

Sprawozdanie bankITup

Projekt Gravity

Gauss Team

15 Czerwiec 2018

1 Wstęp

W poniższym sprawozdaniu opiszemy uzyskane rezultaty podczas przeprowadzania projektu Gravity. Podczas rozwoju naszego projektu skupiliśmy się na dwóch elementach. Pierwszy z nich, na który położyliśmy większy nacisk, to analiza lokali gastronomicznych znajdujących się we Wrocławiu, próba wyjaśnienia czynników wpływających na dany poziom średniej oceny lokalu i ilości ocen na platformie Google Places oraz podanie czynników, na które naszym zdaniem, powinien zwrócić przyszły przedsiębiorca zakładający nowy lokal gastronomiczny. Druga część projektu to modyfikacja interfejsu graficznego, który jest bardziej nastawiony pod użytkownika. Pobiera on lokalizację użytkownika poprzez kliknięcie na mapę (w dalszym zamyśle pobiera ją z urządzenia użytkownika automatycznie) oraz sugeruje on najlepsze lokale w najbliższej okolicy.

2 Wypracowane rozwiązanie

Podczas pracy nad projektem padło wiele pomysłów w jaki sposób dobrze wykorzystać przeprowadzane przez nas analizy. W trakcie okresu zbierania danych zwróciliśmy szczególną uwagę na trudność w pozyskiwaniu danych odnośnie lokali gastronomicznych. Wiele serwisów nie posiada API, bądź usługa ta jest bardzo uboga lub posiada zbyt małą liczbę opinii oraz ocen. Ostatecznie zdecydowaliśmy się na wykorzystanie Google Places, które owszem oferuje API, lecz ilość informacji przez nie dostarczana była zbyt mała, stąd musieliśmy zdobyć nasze dane metodą manualną poprzez kopiowanie danych wprost z przeglądarki do pliku csv. Ostatecznie w ten sposób udało nam się uzyskać 222 rekordy oczyszczonych danych. W celu przeanalizowania naszych danych stworzyliśmy dwie klasyfikacje dobrych lokali:

- Lokale, których średnia opinia jest większa równa 4.4 oraz ilość opinii wynosi ponad 150,
- Lokale, których średnia opinia jest większa równa 4.5 oraz ilość opinii wynosi ponad 50 lub w ogólności ilość opinii danego lokalu wynosi ponad 300.

Pierwsza klasyfikacja umożliwiła nam wychwycenie bardzo dobry oraz sprawdzonych lokali, natomiast druga z nich pozwoliła nam wychwycić lokale bardzo dobre, lecz niekoniecznie sprawdzone, oraz lokale bardzo popularne. Z przeprowadzonych analiz, które znajdują się w pliku *Analysis.ipynb* wysunęliśmy następujące wnioski:

- Dobre lokale cechują się dostępem do piwa oraz wina w lokalu.
- Dobre lokale są przyjazne dla dzieci, rodzin oraz grup.
- Dobre lokale mają oznaczenie 'cosy' oraz 'casual' na portalu Google Maps.
- Dobre lokale podają wyśmienite koktajle.
- Dobre lokale częściej mają w ofercie dania wegetariańskie.
- Dobre lokale znacznie częściej mają stronę na facebooku.
- Dobre lokale pracują średnio dłużej o ok. 1h w weekendy oraz ok. 0,5h w tygodniu.
- Dostępność usługi dostawy podobnie często pojawiają się w obu rodzajach lokali.
- Stwierdzono korelację na poziomie ok. 0.3 pomiędzy dostępnością lokali dla grup a ilością recenzji, a ponadto korelację na podobnym poziomie między przyjaznością dla dzieci oraz oznaczeniem 'cosy' a drugim warunkiem klasyfikacji. Ponadto są to najwyższe odnotowane korelacje w danych.

