Laporan Prediksi Subject Peneliti Berdasarkan Rekan

Farrel Alfarabi Saleh	1806186622
Helmi Alfarel	1806235896

1. Metodologi Pengumpulan Data

Data yang digunakan adalah data author pada bidang Artificial Intelligence, Machine Learning, Computer Vision, Natural language processing,yang ada di https://dl.acm.org/people/ai . Pada ACM, terdapat data profil seorang researcher yang berisi subject-subject dari paper-paper yang dia pernah kerjakan, beserta data daftar researcher yang pernah jadi co-authors nya.

Menggunakan package requests dan BeautifulSoup4, saya bisa mendapatkan html yang berisi daftar author beserta link menuju halaman profilnya terurut dari jumlah publikasi paling banyak, menggunakan url: https://dl.acm.org/people/ai?pageSize=1000&sortBy=count . setelah ini, saya bisa mengakses setiap halaman profilnya dan melakukan scraping data subjects, dan data co-authors nya untuk setiap researcher.

2. Rancangan Graph Mining Task

Task yang dikerjakan berjenis Node Classification. Sebuah vertex pada graph merepresentasikan seorang researcher. Sebuah edge antar 2 vertex merepresentasikan kedua researcher itu co-author. Graph adalah Weighted Undirected Graph, dengan weight sebuah edge adalah jumlah paper yang ditulis oleh 2 researcher. Graph berjenis directed.

Setiap vertex memiliki vektor subjects, yang merepresentasikan subject-subject yang dimiliki oleh seorang researcher. Sebagai contoh, misalnya pada graph ada 3 opsi subject, Image Processing, Text Processing, dan Big Data. jika researcher 1 memiliki subject Image Processing dan Text Processing, vektor subjects researcher 1 adalah [1, 1, 0]. Nilai dalam vektor subject tidak 1 seperti di contoh, melainkan jumlah paper yang berisi subject tersebut, misalnya [5, 1, 0] berarti researcher memiliki 5 paper yang bersubject Image Processing, 1 paper yang bersubject Text Processing, dan 0 paper yang bersubject Big Data.

Menggunakan edge dari sebuah vertex, kita dapat memprediksi vektor subject vertex/researcher itu, dengan memperhatikan vektor subject dari researcher co-author nya/ vertex tetangga nya.

Karena edge memiliki weight, ide awal kami dalam memprediksi vektor subject sebuah vertex adalah dengan menjumlahkan vektor subject dikali weight, untuk setiap vertex tetangganya.

3. Pengumpulan, Pemformatan, dan Pengolahan Data

Data yang berhasil dikumpulkan adalah sebanyak 1653 data subject researcher. Dari 1658 data tersebut, terdapat 213 data researcher yang dimiliki data daftar kolega nya, sehingga 213 ini yang akan dijadikan data untuk prediksi.

Data researcher dan kolega nya disimpan dalam bentuk adjacency matrix, dengan ukuran 213x1658. Setiap cell pada matrix berisi jumlah paper yang di co-author kedua researcher.
Vektor subject dari researcher-researcher disimpan dalam tabel (adjacency matrix) berukuran 1653x1245, karena himpunan subject dari 1658 researcher itu ada 1245 subject.

Untuk setiap nilai pada tabel researcher dan koleganya, nilainya dibagi jumlah dengan jumlah paper. Ide nya seperti itu. Namun terdapat kesalahan pada pengumpulan data, sehingga data jumlah paper yang pernah ditulis oleh researcher tidak disimpan. Jadi nilai pada tabel dibagi jumlah total setiap cell pada row. Sama halnya dengan tabel researcher dan subjectnya.

4. Eksperimen

Terdapat 2 cara yang saya lakukan pada prediksi subject researcher. Yang pertama adalah dengan menggunakan vektor subject yang dideskripsikan diatas, masing2 element di vektor adalah proporsi subject itu dibagi jumlah total vektor (vector.sum()). Yang kedua adalah menciptakan cluster subject sehingga sebuah vector subject bisa diberikan label dan pengerjaan lebih mirip classification.

Kemudian, ada variasi saat melakukan normalisasi. Cara normalisasi yang pertama adalah dengan membagi nilai pada cell dengan sum of row nya. Hal ini dilakukan pada data researcher dan koleganya dan data researcher dan subject nya. Cara normalisasi kedua adalah dengan melakukan Min Max Scaling pada setiap row di tabel researcher kolega dan tabel researcher subject.

5. Evaluasi

Pada cara prediksi subject researcher menggunakan vektor subject, metode evaluasi yang digunakan tidak bisa menggunakan metode evaluasi klasifikasi. Sehingga saya mengukur error vektor prediksi dengan menghitung jarak vektor prediksi dengan jarak vektor subject asli. Jaraknya menggunakan Euclidean Distance. Pada cara mengubah vektor subject menjadi cluster dengan KMeans, evaluasi yang digunakan adalah akurasi.

Pada data yang dinormalisasi dengan sum, Rata-rata Euclidean Distance dari prediksi vektor subject dan vektor subject asli adalah 0.268. Karena sum dari vektor subject maksimalnya adalah 1 (karena dibagi sum), nilai error ini sebenarnya cukup buruk, tapi tidak terlalu buruk. Sedangkan akurasi nya adalah 0.65.

Pada data yang dinormalisasi dengan max, Rata-rata Euclidean Distance dari prediksi vektor subject dan vektor subject asli adalah 1.47288. Sedangkan Akurasi nya adalah 0.33.

Kontribusi:

Helmi Alfarel	Mendefinisikan task, membuat Script pengumpulan data, melakukan scraping, meng-compile data yang dikumpulkan, mengolah data, membuat fungsi untuk memprediksi subject researcher, membuat laporan
Farrel Alfarabi Saleh	Beban