EKSTRAKSI INFORMASI DARI DOKUMEN

PEMROSESAN BAHASA ALAMI

Disusun oleh:

Arif Athaya Harahap	235150200111056
Ariiq Tsany Zu	235150200111049
Fadhlullah Akmal	235150207111068
Muhammad Rafly Ash Shiddiqi	235150207111062
Rashky Rahmadian Jauhara	235150201111066

TEKNIK INFORMATIKA

DEPARTMEN TEKNIK INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS BRAWIJAYA

MALANG

2025

1. Pendahuluan

Proyek ini bertujuan untuk mengekstrak teks dari dokumen PDF dan DOCX, kemudian menyusunnya menjadi format terstruktur yang mudah dibaca dan dianalisis. Hasil ekstraksi disimpan dalam dua format: Markdown (.md) untuk pembacaan manusia dan JSON (.json) untuk pemrosesan programatik.

2. Alur Kerja Sistem

2.1 Ekstraksi Teks

- **PDF (pdf_extractor.py)**
  - Menggunakan library pdfplumber untuk membaca halaman demi halaman.
  - Teks setiap halaman dikumpulkan ke dalam daftar pages dan digabung menjadi string penuh full.
- **DOCX (docx_extractor.py)**
  - Menggunakan library python-docx.
  - Setiap paragraf diekstrak bersama informasi gaya (style), apakah paragraf merupakan heading, dan level heading jika ada.
  - Informasi ini digunakan untuk membangun hierarki dokumen.

2.2 Klasifikasi dan Struktur Dokumen

- **Lines processing (hierarchy.py)**
  - Dokumen dipecah menjadi baris demi baris.
  - Fungsi classify_line_heading menentukan apakah suatu baris merupakan heading, beserta levelnya, menggunakan:
    - Pola numbering (contoh: BAB I, 1.1)
    - Huruf kapital, jumlah kata
    - Roman numerals
  - Fungsi build_hierarchy_from_lines menyusun tree dokumen dengan format:

```
{
    'title': 'Judul',
```

```
    'level': n,
    'content': 'Isi paragraf',
    'children': [...]
}
```

- Jika input berasal dari DOCX, level heading yang sudah terdeteksi dari style akan diprioritaskan.

2.3 Konversi ke Markdown

- Fungsi tree_to_markdown mengubah tree dokumen menjadi teks Markdown.
  - Level heading diterjemahkan menjadi jumlah tanda #.
  - Isi paragraf ditempatkan di bawah heading yang sesuai.
  - Markdown yang dihasilkan mudah dibaca dan siap untuk publikasi atau dokumentasi.

2.4 Penyimpanan Output

- Fungsi save_outputs menyimpan:
  - Markdown (.md) → untuk pembaca manusia
  - JSON (.json) → untuk analisis lebih lanjut atau integrasi sistem lain
- File output diberi nama berdasarkan nama file input ditambah suffix _structured, contohnya:

```
input.pdf → input_structured.md, input_structured.json
```

2.5 Integrasi Sistem

- main.py mengelola alur utama:
  - Deteksi tipe file (PDF atau DOCX)
  - Pemanggilan fungsi ekstraksi sesuai tipe file
  - Pembuatan tree hierarki
  - Konversi ke Markdown
  - Penyimpanan output

- run_local dapat menerima daftar file dan memproses semuanya sekaligus.

3. Hasil

Percobaan kali ini menggunakan dokumen input.pdf dengan isi sebagai berikut.

3.1 Input

Input disini menggunakan cuplikan dari dokumen input.pdf pada bagian Pernyataan Orisinalitas.

PRAKATA

Puji syukur penulis panjatkan ke hadirat Allah SWT yang telah melimpahkan rahmat dan hidayah-Nya sehingga penulis dapat menyelesaikan skripsi yang berjudul "Implementasi *Improved Sqrt-Cosine Similarity* Untuk Pemeringkatan Resume Berdasarkan Kualifikasi Lowongan Kerja". Penulis menyadari bahwa dalam penyusunan skripsi tidak terwujud tanpa adanya dukungan, bimbingan, arahan, serta doa yang tiada hentinya dari berbagai pihak. Pada kesempatan kali ini penulis mengucapkan terima kasih sebesar-besarnya kepada:

