হাতেকলমে বাংলা ন্যাচারাল ল্যাঙ্গুয়েজ প্রসেসিং

'রিকারেন্ট নিউরাল নেটওয়ার্ক' থেকে 'ট্রান্সফরমার'

রকিবুল হাসান

আদর্শ

প্রকাশক: **আদর্শ**
৩৮ পি. কে. রায় রোড, বাংলাবাজার (২য় তলা), ঢাকা ১১০০
(+০২-৯৬১২৮৭৭, ০১৭৯৩২৯৬২০২, ০১৭১০৭৭৯০৫০
info@adarsha.com.bd
www.adarsha.com.bd
www.facebook.com/AdarshaFb

হাতেকলমে বাংলা ন্যাচারাল ল্যাঙ্গুয়েজ প্রসেসিং
১ম প্রকাশ: ৮ মাঘ ১৪২৭; ২২ জানুয়ারি ২০২০

মুদ্রণ ব্যবস্থাপনা: আদর্শ প্রিন্টার্স
রকমারিতে আদর্শর বই: www.rokomari.com/adarsha

Hatekalame Bangla Natural Language Processing (Published in Bengali)
by *Rakibul Hassan*
Published by Adarsha
38 P. K. Ray Road, Banglabazar (1st floor), Dhaka 1100
ISBN: 978-984-95324-0-8

উৎসর্গ

স্বাতী, যাকে দেখে আমার মনে বিশ্বাস জন্মেছে, যার নিজ ভাষায় ভিত ভালো, সে সব জায়গায় ভালো।

যিনি একাধারে সফল ব্যাংকার, মমতাময়ী মা এবং হাসেন সারা দিন। তবে, এমনই বইপোকা, বই খুলে না বসলে খেতে পারেন না একটুকুও।

রকিবুল হাসানের অন্য বইসমূহ

হাতেকলমে মেশিন লার্নিং (২য় সংস্করণ)
শূন্য থেকে পাইথন মেশিন লার্নিং (২য় সংস্করণ)
হাতেকলমে পাইথন ডিপ লার্নিং (২য় সংস্করণ)

সূচি

হাতেকলমে ন্যাচারাল ল্যাঙ্গুয়েজ প্রসেসিং

মানুষের এক্সটেনশন

Our intelligence is what makes us human, and AI is an extension of that quality.
–Yann LeCun, Professor, New York University

অলসতা এবং উদ্ভাবনা

ঝামেলার শুরু

আমি একজন অলস মানুষ। সেটার 'সার্টিফিকেট' পাবেন স্বাতীর কাছ থেকে। দিনের বড় একটা সময় 'নষ্ট' করি, কীভাবে পৃথিবীর অন্য মানুষদেরও অলস বানানো যায়। সত্যি বলছি! যেমন, (উদাহরণ হিসেবে বলছি) আমাদের কেন একটা কাজে এই অফিস ওই অফিস ছোটাছুটি করতে হবে? কেন বাসায় বসে সরকারি/বেসরকারি সার্ভিস পেতে পারি না? বাসায় বসে অ্যাপে জিনিসপত্র অর্ডার, টাকা পাঠানো, স্কুলের বেতন দেওয়া যায়, তাহলে অন্য কাজগুলো হচ্ছে না কেন?

আমার ধারণা, মানুষের কাজ আরও অনেক বড়। কোথায় আমরা ছুটব গ্রহ থেকে গ্রহান্তরে, সাগর থেকে সাগরের অতলান্তে, প্রকৃতির সঙ্গে যুদ্ধ করে, বিশেষ করে, নিজেদের খুঁজতে– সেখানে আমরা এখনো পড়ে আছি সেই একঘেয়ে একই কাজ করতে– প্রতিদিন। আমাকে এমন একটা 'রিপিটেটিভ' কাজের নাম বলুন, যেটাকে পাঠানো যায়

না যন্ত্রের কাছে? আজকে না গেলেও একসময় তো যাবেই। সেই যন্ত্রের কাছে কাজ পাঠানোর কাজই করি সময় পেলে।

