Progetto Python, Prof Alfio Ferrara

Gianpaolo Coppola, Ivano Contu

Analisi degli annunci di AirBnb nella città di Milano

L'idea principale alla base di questo progetto è di effettuare un'analisi degli annunci presenti sul sito di AirBnb per quanto riguarda Milano. I principali obbiettivi di questa analisi sono:

- Trovare le correlazioni presenti tra appartamenti che riscuotono successo sulla piattaforma e le diverse feature analizzate
- Analizzare i titoli degli annunci e le recensioni col fine di trovare le parole chiave che possono portare ad un vantaggio competitivo su Airbnb
- Analizzare la posizione geografica degli appartamenti per studiare il peso rappresentato dalla distanza da luoghi cardine come le fermate della metro o diverse zone importanti a Milano

Per quanto riguarda la strategia operativa, i punti principali definiti sono:

- La ricerca e l'identificazione di dataset affidabili di annunci, recensioni e posizioni geografiche
- Creazioni di classi di funzioni per lo svolgimento dei punti successivi
- Data Cleaning per feature non rilevanti, Data Preprocessing per dati in formato non adatto e Text Cleaning per le analisi testuali
- Divisione del dataset in appartamenti considerati di successo e di insuccesso, la condizione scelta per la divisione è stata di almeno 50 recensioni ricevute e di una disponibilità per il prossimo mese di meno di 10 giorni
- Vari tipi di analisi statistiche riguardanti lo split del dataset su feature considerate importanti
- Analisi di testo tramite funzioni di conteggio di parole
- Analisi grafica per la posizione geografica dei due dataset, fermate della metro e luoghi di interesse di Milano
- Analisi statistica riguardante la distanza fra i vari appartamenti dei due dataset e i luoghi di interesse

Le ipotesi pre-progetto che son state successivamente confermate con lo svolgimento riguardano in primo luogo il prezzo e la vicinanza alle fermate della metropolitana, infatti in media gli appartamenti di successo sulla piattaforma hanno un prezzo più basso e si trovano più vicini alle metro.

Un'ulteriore analisi è stata condotta sulla distanza degli appartamenti in confronto a 5 luoghi chiave di Milano: Duomo, Parco Sempione, stazione centrale, San Siro e Porta Romana, anche in questo caso gli appartamenti di successo si trovano mediamente più vicino, tuttavia la differenza non è significativa.

Si è riusciti inoltre a trovare una lista di parole ricorrenti negli annunci che sono statisticamente rilevanti in quanto presenti in quantità fortemente maggiore nel dataset degli appartamenti di successo che sul secondo dataset.

A livello di codice Python le principali librerie utilizzate sono Pandas per la creazione e manipolazione di dataframe, Numpy e Scipy per le analisi statistiche, Matplotlib per la creazione dei grafici e mappe e Re per le operazioni con le espressioni regolari. Ulteriori librerie e moduli utilizzati sono indicati all'interno del file principale e di utils.py, la scelta delle funzioni utilizzate e create è spiegata direttamente all'interno dei due file.