1. Bapak Rizal Setya Perdana, S.Kom., M.Kom., Ph.D. selaku dosen pembimbing satu yang telah menyetujui dan mengarahkan penulis sehingga dapat menyelesaikan skripsi ini.
2. Ibu Ir. Indriati, S.T., M.Kom. selaku dosen pembimbing dua yang telah menyetujui dan membimbing dalam penulisan untuk pengerjaan skripsi ini.
3. Bapak Bayu Priyambadha, S.Kom., M.Kom., Ph.D. selaku Ketua Program Studi Teknik Informatika Fakultas Ilmu Komputer Universitas Brawijaya.
4. Bapak Sabriansyah Rizqika Akbar, S.T., M.Eng., Ph.D. selaku Ketua Departemen Teknik Informatika Fakultas Ilmu Komputer Universitas Brawijaya.
5. Rajiv Maulana selaku validator dalam skripsi ini, serta seluruh rekan kerja penulis yang telah berkontribusi dalam memperluas wawasan dan pengetahuan penulis selama proses penelitian.
6. Ayah Tripinto Laksono, S.Kom. dan Bunda Dian Laksono selaku kedua orang tua penulis, Sangkalsar Laksono selaku adik penulis, dan seluruh keluarga penulis yang senantiasa memberikan dukungan, doa, dan motivasi sehingga dapat menyelesaikan skripsi ini.
7. P4OP Dinas Pendidikan Jakarta selaku penyelenggara beasiswa KJMU yang membantu penulis dalam menyelesaikan studi sarjana.
8. Seluruh teman tercinta penulis hingga saat ini yang telah menjadi teman diskusi selama proses penelitian, teman seperjuangan, serta sumber motivasi, terutama Salsabila Rachmayani, Kirana Alivia, Nathania Putri, Aidah Az Zahra, Raditya Atmaja, Rolyan Zain, Ade Arya, Nadhira Nurannisa, Saqina Salsabila, Ghania Tanziela, Gustav Ali, Emilia Putri, Farel Rakha, Aldiansyah, Dzaki Rafif, Bagas Antarino, Safia Putri, Rayshanda Yuwandina, Arkan, Alka, Faqih, Audrey, Aelissa, Dina, Kurnia, dan Zahra.

Malang, 25 Juni 2025

Penulis

khansalaksono@gmail.com

iv

3.2 Hasil Output Markdown

Hasil output markdown disimpan dalam input_structured.md

PRAKATA

Puji syukur penulis panjatkan ke hadirat Allah SWT yang telah melimpahkan

rahmat dan hidayah-Nya sehingga penulis dapat menyelesaikan skripsi yang

berjudul "Implementasi Improved Sqrt-Cosine Similarity Untuk Pemeringkatan

Resume Berdasarkan Kualifikasi Lowongan Kerja". Penulis menyadari bahwa

dalam penyusunan skripsi tidak terwujud tanpa adanya dukungan, bimbingan,

arahan, serta doa yang tiada hentinya dari berbagai pihak. Pada kesempatan kali

ini penulis mengucapkan terima kasih sebesar-besarnya kepada:

1. Bapak Rizal Setya Perdana, S.Kom., M.Kom., Ph.D. selaku dosen pembimbing

satu yang telah menyetujui dan mengarahkan penulis sehingga dapat

menyelesaikan skripsi ini.

2. Ibu Ir. Indriati, S.T., M.Kom. selaku dosen pembimbing dua yang telah

menyetujui dan membimbing dalam penulisan untuk pengerjaan skripsi ini.

3. Bapak Bayu Priyambadha, S.Kom., M.Kom., Ph.D. selaku Ketua Program Studi

Teknik Informatika Fakultas Ilmu Komputer Universitas Brawijaya.

4. Bapak Sabriansyah Rizqika Akbar, S.T., M.Eng., Ph.D. selaku Ketua Departemen

Teknik Informatika Fakultas Ilmu Komputer Universitas Brawijaya.

5. Rajiv Maulana selaku validator dalam skripsi ini, serta seluruh rekan kerja

penulis yang telah berkontribusi dalam memperluas wawasan dan

pengetahuan penulis selama proses penelitian.

6. Ayah Tripinto Laksono, S.Kom. dan Bunda Dian Laksono selaku kedua orang tua

penulis, Sangkaisar Laksono selaku adik penulis, dan seluruh keluarga penulis

yang senantiasa memberikan dukungan, doa, dan motivasi sehingga dapat

menyelesaikan skripsi ini.

7. P4OP Dinas Pendidikan Jakarta selaku penyelenggara beasiswa KJMU yang

membantu penulis dalam menyelesaikan studi sarjana.

