On Web-based Plagiarism Analysis

A. Kleppe, D. Braunsdorf, Chr. Lössnitz, S. Meyer zu Eissen

Bauhaus University Weimar
Web-based Information Systems

What is Plagiarism?

"Plagiarism refers to the use of another's ideas, information, language, or writing, when done without proper acknowledgment of the original source"

[Wikipedia]

What is Plagiarism?

"Plagiarism refers to the use of another's ideas, information, language, or writing, when done without proper acknowledgment of the original source"

[Wikipedia]

Plagiarism analysis:

Given. A suspicious document.
Task. Find potentially copied parts, and provide references to original sources.

What is Plagiarism?

"Plagiarism refers to the use of another's ideas, information, language, or writing, when done without proper acknowledgment of the original source"

[Wikipedia]

Plagiarism analysis:

Given.	A suspicious document.
Task.	Find potentially copied parts, and provide references to original sources.

Fact: About 50% of the students admit to plagiarize from Internet documents (study on 18,000 students).

[McCabe 2005]

Current Research on Plagiarism Analysis

Current research is corpus-oriented.

Given. A suspicious document
and a corpus of original documents.
Task. Find potentially copied parts *in the corpus*,
and provide references to original sources.

Current Research on Plagiarism Analysis

Current research is corpus-oriented.

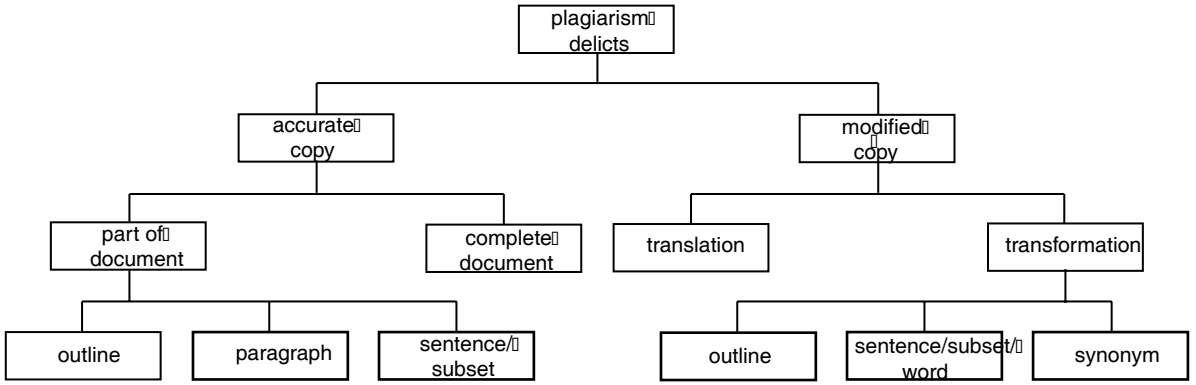
Given. A suspicious document
and a corpus of original documents.
Task. Find potentially copied parts *in the corpus*,
and provide references to original sources.

Research questions:

- How can a corpus of potentially original documents be constructed from the Web?
- Can plagiarized parts be detected *without* a corpus?

Plagiarism Forms

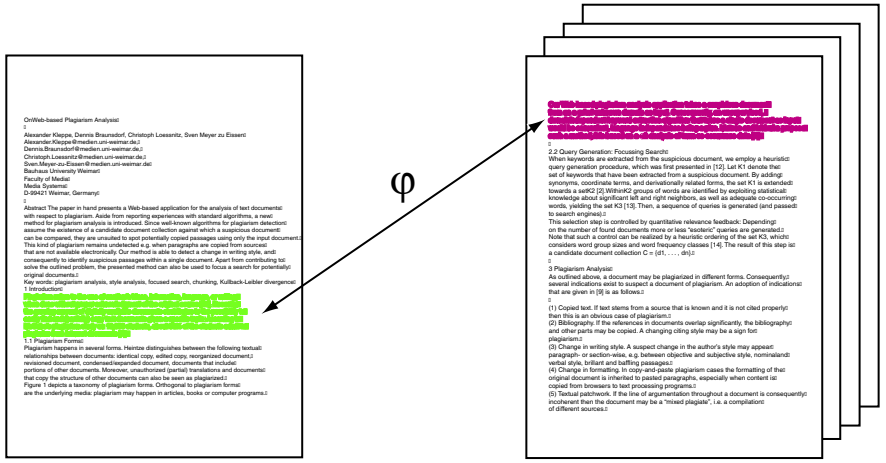
Plagiarism may happen in manyfold variants:



Plagiarism Analysis against a Corpus (1)

Intuition:

- Partition each document in meaningful units (“chunks”).
- Compare them with a similarity function φ (pairwise).