আমি বসে আছি কবে আমরা হব 'ফ্ল্যাশ গর্ডন', যাব পৃথিবীর মতো আরও কিছু বসবাসযোগ্য গ্রহের খোঁজে। পাশাপাশি, মহামারি এবং প্রাকৃতিক বিপর্যয়গুলো থেকে বাঁচার উদ্ভাবনা আনতে হবে আমাদেরই। তবে, আমাদের জীবদ্দশায় 'সর্বজনীন ন্যূনতম আয়' অর্থাৎ 'ইউনিভার্সাল বেসিক পে' চলে এলে মুক্তি পাব এই একঘেয়ে কাজ থেকে। কাজ করবে যন্ত্র, মানুষ করবে উদ্ভাবনা, ক্রিয়েটিভ কাজ।

তবে, আমাদের প্রতিদিনের যে কাজগুলো 'রিপিটেটিভ' অর্থাৎ যে কাজগুলো বারবার করতে হয়, সেগুলোকে যন্ত্রের কাছে পাঠানোর জন্য আমার 'অলসতা' নেই। এই যে বইটা লিখছি, আগে টাইপ করে লিখে পাঠাতাম প্রকাশকের কাছে, এখন টাইপের মতো জিনিসগুলোকেও অসহ্য মনে হয়। মানুষ ক্রিয়েটিভ, সে কম্পিউটারের সামনে বসে কেন আঙুল দিয়ে টাইপ করবে? সে মুখে বলবে আর কাজ হয়ে যাবে। যেমন, আমি মুখে বলছি বলে বইটা লেখা হয়ে যাচ্ছে। অথবা, সামনে চিন্তা করতে করতেই লেখা হয়ে যাবে বই, হয়ে যাবে কাজ। বেশি দেরি নেই সেই সময়ের। নিউরোসায়েন্স এবং নিউরাল নেটওয়ার্ক চলে আসছে পাশাপাশি। এদিকে ই-কমার্স প্ল্যাটফর্মগুলো যখন জেনে যাচ্ছে আপনি কাল কী অর্ডার করবেন, তাহলে আর বাকি থাকল কী?

যন্ত্র এবং কাজ

তবে, ফিরে আসি এই বই লেখায়। কথা দিয়ে লেখা তো আর এমনি এমনি হচ্ছে না? যন্ত্র আমার মুখের আওয়াজ থেকে শব্দগুলোকে ধাপে ধাপে বুঝে নিয়ে লিখে ফেলার চেষ্টা করছে। প্রচুর ভুল হচ্ছে, আর সেই ভুল থেকেই শিখছে যন্ত্র। শুরুতে ভুলগুলোকে ঠিক করে দিচ্ছি আমরা মানুষ। যেভাবে বাচ্চাদের আমরা শেখাই। যখন সে আমার কথা বুঝতে পারছে না, তখন সেই শব্দের ডিকশনারির কাছাকাছি কোনো শব্দকে ব্যবহার করতে পারে, সেটার 'প্রবাবিলিটি ডিস্ট্রিবিউশনে' ফেলে দেওয়ার চেষ্টা করছে।

এই ভুল হওয়ার এবং লার্নিং প্রসেসের কারণে টাইপ করে এই বইটা লিখতে যে সময় লাগত, তার থেকে চার-পাঁচ গুণ বেশি সময় লাগছে এ মুহূর্তে। কারণ, যন্ত্র এখনো শিখছে, মানে বাংলায় শেখার জন্য এখনো তার শৈশবকাল চলছে। তবে, এ মুহূর্তে এর পেছনে চার গুণ সময় দেওয়ার উদ্দেশ্য হচ্ছে ভবিষ্যতে সে কমিয়ে আনবে সময়। টাইপ থেকেও। এই যন্ত্রকে আরও ভালো করে 'বাংলা' শেখানোর জন্য সাহায্য চাইছি আপনাদের। আপনাদের হাত দিয়ে আসবে প্রচুর বাংলা ডেটাসেট, সামনে।

