AudiTREE

Albuddy's Solution for:
BPK HACKATHON 2020

Jan 13 2021

Character Introduction

Audi is a professional auditor

He works more than 8 years in a well-known Audit Board in a country

During audit period, he spends a considerable amount of time and energy to process files and documents from auditee

He thinks whether his works should always last like this...

Then he meets his pal, Albuddy ..

He states his problems are:

- His physical files and documents are **SCATTERED** everywhere.
- He mostly uses his intuitive on **TEXT REPORTS** for Compliance Analysis.
- He search files and documents based on **TOKEN MATCHING**.
- He find it so hard to predict fraud/**FAKE NUMERICAL** value.

Albuddy introduces AudiTREE

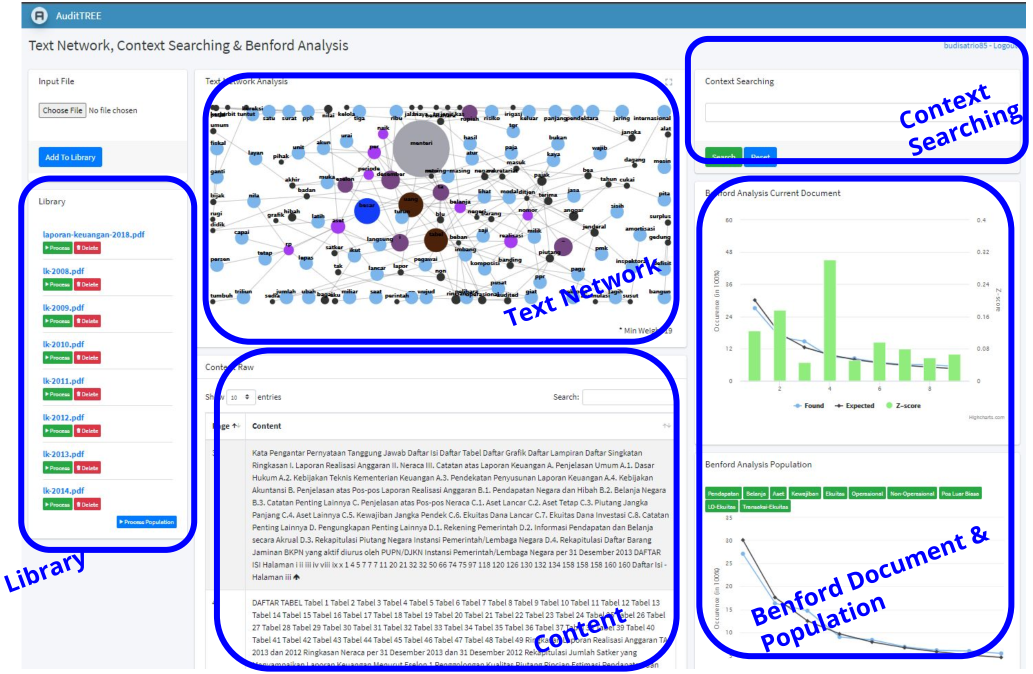
AudiTREE has the solutions of his problems:

- An **ELECTRONIC LIBRARY** of files and documents
- A text processor to enrich analysis with **NETWORK GRAPH VISUALIZATION**.
- A searching tool based on **CONTEXT MATCHING** of his files and documents.
- An intuitive chart and value of **BENFORD ALGORITHM** to predict fake numerical value.

Audi doing trial

- He gathers PDF document of Annual **FINANCIAL REPORT**S
- Documents are sampled from **21 MINISTRIES** where extractable
- He inserts financial report of a ministry from 2008 to 2014, and 2018 as his trial