Ponadto testowo wykorzystaliśmy kilka modeli uczenia maszynowego: Drzewa decyzyjne, Lasy losowe, Gradient Boosting, aby sprawdzić które cechy modele przede wszystkim wykorzystują do podejmowania decyzji w trakcie klasyfikacji. Wśród wyników otrzymywaliśmy przede wszystkim wszystkie kolumny dotyczące ilości lokali w okolicy, godziny pracy, cechy 'cosy', strony na facebooku. Analiza w tym obszarze nie została wykonana bardzo obszernie, lecz potwierdziła kilka naszych wcześniejszych wniosków dot. ważności godzin pracy, cechy 'cosy' oraz strony na facebooku.

Analiza kończy się wizualizacją wszystkich lokali wraz z wyróżnieniem dobrych lokali wg kategorii drugiej. Na jej podstawie możemy także wyciągnąć wnioski, które miejsca są przesycane dobrymi lokalami, a które czekają na nowe, dobre inwestycje.

Warto zaznaczyć, że powyższe analizy są wykonane na niepełnej bazie danych lokali we Wrocławiu (brakuje przede wszystkim centrum miasta i mogą istnieć drobne luki w pozostałych obszarach miasta), lecz pokazują, że widać w tych danych potencjał do zdobycia wiedzy dot. branży restauracyjnej we Wrocławiu.

Druga część zrealizowanej projektu, to system rekomendacji, który dla użytkownika w zaznaczonym miejscu wskazuje 5 najlepszych lokali w jego pobliżu.

Algorytm jest prosty, lecz daje bardzo zrównoważone wyniki rekomendacji oraz daje duże możliwości dalszego rozbudowania. Reguła opisana jest wzorem:

$$\frac{\text{ilość recenzji} \cdot \text{średnia ocena}}{\text{odległość użytkownika od lokalu}}$$

3 Potencjał rozwojowy

Nasze doświadczenie w zbieraniu danych oraz popularność serwisów typu social media doprowadziła Nas do wniosku, że warto by było stworzyć serwis, który oferowałby możliwość wrzucania zdjęć z restauracji wraz z jej oceną oraz recenzją dla użytkowników oraz możliwość prowadzenia strony lokalu dla przedsiębiorców. Ponadto ważną częścią tej aplikacji byłaby możliwość wyszukiwania najlepszych lokali w okolicy. Z przeprowadzonych badań wśród najbliższych okazuje się, że jest duże zapotrzebowanie na tego typu usługę.

Dzięki uzyskanym w ten sposób danym moglibyśmy przeprowadzić dokładniejsze analizy pośród lokali gastronomicznych, nie tylko we Wrocławiu, a jak widać z naszej analizy, jest potencjał w tego typu badaniach. Dodatkowo posiadane recenzje umożliwiłyby nam przeprowadzenie analizy sentymentów, która mogłaby wprowadzić kolejny ważny czynnik do analiz.

Usprawnienia mogły by być także przeprowadzone w algorytmie rekomendacji. Dzięki wykorzystaniu serwisów zewnętrznych moglibyśmy zastąpić odległość użytkownika w linii prostej od lokalu informacją o czasie dotarcia do danego miejsca (szczególnie ważne przy badaniu polecanych lokali, gdy znajdujemy się w pobliżu obiektów nieprzekraczalnych, np. rzeka). Ponadto na podstawie analizy historii odwiedzin użytkownika oraz jego profilowania moglibyśmy nadać wagi odpowiednim typom lokali i trafniej polecać kolejne. Warto zauważyć, że tego typu aplikacja byłaby szczególnie cenna podczas wyjazdów do innych miast, gdyż dzięki znanym preferencjom dostalibyśmy bardzo dobrą rekomendację w mgnieniu oka, co aktualnie jest niemożliwe.

4 Wykorzystane technologie

W naszym rozwiązaniu korzystaliśmy przede wszystkim z Pythona (Django, pandas, plotly, numpy) oraz JavaScriptu (Leaflet). Dodatkowo posiłkowaliśmy się niezastąpionym HTMLem oraz CSSem.