ভাষা একটা 'কমপ্লেক্স' কাজ

মেশিন লার্নিংয়ের যে দক্ষতাটা এখন শিল্পের পর্যায়ে গেছে— বিশেষ করে মানুষের সঙ্গে যন্ত্রের যোগসূত্র স্থাপনে, সেখানে 'ন্যাচারাল ল্যাঙ্গুয়েজ প্রসেসিং' অংশটা অন্যতম। যন্ত্রকে মানুষের ভাষার মতো এত কমপ্লেক্স জিনিসকে যে শেখানো যায়, সেটা একটা বড় অর্জন। এই ভাষার ব্যাপারে আমি এখনো অনেক কিছু বুঝতে হিমশিম খেয়ে যাই। যেমন, কেউ একটা কথা বলল, আমি শুনলাম— তবে সেই কথাটা যে খোঁচা ছিল, সেটা মাথায় খেলল রাতে ঘুমানোর আগে। বুঝতে পারি, ভাষার ব্যাপারে আমি হয়তোবা এখনো 'টিউবলাইট'। অথচ, এখন মানুষের বলার ভঙ্গি, চেহারার 'অঙ্গভঙ্গিমা' থেকেই যন্ত্র 'সার্কাজম', 'ক্ল্যাসিফাই' করতে পারে কোন ক্যাটাগরিতে পড়বে সেই বলা কথাগুলো।

ভাষা কী?

ব্যাপারটা বোঝানো কষ্টের। দেখুন, এই ভাষার ডেফিনেশন দিতে গিয়ে উইকিপিডিয়াও হিমশিম খেয়ে যাচ্ছে।

ভাষার ডেফিনেশন

ভাষা ধারণাটির কোনো সুনির্দিষ্ট, যৌক্তিক ও অবিতর্কিত সংজ্ঞা দেওয়া কঠিন, কেননা যেকোনো কিছুর সংজ্ঞা ভাষার মাধ্যমেই দিতে হয়। তাই ভাষার আত্মসংজ্ঞা প্রদান দুরূহ। তবে ভাষার একটি কার্যনির্বাহী সংজ্ঞা হিসেবে বলা যায় যে ভাষা মানুষের মস্তিষ্কজাত একটি মানসিক ক্ষমতা, যা অর্থবাহী বাকসংকেতে রূপায়িত (বাগযন্ত্রের মাধ্যমে ধ্বনিভিত্তিক রূপ বা রূপে) হয়ে মানুষের মনের ভাব

প্রকাশ করতে এবং একই সমাজের মানুষের মধ্যে যোগাযোগ স্থাপনে সহায়তা করে। ভাষা মানুষে-মানুষে যোগাযোগের প্রধানতম বাহন।
—উইকিপিডিয়া

আমরা অঙ্কের মানুষ, ব্যাপারটাকে একটু সোজা করে নিয়ে আসি।

একটা ভাষায় যোগাযোগ করতে হলে শুরুতে দরকার একটা নির্দিষ্টসংখ্যক ভোকাবুলারি। একটা নির্দিষ্ট সমাজের মধ্যে নিজেদের চিন্তাভাবনা প্রকাশ করতে এই 'ভোকাবুলারি'র মধ্যেই আমরা থাকি। 'ভোকাবুলারি'র বাইরে কিছু বললে আপনি হয়তোবা সেটা বুঝবেন না। আমরা বাংলা ভাষায় কথা বললে সাধারণত বাংলা শব্দকোষের বাইরের শব্দগুলো নিয়ে নিজেদের ভাব প্রকাশ করব না। এর অর্থ হচ্ছে, প্রতিটা ভাষায় আমাদের এই 'ভোকাবুলারি' প্রায় নির্দিষ্ট। মানুষ শৈশবকালে ধীরে ধীরে এই ভোকাবুলারি অংশবিশেষ শিখেই বড় হতে থাকে। তবে, বড় হয়েও যখন আমরা এক শব্দের অর্থ বুঝি না, তখন রেফারেন্স হিসেবে অভিধান, অথবা শব্দকোষ ব্যবহার করি, যাতে ভবিষ্যতে সেই শব্দগুলোকে ব্যবহার করতে পারি।

কম্পিউটার কীভাবে ভাষা বোঝে?