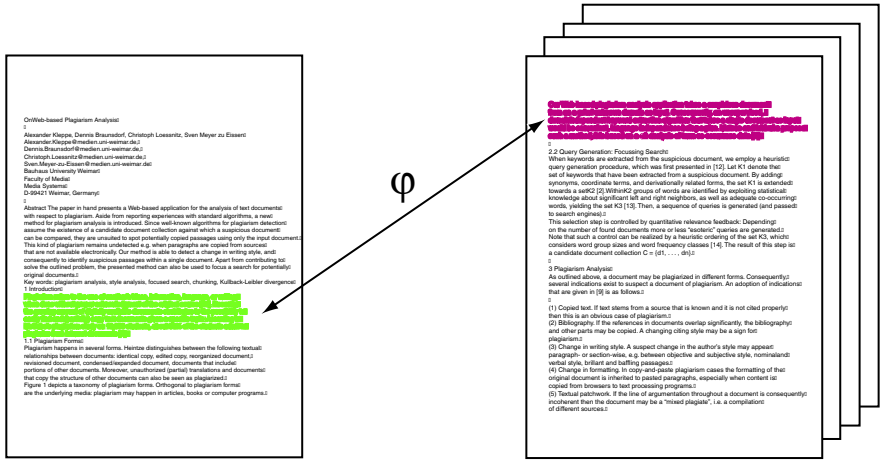


suspicious document corpus documents

Plagiarism Analysis against a Corpus (1)

Intuition:

- Partition each document in meaningful units (“chunks”).
- Compare them with a similarity function φ (pairwise).



suspicious document corpus documents

Complexity:

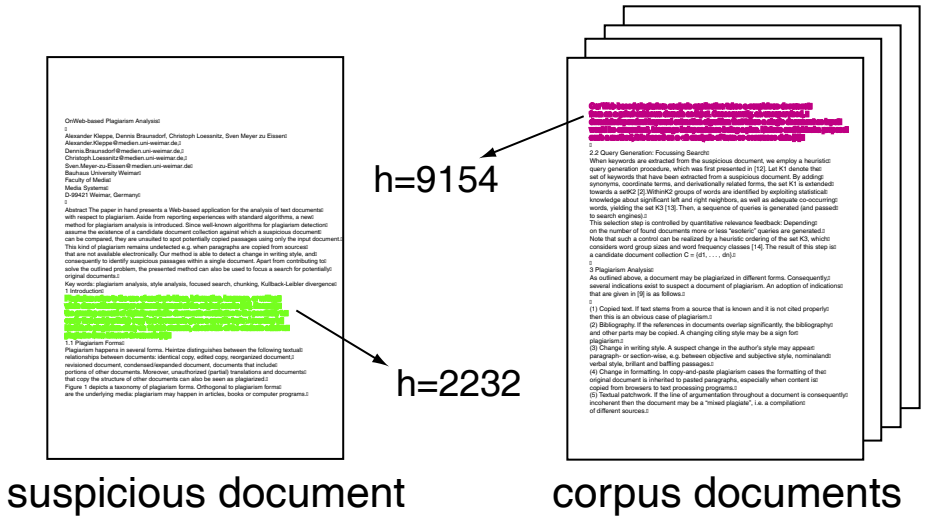
n documents in corpus, c chunks per document on average

➜ $O(n \cdot c^2)$ comparisons

Plagiarism Analysis against a Corpus (2)

Text comparison with digital fingerprints:

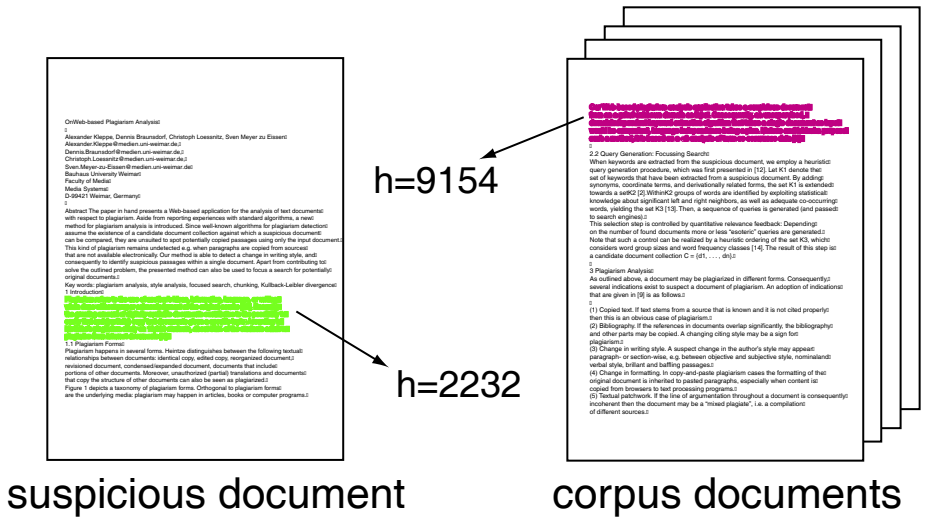
- Partition each document in meaningful units (“chunks”).
- Compute fingerprints of the chunks using a hash function h.
- Put all hashes into a hashtable. A collision indicates matching chunks.



Plagiarism Analysis against a Corpus (2)

Text comparison with digital fingerprints:

- Partition each document in meaningful units (“chunks”).
- Compute fingerprints of the chunks using a hash function h.
- Put all hashes into a hashtable. A collision indicates matching chunks.



Complexity:

n documents in corpus, c chunks per document on average

➜ $O(n \cdot c)$ operations (fingerprint generation, hashtable operations).

Plagiarism Analysis against a Corpus (3)

Discussion :

- Hashing is fast, but sensitive to (even small) changes:
  $h(c_1) = h(c_2) \Rightarrow c_1 = c_2$ (with very high probability).
- Pairwise comparisons based on similarity-function φ are too expensive.

➜ Past research focussed on chunking strategies.

Plagiarism Analysis against a Corpus (3)

Discussion :

- Hashing is fast, but sensitive to (even small) changes:
  $h(c_1) = h(c_2) \Rightarrow c_1 = c_2$ (with very high probability).
- Pairwise comparisons based on similarity-function φ are too expensive.

➜ Past research focussed on chunking strategies.

Current research:

- Focus on *fuzzy* hash functions h_F:
  $h_F(c_1) = h_F(c_2) \Rightarrow \varphi(c_1, c_2) \geq 1 - \varepsilon$ [Stein 2005]
- Fuzzy hash functions allow for big chunk sizes (speed-up) and are not sensitive to changes.

Indications for Plagiarism

Text similarity is not the only indication for plagiarism.

Indications include:

- Changes in formatting.
- *Changes in writing style*.
- Broken argumentation.
- Inconsistent spelling.
- Outmoded diction.

These indications can be detected (by humans) without corpora.

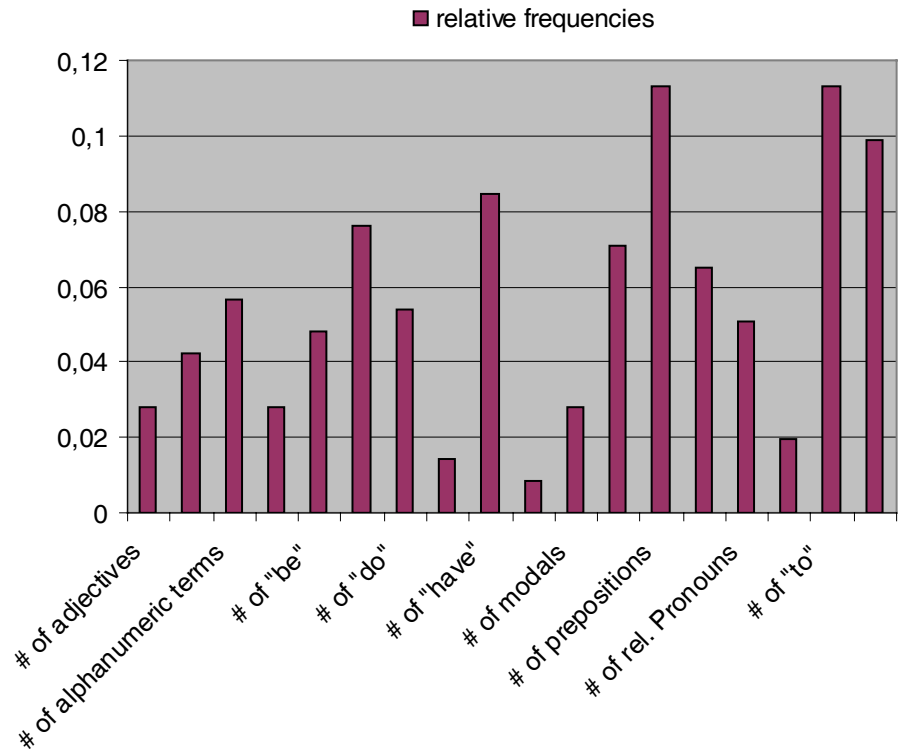
➜ How can we operationalize the detection of these indications?

