Geoparsing a jeho Využití v Datové Analýze

Co je geoparsing?

Geoparsing je proces extrakce geografických entit (např. názvů měst, zastávek MHD, ulic) z nestrukturovaného textu a jejich následné přiřazení ke geografickým souřadnicím. Tento přístup umožňuje **automaticky identifikovat a geokódovat místa zmíněná v textových datech**, což je klíčové pro prostorovou analýzu.
V datové analýze se geoparsing využívá pro vizualizaci geografického rozložení dat, analýzu pohybu uživatelů, sledování sentimentu v konkrétních lokalitách a predikci trendů. Tento projekt demonstruje aplikaci geoparsingu na dataset tweetů souvisejících s veřejnou dopravou v Praze.

Co tímto projektem demonstruji?

Tento projekt slouží jako ukázka aplikace geoparsingu v reálných datech a zároveň demonstruje několik klíčových aspektů datové analýzy:
Práce s rozsáhlejšími daty → Zpracování a analýza tisíců tweetů o veřejné dopravě v Praze.
Vlastní postup řešení → Od ručního vytvoření slovníku zastávek po vývoj vlastního geoparsovacího algoritmu.
Kreativní přístup → Kombinace různých přístupů k rozpoznávání míst (Named Entity Matching, fuzzy matching).
Netroviální zpracování dat → Čištění, transformace, sentimentová analýza, vizualizace aktivity uživatelů.
Projekt propojuje textovou analytiku, geoparsing a vizualizaci, čímž vytváří unikátní analýzu tweetů o veřejné dopravě.

📂 Struktura souborů

Soubor	Popis
Praha_tweety.csv	Zdrojová data – tweets o veřejné dopravě v Praze
praha.csv	Slovník zastávek MHD (původně extrahovaný z GTFS souboru ve formátu JSON)
praha_processed.csv	Zpracovaná data – tweety s přiřazenými geografickými lokalitami
geoparsing_praha.py	Hlavní geoparsovací skript
sentiment_tyden.py	Analýza průběhu sentimentu tweetů v čase
aktivita_graf.py	Vizualizace vývoje aktivity uživatelů
README.pdf	Dokument popisující celý projekt