কম্পিউটারের নিজের ভাষা হচ্ছে সংখ্যা। আর, সে কারণেই আমরা যেভাবে কথা বলি অথবা লিখি, সেগুলোকে কম্পিউটারে দেওয়ার আগে সংখ্যায় পাল্টে নিতে হয়। যন্ত্র যেহেতু অঙ্কের মডেলিংয়ে ভালো, সে কারণেই এই ভাষাকে যান্ত্রিক ভাষায় অর্থাৎ সংখ্যায় পরিবর্তন করে নেওয়ার ধারণাকে যন্ত্রকে শিখিয়ে দিলেই আমাদের কাজ কমে আসে। কম্পিউটারকে একটা ভাষা শেখানোর জন্য একটা ভাষার সবচেয়ে ছোট ইউনিট অক্ষর হলেও একটা বাক্যের মধ্যে শব্দগুলো কোনটা কোথায় বসেছে, সেটার ওপর আলাদা 'ইন্টারপ্রিটেশন' আসে।

ছোটবেলায় আমরা যখন কথা বোঝা বা বলা শুরু করি, তখন আমাদের অক্ষরজ্ঞান এবং ব্যাকরণ প্রয়োজন হয় না। সেই একইভাবে যন্ত্রকে যখন আমরা ভাষা শেখাতে যাব, তখন অক্ষরজ্ঞান অথবা ব্যাকরণ থেকেও একটা বাক্যে শব্দগুলোর 'প্লেসমেন্ট' ভাষাকে বোঝাতে সাহায্য করে। আমি যে জিনিসটা বলতে চাইছি— মানুষ যেভাবে শেখে, একটা যন্ত্রকেও সেভাবে শেখাতে হবে। অক্ষর দিয়ে

যেভাবে কোনো ভাষার আদান-প্রদান করা যায় না, সেভাবে একটা বাক্যের ভেতরে শব্দগুলো কখন, কীভাবে অথবা কতবার ব্যবহার হয়েছে, সেটার ওপরে মনের ভাব প্রকাশ করার একটা ধারণা তৈরি করা যায়। এ ব্যাপারে ভেতরের পুরো ব্যাপারটা হাতেকলমে করতে করতেই শিখে যাব আমরা।

মানুষ পড়তে পারার আগেই কথা বলতে পারে। আমরা ছোটবেলায় কথা বলা শিখেছি, পড়তে পারার আগে। আমরা শব্দকে মুখে উচ্চারণ করেছি এবং আরেকজনের মুখের উচ্চারণ শুনে উত্তর দিয়েছি। যন্ত্রের জন্য সেভাবে ব্যাপারটি শুরু হয়নি। যন্ত্র ভাষা বোঝা শুরু করেছে সরাসরি টেক্সট পড়ে। এর অর্থ হচ্ছে যন্ত্রের লেখা পড়ে ভাষা শিখতে হয়েছে। এর মানে এই নয় যে যন্ত্র টেক্সট পড়ে, ব্যাকরণ শিখে ভাষা শেখা শুরু করেছে। বরং, হাজারো টেক্সট ডেটা থেকে একটা ভাষা কীভাবে লেখা হয় সেটার একটা ধারণা পেয়েছে যন্ত্র। যত বেশি ডেটা, তত বেশি ভালো শিক্ষা। এ কাজটা সে শুরু করে 'টোকেনাইজেশন' করে। এ ব্যাপারগুলো নিয়ে আলাপ করব সামনে।