Style Analysis

Q: How can writing style changes be measured?

A: Not directly, but divergences of word class distributions give hints.

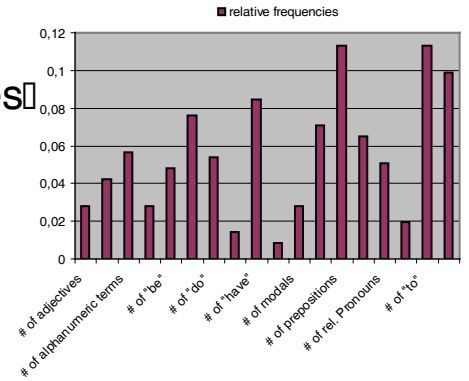
Word class use in a document:



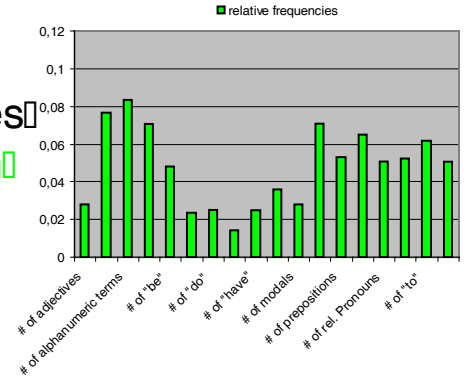
Style Analysis

Idea: Compare the word class distribution of each paragraph to the distribution of the entire document.

word-class frequencies for the document (global distribution)



word-class frequencies for a single paragraph (local distribution)



Style Analysis

The divergence can be measured by means of the Kullback-Leibler divergence.

Let W denote the set of word classes, let $w \in W$,

let $p(w)$ denote local word class frequencies,

let $q(w)$ denote the global word class frequencies.

Style Analysis

The divergence can be measured by means of the Kullback-Leibler divergence.

Let W denote the set of word classes, let $w \in W$,

let $p(w)$ denote local word class frequencies,

let $q(w)$ denote the global word class frequencies.

The Kullback-Leibler divergence measure is defined as

$$KL_W(p,q) = \sum_{w \in W} p(w) \log \frac{p(w)}{q(w)} = H(p,q) - H(p) \in \mathbf{R}_0^+$$

If $KL_W(p,q)$ is significant then the paragraph that is associated with p may be copied.

We found KL_W to work very well when single paragraphs are copied from one document to another.

Web-based Plagiarism Analysis

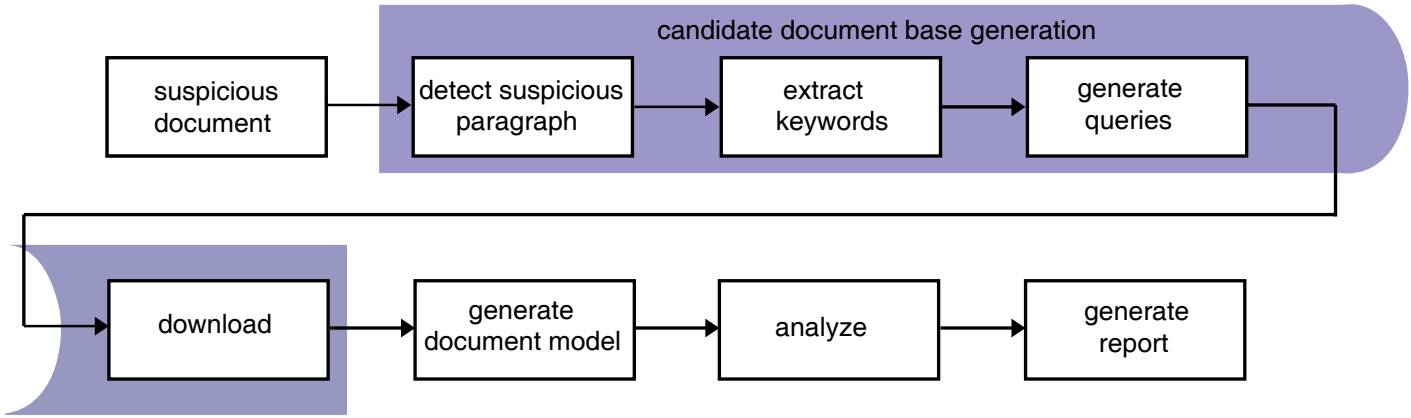
Given. A suspicious document,
and the Web as corpus of original documents.

