Algorithms identifying online plagiarism

Ondrej Vesely,
Mendel University in Brno

Mendel University in Brno

Story

Diploma thesis

IPPHEAE

SEZNAM.CZ

Russia, South Korea, China, Japan, Czech republic.

Algorithms

Just Google it, but...
Just one query per second

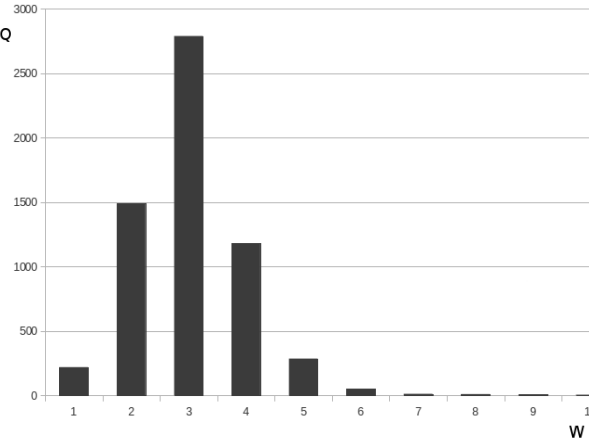
How to find the right length of fragment?

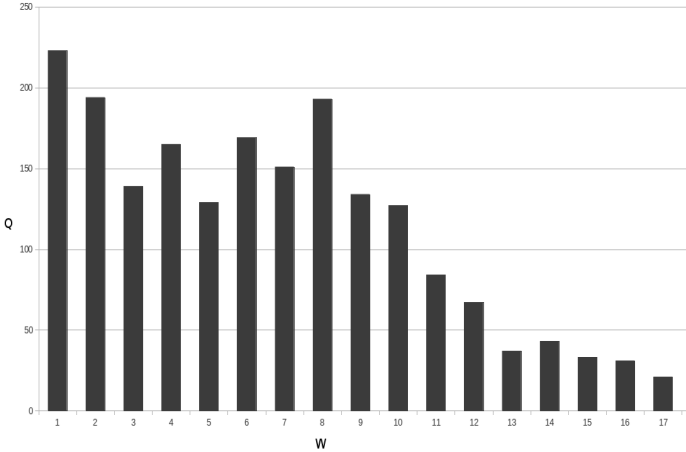
Which fragments should be check first?

How to find the right length of fragment?

- naive
- heuristics
- neural network

(grams –> bigrams)

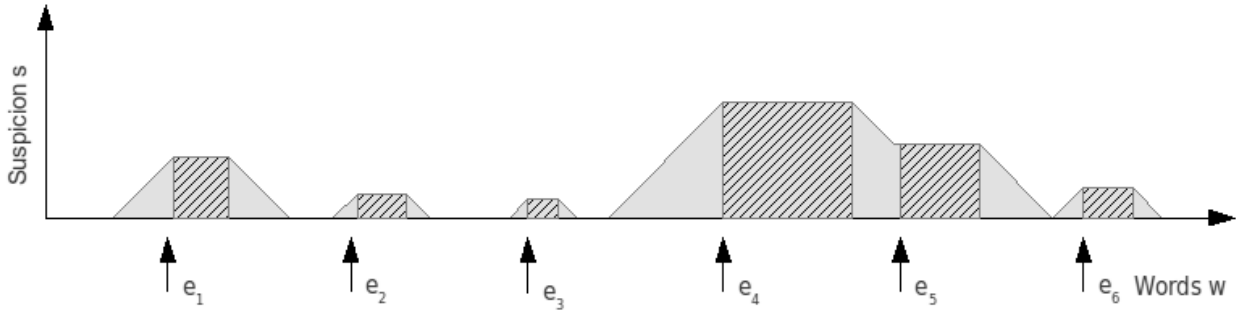




Which fragments should be checked first?

presuspicious index

typical characteristics of plagiarised fragments?



Results & status

- It works! … as a technology demo
- Integrating to university IS
- Sensitivity as good as TurnItIn

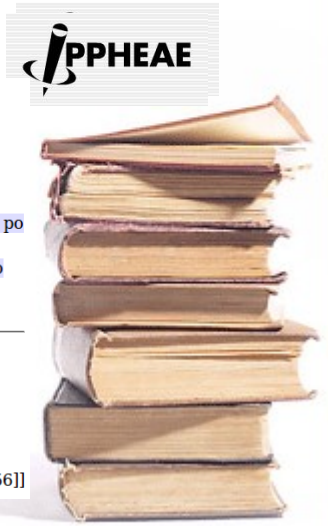
Diplomka: analýza podobnosti závěrečných prací a on-line dokumentů[16] Zadání Navrhněte a implementujte systém pro srovnávání obsahu zadaného[16] dokumentu s dokumenty publikovanými na českém internetu a hledání obsahově[16] podobných fragmentů textu. Práci koncipujte tak, aby byla použitelná pro rozšíření antiplagiátorského modulu v UIS[16] MENDELU. Součástí práce bude analýza časové a prostorové složitosti a návrh způsobů jak tuto složitost[16] snížit. Proveďte analýzu současného stavu řešené problematiky a srovnejte váš[16] systém s jinými dostupnými řešeními „Obě vozidla skončila po srážce mimo vozovku v poli, osobní vozidlo převrácené na střechu. Na místo události vyjížděli hasiči v 10:40 hodin,“[20][20] uvedla mluvčí hasičů Martina Žahourková. Dodala,[20] že hasiči po příjezdu na místo vyprostili řidiče osobního vozu, který utrpěl smrtelné zranění. Řidič kamiónu byl převezen záchrannou službou do nemocnice.[11][11]

Strategie WORDS:

- zdroj: http://www.novinky.cz/krimi/277730-pri-s, index: **57**, slova [[92, 99], [82, 93], [72, 93]]
- zdroj: http://www.hzscr.cz/clanek/ridic-skodovk, index: **36**, slova [[112, 123], [102, 123]]

Strategie PHRASE:

- zdroj: http://www.antiplagiator.cz/, index: **46**, slova [[62, 66], [37, 41], [22, 26], [12, 16], [2, 7], [52, 56]]



Thank you. Questions?

cz.linkedin.com/in/veselyondrej

xorwen@gmail.com

twitter.com/xorwen

www.facebook.com/orwen