একটা কথা বলে রাখি চুপিসারে। যন্ত্র যদি 'ভাষা'র মতো কমপ্লেক্স জিনিস নিয়ে কাজ করতে পারে, তাহলে এই যন্ত্রকে দিয়ে সব কাজ করানো সম্ভব। পৃথিবীর সবচেয়ে কমপ্লেক্স কাজ হচ্ছে আইনি ডকুমেন্টের মধ্যে 'কন্ট্রাক্ট অ্যানালাইসিস', যার ওপর চলছে বড় বড় কোম্পানি, আইনি এবং আর্থিক সংস্থাগুলো। সেখানে যখন 'ন্যাচারাল ল্যাঙ্গুয়েজ প্রসেসিং' সহায়তা দিচ্ছে সংস্থাগুলোকে, তাহলে মানুষকে খুঁজতে হবে নতুন কাজ। তবে, এই অটোমেশন মানুষকে খুলে দিচ্ছে নতুন কাজের সুযোগ, যেগুলো নিয়ে চিন্তা করিনি আমরা। কাজ হারাবে না মানুষ, বরং 're-skilling' হবে খাতজুড়ে। এই যেমন আমি 'ন্যাচারাল ল্যাঙ্গুয়েজ প্রসেসিং' নিয়ে পাচ্ছি না দক্ষ মানুষ। বিশেষ করে বাংলায়। ট্রান্সক্রাইবিংয়ে।

কাদের জন্য এই বইটা প্রযোজ্য?

'ন্যাচারাল ল্যাঙ্গুয়েজ প্রসেসিং' (এনএলপি) নিয়ে এ বইটি লেখা হয়েছে আমার আরেকটি বই 'হাতেকলমে পাইথন ডিপ লার্নিং'য়ের সহযোগী বই হিসেবে। নিউরাল নেটওয়ার্কের ধারণার পাশাপাশি বাংলায় 'ন্যাচারাল

ল্যাঙ্গুয়েজ প্রসেসিং' নিয়ে বেশ কয়েকটা অধ্যায় লিখেছি সেখানে। 'এন্ড টু এন্ড' অর্থাৎ 'সেন্টিমেন্ট অ্যানালাইসিস' অ্যাপ্লিকেশন তৈরি করতে 'ন্যাচারাল ল্যাঙ্গুয়েজ প্রসেসিং' ব্যবহার করা হয়েছে ওই বইটাতে।

তবে, 'হাতেকলমে পাইথন ডিপ লার্নিং' বইটার পৃষ্ঠা সংখ্যা বেড়ে যাওয়ায় বেসিক অর্থাৎ শুরুর দিকে 'ন্যাচারাল ল্যাঙ্গুয়েজ প্রসেসিং' নিয়ে আলাপগুলো কমিয়ে এনেছিলাম তখন। আর এ কারণেই যারা নিউরাল নেটওয়ার্ক সম্বন্ধে জানেন অথবা 'হাতেকলমে পাইথন ডিপ লার্নিং' বইটা শুরু করেছেন, তবে ন্যাচারাল ল্যাঙ্গুয়েজ প্রসেসিং নিয়ে কিছু দ্বিধাদ্বন্দ্বে আছেন, তাদের জন্য এই বইটি প্রযোজ্য। আগের বইগুলোর মতো এখানে তাত্ত্বিক ব্যাপারগুলো থেকে হাতেকলমের ওপরে জোর দেওয়া হয়েছে বেশি। হাতেকলমে এবং নিজের চোখের সামনে যখন একটা জিনিস কাজ করে, তখন সেই জিনিসটি একটা ভালো ফাউন্ডেশন দেয় ভবিষ্যৎ ধারণায়।

কেন দর্শন প্রয়োজন?

আমার বইগুলো গল্পের ছলে লেখা হয় বলে যারা গল্প পছন্দ করেন না, তাদের অনুরোধ করব, অনলাইনে পুরোটুকু পড়ে কেনার সিদ্ধান্ত নিতে। আমি বিশ্বাস করি, সবকিছুর পেছনে দর্শন জানলে সেটা শেখার জন্য আগ্রহ জন্মায়। আমি 'প্রযুক্তি' নিয়ে লিখি ঠিকই, তবে পৃথিবীর সব শেখার পেছনে আছে দর্শন। এটার সত্যতা পাওয়া গেছে উইকিপিডিয়ার আর্টিকেলগুলোতে।

আমরা জানি, পৃথিবীর প্রায় সবকিছু নিয়েই কিছু না কিছু লেখা হয়েছে উইকিপিডিয়ায়। আমরা যখন কোনো (যেকোনো) উইকিপিডিয়া আর্টিকেলের প্রথম লিংক ধরে ক্লিক করে একটার পর একটা আর্টিকেলে (সার্চ ইঞ্জিনের মতো ক্রল করে) গেলে, সেটা শেষে গিয়ে পৌঁছাবে দর্শন আর্টিকেলে। এটা ট্রেন্ডটা বাড়ছে আস্তে আস্তে। এটার একটা অর্থ আছে। দেখুন, উইকিপিডিয়া এন্ট্রি কী বলছে?

