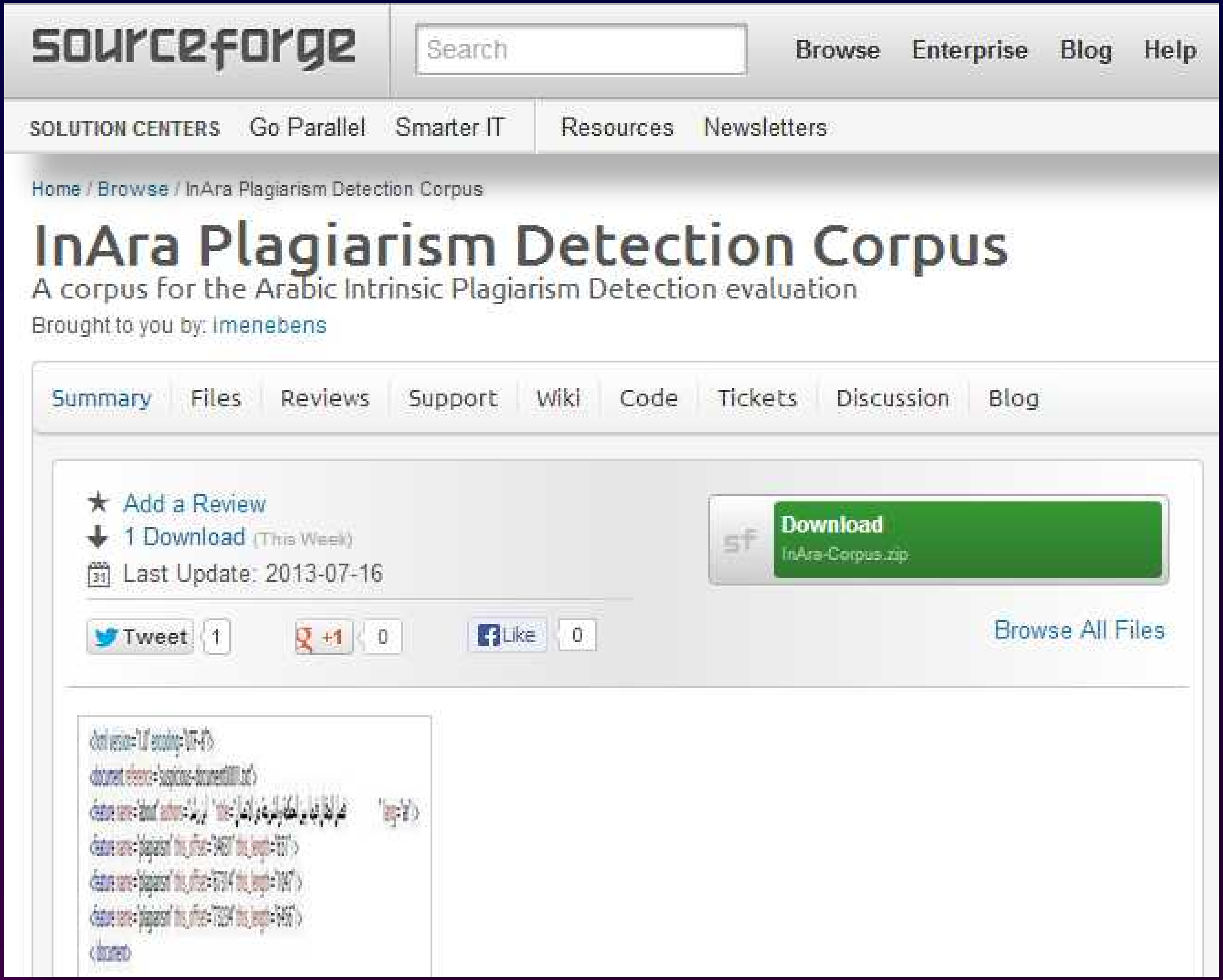

The release of the 1st PAN corpus (standardized and publically available)

InAra Corpus Statistics



Document statistics			
Total number of documents	1024		
Plagiarism percentage per document		**Document length**	
Null (0%)	20%	Very Short (1-3 pages)	46%
Hardly]0% 10%]	24%	Short (3-15 pages)	37%
Few]10% 30%]	32%	Medium (15-100 pages)	12%
Medium]30% 60%]	24%	Long (>100 pages)	05%
Plagiarism cases statistics			
Total number of plagiarism cases	2833		
Plagiarism cases length		**Number of plagiarism cases per document**	
Very short (some sentences)	09%	Null (0)	20%
Short (some paragraphs)	40%	Few]0 5]	69%
Medium (around 1 page)	21%	Medium]5 15]	08%
Long (many pages)	30%	Much]15 45]	03%

How to Get InAra Corpus ?



Home / Browse / InAra Plagiarism Detection Corpus

InAra Plagiarism Detection Corpus

A corpus for the Arabic Intrinsic Plagiarism Detection evaluation

Brought to you by: Imenebens

★ Add a Review

1 Download (This Week)

Last Update: 2013-07-16

```
<?xml version="1.0" encoding="UTF-8"?>
<document reference="suspicious-document[illegible].txt">
<feature name="about" author="[illegible]" title="[illegible]" lang="ar"/>
<feature name="plagiarism" this_offset="[illegible]" this_length="[illegible]"/>
<feature name="plagiarism" this_offset="[illegible]" this_length="[illegible]"/>
<feature name="plagiarism" this_offset="[illegible]" this_length="[illegible]"/>
</document>
```

http://sourceforge.net/projects/inaracorpus

Conclusion

Intrinsic plagiarism detection is a challenging task that need more attention from the research community

We hope InAra corpus will help to foster research in this area in Arabic language

Thank you !

Questions ?