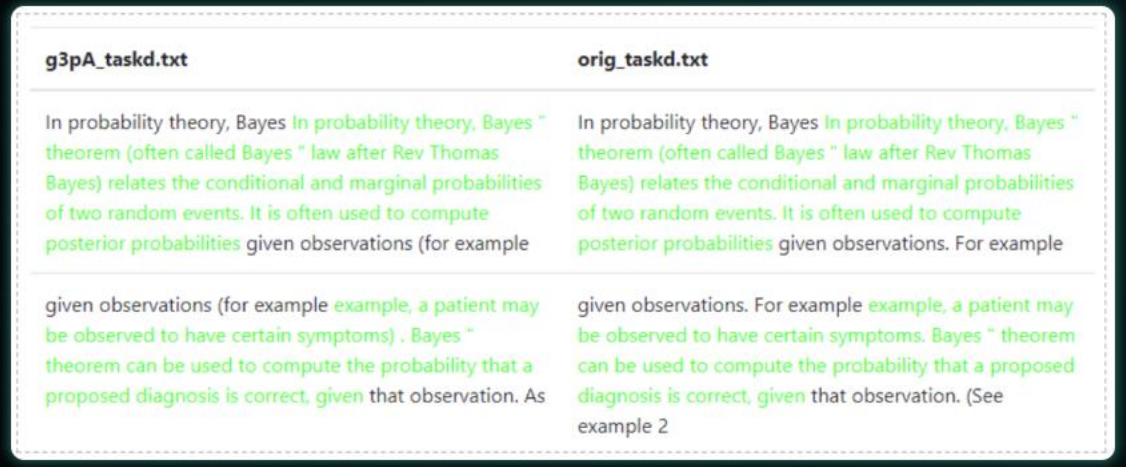
machine learning
excellence delivered

Plagiarism detection system

Industries: Education, Text analysis, SaaS



g3pA_taskd.txt	orig_taskd.txt
In probability theory, Bayes In probability theory, Bayes ' theorem (often called Bayes ' law after Rev Thomas Bayes) relates the conditional and marginal probabilities of two random events. It is often used to compute posterior probabilities given observations (for example	In probability theory, Bayes In probability theory, Bayes ' theorem (often called Bayes ' law after Rev Thomas Bayes) relates the conditional and marginal probabilities of two random events. It is often used to compute posterior probabilities given observations. For example
given observations (for example example, a patient may be observed to have certain symptoms) . Bayes ' theorem can be used to compute the probability that a proposed diagnosis is correct, given that observation. As	given observations. For example example, a patient may be observed to have certain symptoms. Bayes ' theorem can be used to compute the probability that a proposed diagnosis is correct, given that observation. (See example 2

Technology stack: Python, NLP, FastAPI, PostgreSQL, Celery, Redis, TET, Docker

The solution is able to process several hundreds of exam scripts in hours, finding for the client staff only few highly similar pairs of exam scripts for manual inspection.
The antiplagiarism solution based on the document processing and analysis platform. It consisted of a custom preprocessing of PDF formats, text comparison engine, and visualization of results for analysis.
Because the exam scripts may contain legitimate common pieces, the final solution acts more like a filter: it locates suspicious pairs of documents that reach a specified similarity threshold. After such filtering, it only takes a couple of hours for a human professional to manually review 20-30 pairs of documents.

The link to our plagiarism detection system: https://silkdata.tech/products/plagiarism-detection-system

Plagiarism detection system

- Quarantines, lockdown, online-first in 2020
- How to ensure no exam essays are not plagiarized
- Requirements
  - Fast processing of a single exam
  - Few hundreds of exam scripts (950 – 2400) per exam
  - Each exam script is about 15 – 45 pages
  - From business perspective, identical half of a (full!) page can be plagiarism
  - Possible paraphrases
  - Lots of repeated text (conversation formulas, legal text)
- Not addressed by other solutions

Possible billions of page pairs

Plagiarism detection system

- Special preprocessing (footers/headers, almost empty pages, …)
- Identification on full duplicates (MD5 hash)
- Based on common word triplets
- For performance, documents with similarity below 50%, only approx. algorithm was used.
- Doc similarity is maximal similarity between pages
- Special algorithm for OCRed essays
- Acts as a filter and finds few candidates for manual inspection
- Analysis time: typically, < 1h on our servers
- Customer happy ;)
- Case study: https://www.silkdata.ai/products/compare-text/apt-case-study

Plagiarism detection system: Plans

Performance

- Using online text collections (big data)
- Visualization improvements (requested)
- Extra languages
- Paraphrase detection
- Multi-language matching
- Source-code plagiarism (CS students)

Integrations

- Reporting system
- Moodle plugin
- DMS integrations (standards)
- Web site checks (SEO tasks)
- Self-service support

www.silkdata.tech

info@silkdata.tech

+48 452 380 167

Poland

Silk Data sp. z o o.
Domaniewska 17/19
off. 13302-672 Warsaw
Poland +48 452 380 167

Germany

SilkCode GmbH
Luisenstraße 62
D-47799 Krefeld
Germany +49 2151 387 3531