> Clicking on the first link in the main text of a Wikipedia article, and then repeating the process for subsequent articles, usually leads to the Philosophy article. In February 2016, this was true for 97% of all articles in Wikipedia, an increase from 94.52% in 2011.
>
> –Wikipedia: Getting to Philosophy

কৃত্রিম বুদ্ধিমত্তার বিভিন্ন অ্যাপ্লিকেশনের পেছনে রয়েছে অনেকগুলো বিশাল দর্শন। সেটা না জেনে এটার ভেতরে ঢোকা দুষ্কর।

‘ন্যাচারাল ল্যাঙ্গুয়েজ প্রসেসিং’ জিনিসটা কী?

যন্ত্রকে ক্ষমতা দেওয়া

Natural Language Processing or NLP is a field of Artificial Intelligence that gives the machines the ability to read, understand and derive meaning from human languages.

কৃত্রিম বুদ্ধিমত্তার যে অংশটুকু যন্ত্রকে মানুষের মতো করে মানুষের ভাষা ব্যবহার করার ক্ষমতা দেয়, সেটাকে ‘ন্যাচারাল ল্যাঙ্গুয়েজ প্রসেসিং টুল’ বলতে পারি। এই টেকনিক ব্যবহার করে ভাষার ভেতরের ‘লিঙ্গুইস্টিকস’ অর্থাৎ ‘ভাষাতত্ত্ব’ জ্ঞানসহ অথবা ছাড়াই একটা অ্যাপ্লিকেশন বিভিন্ন স্ট্যাটিসটিক্যাল মডেল ব্যবহার করে বাস্তবসম্মত সমস্যাগুলোকে সমাধান করতে পারে। এই মানুষের ভাষাকে বুঝতে, জানতে, শিখতে, লিখতে, পড়তে এবং হৃদয়াঙ্গম করার জন্য যত ধরনের টুল ব্যবহার করা হয়, সেগুলোকে ন্যাচারাল ল্যাঙ্গুয়েজ প্রসেসিংয়ের সহায়ক টুল বলতে পারি।

এই বইটা হাতে নেওয়ার অর্থ হচ্ছে, আপনি তিন রাস্তার মোড়ে দাঁড়িয়ে আছেন। কম্পিউটার সায়েন্স, কৃত্রিম বুদ্ধিমত্তা এবং ভাষাতত্ত্বের এই তিন জগৎ আপনার সামনে উপস্থিত। তবে, এই নিউরাল নেটওয়ার্কের যুগে ভাষাতত্ত্ব অতটা না বুঝলেও চলবে, কারণ নিউরাল নেটওয়ার্কগুলো বিশাল ডেটা থেকে প্যাটার্ন বুঝতে ওস্তাদ। সেই ফিচার খুঁজে বের করে ফেলবে।

আগে ন্যাচারাল ল্যাঙ্গুয়েজ প্রসেসিং মানে আমরা ইংরেজি ভাষায় সবকিছু বুঝতাম, তবে এখন নিউরাল নেটওয়ার্ক আসার পরে এই বিষয়গুলো ভাষা স্পেসিফিক অর্থাৎ কোনো একটা ভাষার মধ্যে সীমাবদ্ধ না থেকে ডেটা সেটের ওপর ভিত্তি করে সেই মডেলগুলোকে তৈরি করা সহজ হয়ে গেছে।

এটা কম্পিউটার সায়েন্সের ফিল্ড হলেও এর মধ্যে মানুষের ভাষা সম্পর্কিত যতগুলো পদ্ধতি ব্যবহার করা হচ্ছে, সেগুলো যখন ভাষাকে দরকারমতো ‘অ্যানালাইজ’ এবং মডেলিং করে, তখন মানুষের