Task. Generate a *candidate document base* from the Web, find potentially copied parts in the base documents, and provide references to original sources.

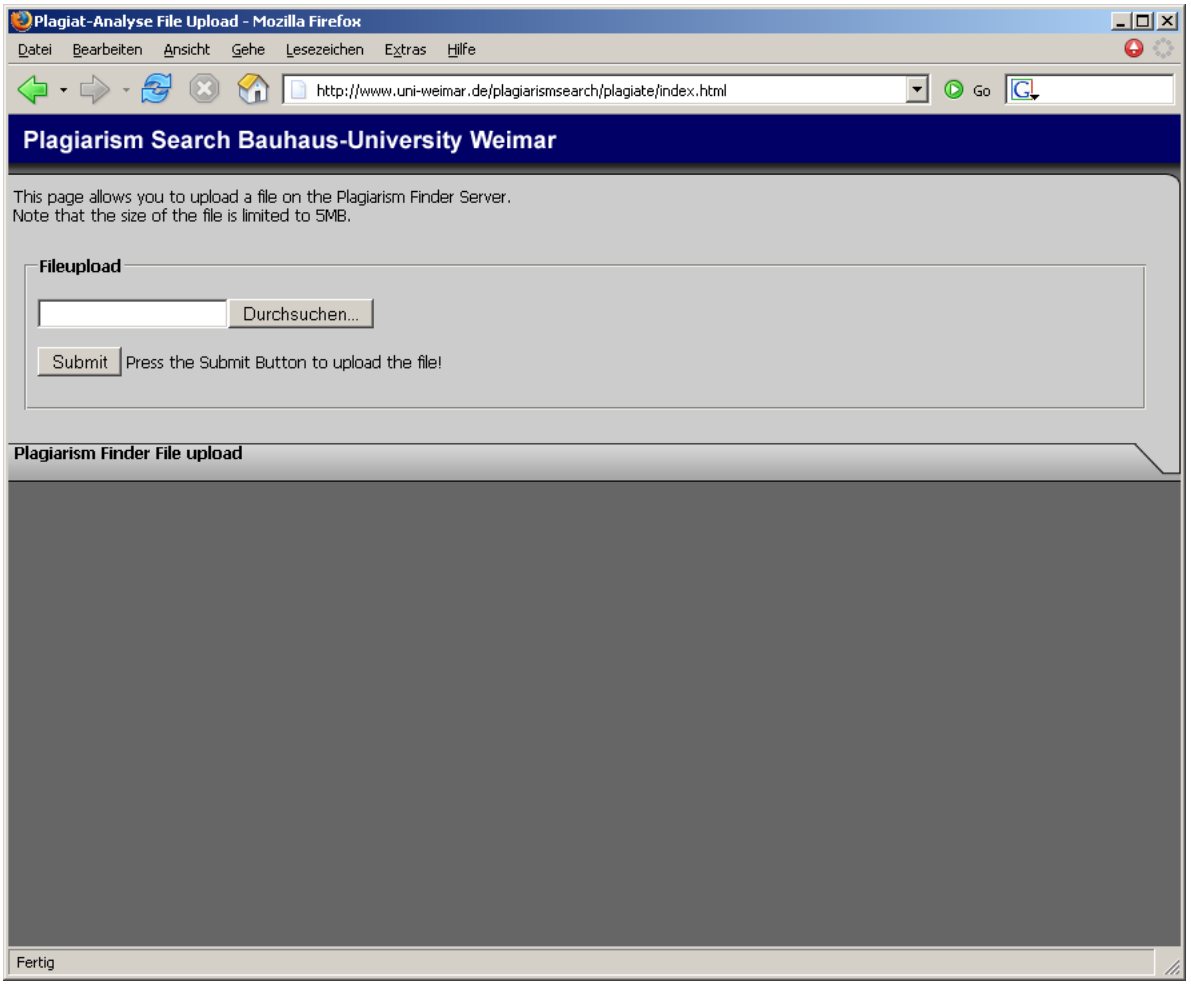
Web-based Plagiarism Analysis

Given. A suspicious document,
and the Web as corpus of original documents.