Jos!, said Audi to the results



Text Network, Context Searching & Benford Analysis

Page	Content
3	Kata Pengantar Pernyataan Tanggung Jawab Daftar Isi Daftar Tabel Daftar Grafik Daftar Lampiran Daftar Singkatan Ringkasan I. Laporan Realisasi Anggaran II. Neraca III. Catatan atas Laporan Keuangan A. Penjelasan Umum A.1. Dasar Hukum A.2. Kebijakan Teknis Kementerian Keuangan A.3. Pendekatan Penyusunan Laporan Keuangan A.4. Kebijakan Akuntansi B. Penjelasan atas Pos-pos Laporan Realisasi Anggaran B.1. Pendapatan Negara dan Hibah B.2. Belanja Negara B.3. Catatan Penting Lainnya C. Penjelasan atas Pos-pos Neraca C.1. Aset Lancar C.2. Aset Tetap C.3. Piutang Jangka Panjang C.4. Aset Lainnya C.5. Kewajiban Jangka Pendek C.6. Ekuitas Dana Lancar C.7. Ekuitas Dana Investasi C.8. Catatan Penting Lainnya D. Pengungkapan Penting Lainnya D.1. Rekening Pemerintah D.2. Informasi Pendapatan dan Belanja secara Akrual D.3. Rekapitulasi Piutang Negara Instansi Pemerintah/Lembaga Negara D.4. Rekapitulasi Daftar Barang Jaminan BKPN yang aktif diurus oleh PUPN/DJKN Instansi Pemerintah/Lembaga Negara per 31 Desember 2013 DAFTAR ISI Halaman i ii iii iv viii ix x 1 4 5 7 7 7 11 20 21 32 32 50 66 74 75 97 118 120 126 130 132 134 158 158 158 160 160 Daftar Isi - Halaman iii
4	DAFTAR TABEL Tabel 1 Tabel 2 Tabel 3 Tabel 4 Tabel 5 Tabel 6 Tabel 7 Tabel 8 Tabel 9 Tabel 10 Tabel 11 Tabel 12 Tabel 13 Tabel 14 Tabel 15 Tabel 16 Tabel 17 Tabel 18 Tabel 19 Tabel 20 Tabel 21 Tabel 22 Tabel 23 Tabel 24 Tabel 25 Tabel 26 Tabel 27 Tabel 28 Tabel 29 Tabel 30 Tabel 31 Tabel 32 Tabel 33 Tabel 34 Tabel 35 Tabel 36 Tabel 37 Tabel 38 Tabel 39 Tabel 40 Tabel 41 Tabel 42 Tabel 43 Tabel 44 Tabel 45 Tabel 46 Tabel 47 Tabel 48 Tabel 49 Ringkasan Laporan Realisasi Anggaran TA 2013 dan 2012 Ringkasan Neraca per 31 Desember 2013 dan 31 Desember 2012 Rekapitulasi Jumlah Satker yang Menyampaikan Laporan Keuangan Menurut Eselon 1 Penggolongan Kualitas Piutang Rincian Estimasi Pendapatan

Audi finds AudiTREE really helps him accelerate his job

He tells his boss and colleagues about it and also proposes it as a central tools to be used when auditing

Audi now designs the Future Works

- **OPTIMIZE** the **CONTEXT** search by fine tuning on financial domain
- **ENRICH** data **GATHERING** to include scanned images using OCR
- Add **NER** (Named Entity Recognition) into Text Network Analysis

Ke Sabah membeli rotan,
pulang petang bawa selasih.
Ada salah mohon benarkan,
ini sekian **TERIMA KASIH**.

Technology Stack

- Python
- Flask
- AdminLTE by AppSeed
- Networkx
- Highcharts
- Faizz
- Benford-Py

Data Pipeline

1) **Search Tool**

   PDF → Convert to Text → Cleansing → Sentence/Paragraph/Summary list → BERT Embedding list → FAISS Clustering → Model Serialization → Search API → Inference Task

2) **Text Network Analysis**

   PDF → Convert to Text → Cleansing → Summary → Most relevant topics and their relations → visualize by Networkx→ Text Network Analysis

3) **Benford Histogram**

   PDF → Convert to Text → Cleansing → Crawl variable & numerical value from balance sheet → Cluster into Asset, Liability, Debts, Income, etc → Collect previous data → Calculate histogram

Process

1. Research of audit documents
2. Scraping data
3. Cleaning
4. Text Analysis
5. Visualization

Team Roles

1. Data Modeling/Mining : **AL**wi husada
2. LEAD, FrontEnd Engineer : **BUD**i satrio
3. Data Engineering : re**DY** andriyansah

Tools

- **IndoBERT** untuk text embedding
- BERT extractive **Summarizer**
- Facebook AI Similarity Search (**FAISS**)
- **NeoVis** untuk Text Graph Analysis
- **Uvicorn** Web Server