ভাষাকে বোঝার চেষ্টার পেছনে থাকছে এই কাজ। বর্তমানের প্রতিটা বুদ্ধিমান অ্যাপ্লিকেশন, যা মানুষের ভাষা নিয়ে কাজ করছে, সেগুলোর পেছনে আছে এই 'এনএলপি'।

একটা আক্ষরিক শব্দকে বুঝতে সেই শব্দের ভেতরে যে ট্রান্সফরমেশন প্রয়োজন, সেটাকে আমরা 'কম্পিউটেশনাল রিপ্রেজেন্টেশন' বলতে পারি। আবার, এই শব্দের 'রিপ্রেজেন্টেশন'কে ঠিকমতো বুঝতে, বিশেষ করে ডেটা থেকে শেখার জন্য প্রয়োজন 'মেশিন লার্নিং'। যন্ত্রকে মানুষের ভাষা বুঝতে শুরু থেকে শেষ পর্যন্ত অনেক ট্রান্সফরমেশনের মধ্য দিয়ে যেতে হয়, সেগুলোর আলাপ করব রাস্তায় নামলে। সামনে টোকেনাইজেশন, ভেক্টর, এমবেডিং ইত্যাদি টার্ম নিয়ে হাতেকলমে দেখব।

কেন 'ন্যাচারাল ল্যাঙ্গুয়েজ প্রসেসিং' শিখবেন?

'ওকে গুগল', 'সিরি', 'অ্যালেক্সা', 'গুগল ট্রান্সলেট' ইত্যাদি ধরনের প্রোডাক্ট এখন ব্যবহার হচ্ছে আমাদের বাসায় ও অফিসে। আমাদের পরবর্তী প্রজন্ম ডেস্কটপ কম্পিউটারের সামনে বসলেও গুগল সার্চ করে 'ওয়েব' ব্রাউজারের মাইক্রোফোনে চাপ দিয়ে। এদিকে 'ওকে গুগল', 'সিরি', 'অ্যালেক্সা' দিয়ে চালাতে পারেন বাসার অফিসের লাইট, ফ্যান, টিভি, গ্যারেজ খোলার মতো হাজারো অ্যাপ্লিকেশন। গুগল সার্চ করার সময় সেই শব্দকে মিলিয়ে আর কোন কোন বাক্য সার্চ করা হয়েছে, সেই প্রেডিকশন সময় কমিয়ে নিয়ে আসছে আমাদের সার্চে। এই সবগুলো প্রোডাক্টের পেছনে কাজ করছে 'ন্যাচারাল ল্যাঙ্গুয়েজ প্রসেসিং'।

৪১ বিলিয়ন ডলারের বাজার

The global Natural Language Processing (NLP) market size is expected to surpass USD 41 billion by 2025, at a CAGR of ~23%. This is owing to the growing demand for analyzing the data generated from conversations, social media, and other sources to enhance the customer experience.

—Adroit Market Research

বাংলায় 'ন্যাচারাল ল্যাঙ্গুয়েজ প্রসেসিং' নিয়ে ইংরেজির মতো সে রকম রিসোর্স না থাকলেও সেটা তৈরি করতে হবে আমাদেরই। আর, সেটা নিয়ে প্রচুর অ্যাপ্লিকেশন অপেক্ষা করছে সামনেই। বাংলায় আইনি সাহায্য নিয়ে প্রচুর কাজ পড়ে আছে সামনে। একটা বড় বাজার অপেক্ষা করছে, সেটার জন্য তৈরি হতে হবে আমাদের। ইংরেজি, বাংলা যা-ই হোক, ন্যাচারাল ল্যাঙ্গুয়েজ প্রসেসিং ঢুকে যাচ্ছে সব জায়গায়। আমি চাইছি, বাংলায় প্রচুর ডেটাসেট আসুক আপনাদের হাত ধরে। তাহলেই বাংলা হবে আসল 'প্রযুক্তির ভাষা', তখন সেটার ব্যবহার হবে সব মাত্রায়।