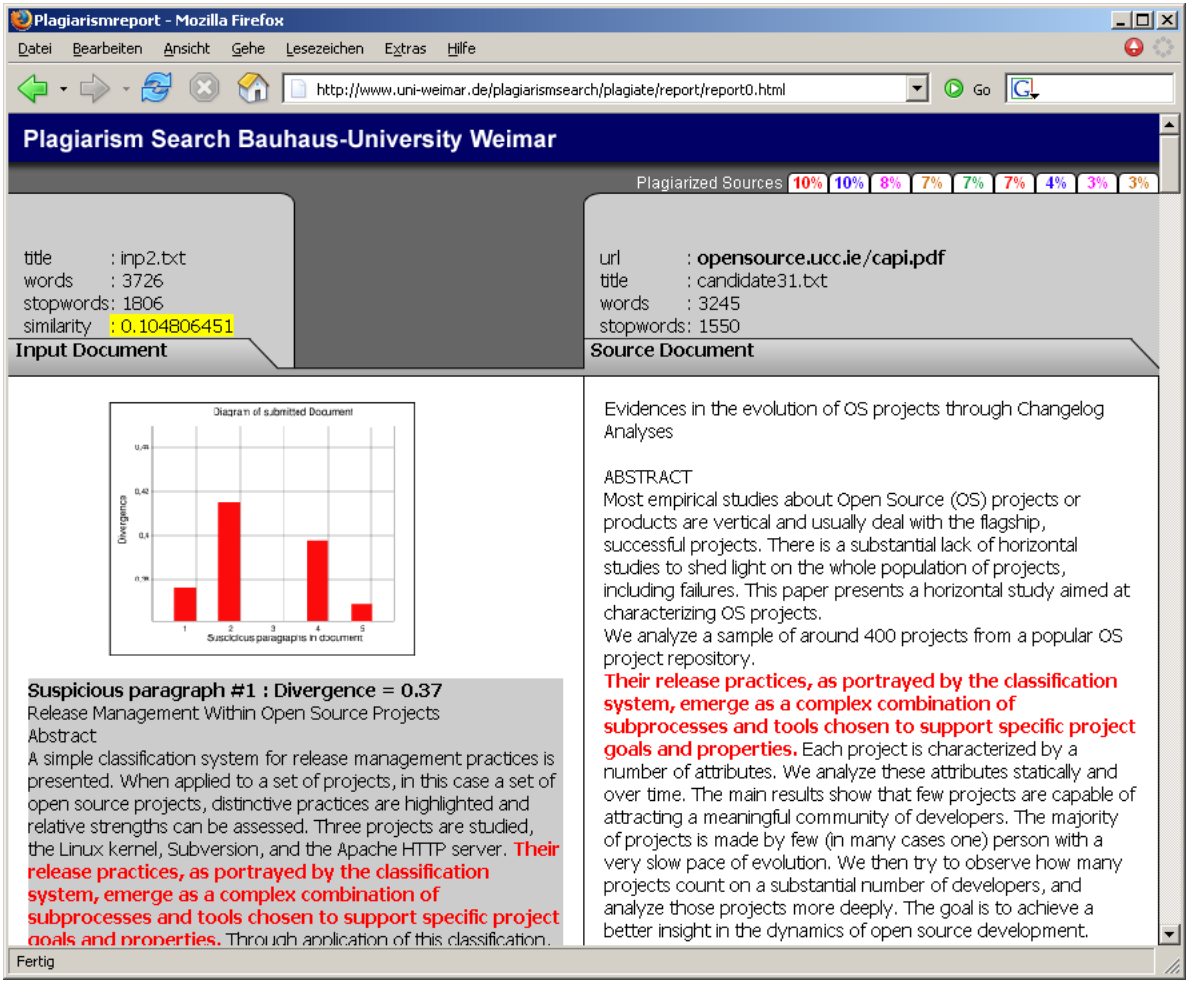
Task. Generate a *candidate document base* from the Web, find potentially copied parts in the base documents, and provide references to original sources.



Prototypic Implementation



Prototypic Implementation



Thank You!

Questions?