8. Seluruh teman tercinta penulis hingga saat ini yang telah menjadi teman diskusi

selama proses penelitian, teman seperjuangan, serta sumber motivasi,

terutama Salsabila Rachmayani, Kirana Alivia, Nathania Putri, Aidah Az Zahra,

Raditya Atmaja, Roiyan Zain, Ade Arya, Nadhira Nurannisa, Saqina Salsabila,

Ghania Tanziela, Gustav Ali, Emilia Putri, Farel Rakha, Aldiansyah, Dzaki Rafif,

Bagas Antarino, Safia Putri, Rayshanda Yuwandina, Arkan, Alka, Faqih, Audrey,

Aelissa, Dina, Kurnia, dan Zahra.

```
Malang, 25 Juni 2025
```

3.3 Hasil Output JSON

Hasil output JSON disimpan dalam input_structured.json


```
{
      "title": "PRAKATA",
      "level": 1,
      "content": "Puji syukur penulis panjatkan ke hadirat Allah SWT yang telah melimpahkan\nrahmat dan hidayah-Nya sehingga penulis dapat menyelesaikan skripsi yang\nberjudul “Implementasi Improved Sqrt-Cosine Similarity Untuk Pemeringkatan\nResume Berdasarkan Kualifikasi Lowongan Kerja”. Penulis menyadari bahwa\ndalam penyusunan skripsi tidak terwujud tanpa adanya dukungan, bimbingan,\narahan, serta doa yang tiada hentinya dari berbagai pihak. Pada kesempatan kali\nini penulis mengucapkan terima kasih sebesar-besarnya kepada:\n1. Bapak Rizal Setya Perdana, S.Kom., M.Kom., Ph.D. selaku dosen pembimbing\nsatu yang telah menyetujui dan mengarahkan penulis sehingga dapat\nmenyelesaikan skripsi ini.\n2. Ibu Ir. Indriati, S.T., M.Kom. selaku dosen pembimbing dua yang telah\nmenyetujui dan membimbing dalam penulisan untuk pengerjaan skripsi ini.\n3. Bapak Bayu Priyambadha, S.Kom., M.Kom., Ph.D. selaku Ketua Program Studi\nTeknik Informatika Fakultas Ilmu Komputer Universitas Brawijaya.\n4. Bapak Sabriansyah Rizqika Akbar, S.T., M.Eng., Ph.D. selaku Ketua Departemen\nTeknik Informatika Fakultas Ilmu Komputer Universitas Brawijaya.\n5. Rajiv Maulana selaku validator dalam skripsi ini, serta seluruh rekan kerja\npenulis yang telah berkontribusi dalam memperluas wawasan dan\npengetahuan penulis selama proses penelitian.\n6. Ayah Tripinto Laksono, S.Kom. dan Bunda Dian Laksono selaku kedua orang tua\npenulis, Sangkaisar Laksono selaku adik penulis, dan seluruh keluarga penulis\nyang senantiasa memberikan dukungan, doa, dan
```

```
motivasi sehingga dapat\nmenyelesaikan skripsi ini.\n7. P4OP Dinas Pendidikan Jakarta selaku penyelenggara beasiswa KJMU yang\nmembantu penulis dalam menyelesaikan studi sarjana.\n8. Seluruh teman tercinta penulis hingga saat ini yang telah menjadi teman diskusi\nselama proses penelitian, teman seperjuangan, serta sumber motivasi,\nterutama Salsabila Rachmayani, Kirana Alivia, Nathania Putri, Aidah Az Zahra,\nRaditya Atmaja, Roiyan Zain, Ade Arya, Nadhira Nurannisa, Saqina Salsabila,\nGhania Tanziela, Gustav Ali, Emilia Putri, Farel Rakha, Aldiansyah, Dzaki Rafif,\nBagas Antarino, Safia Putri, Rayshanda Yuwandina, Arkan, Alka, Faqih, Audrey,\nAelissa, Dina, Kurnia, dan Zahra.\nMalang, 25 Juni 2025",
      "children": [
        {
          "title": "Penulis",
          "level": 2,
          "content": "khansalaksono@gmail.com\niv",
          "children": []
        }
      ]
    }
```


4. Kesimpulan

Berdasarkan hasil pengembangan dan pengujian sistem, dapat disimpulkan bahwa proyek ini berhasil mengekstrak teks dari dokumen PDF maupun DOCX secara otomatis dan menyusunnya menjadi format terstruktur. Sistem mampu mengenali heading dan levelnya dengan memanfaatkan kombinasi pola teks, numbering, huruf kapital, dan style bawaan DOCX, sehingga struktur dokumen asli dapat dipertahankan. Output yang dihasilkan berupa Markdown memudahkan pembacaan manusia, sedangkan JSON memungkinkan pemrosesan lebih lanjut oleh aplikasi lain. Dengan demikian, sistem ini efektif untuk memproses dokumen

berstruktur sederhana maupun kompleks, serta menyediakan hasil yang fleksibel untuk keperluan dokumentasi dan analisis data.