'ন্যাচারাল ল্যাঙ্গুয়েজ প্রসেসিং' শেখার পেছনে কয়েকটা ট্রেন্ডের কথা বলা যেতে পারে। প্রথমত, বর্তমানে এই ইন্টারনেটের যুগে আমরা নিজেরাই জেনারেট করছি প্রচুর ডেটা প্রতিমুহূর্তে। সঙ্গে থাকা প্রতিটা ডিভাইস প্রতি সেকেন্ডে এত লগ জেনারেট করছে, এখন এগুলোই আমাদের জন্য ট্রেনিং ডেটা। ডেটার পাশাপাশি প্রচুর ডেটা স্টোর করার ডিভাইসগুলোর দাম কমছে লাফিয়ে লাফিয়ে। আমরা নিজেদের চোখেই দেখছি, অসাধারণ প্রসেসিং স্পিড এবং ডেটাকে প্রসেস করার জন্য 'স্পেশালাইজড' অর্থাৎ 'জিপিইউ' এবং অনেক প্রসেসর দিয়ে ভর্তি আমাদের কাজের পরিবেশ। এই মুহূর্তে ডেটার এত ক্ষমতা থেকে এই সুবিধাগুলো না নিতে পারলে ভবিষ্যতে আমরা পিছিয়ে পড়ব আরও।

'ম্যাথমেটিকস' অথবা 'লিনিয়ার অ্যালজেবরা' নিয়ে আমাদের কিছুটা ধারণা থাকলেও আমরা চেষ্টা করব 'ন্যাচারাল ল্যাঙ্গুয়েজ প্রসেসিং' ব্যাপারটা হাতেকলমে দেখার জন্য। আমার গত চারটা বই লেখা এবং পাঠকদের সঙ্গে কথা বলার অভিজ্ঞতা থেকে এটা বলতে পারি, আমরা যখন কোনো জিনিস চোখের সামনে হতে দেখি এবং এর ভেতরে ট্রান্সফরমেশনে কী কী ঘটছে, সেটা যদি স্টেপ বাই স্টেপ বুঝতে পারি, তাহলে পুরো ব্যাপারটাই সহজ হয়ে যায় আমাদের জন্য।

শুরুতেই বলে নিচ্ছি, এ ব্যাপারে আপনাদের 'আর্টিফিশিয়াল ইন্টেলিজেন্ট' এক্সপার্ট অথবা 'ন্যাচারাল ল্যাঙ্গুয়েজ প্রসেসিং'-এ বোদ্ধা হওয়ার প্রয়োজন নেই। আমি আপনাদের গাইড এই পুরো রাস্তায়। পুরো ব্যাপারটাকে ঠিকমতো 'কনটেক্সটচুয়ালাইজেশন' লেভেলে আনতে

সাহায্য নিয়েছি গুগলের। বিশেষ করে লরেন্স মরোনির কোর্সেরার কোর্স এবং আরও অনেকের (কৃতজ্ঞতা দেখুন) কাছ থেকে। এখন কোলাবোরেশনের যুগ। একা কাজ করার যুগ শেষ। সেটার একটা ভালো ধারণা পাবেন রাস্তায় নামলে।

আমার একটা বড় সময় কাটছে 'ওপেনএআই'-এর জিপিটি-৩ ল্যাঙ্গুয়েজ মডেলকে বুঝতে। ১৭৫০ কোটি প্যারামিটার নিয়ে এই মডেলটা পৃথিবীর সবচেয়ে ক্ষমতাশালী 'লিঙ্গুইস্টিক' মডেল হলেও এ মুহূর্তে জিনিসটা এপিআই দিয়ে শুধু ইংরেজিতে সীমাবদ্ধ। তবে এর ধারণা অনুপ্রেরণা দিচ্ছে আমাদের প্রতিনিয়ত। সেই ট্রান্সফরমার মডেল নিয়ে ধারণার জন্য 'ওপেনএআই'-এর কাজ দেখতে পারেন তাদের সাইটে।

চলুন, যন্ত্রের সঙ্গে কথা বলি বাংলায়!