UFO w Azure Databricks

(w Stanach Zjednoczonych i nie tylko)

Patryk Krukowski
Marcin Świątkowski

Akademia Górniczo-Hutnicza w Krakowie

Analiza Dużych Zbiorów Danych

Agenda

1. Zbiór danych
2. Metody analizy
3. Wnioski dotyczące danych
4. Wnioski dotyczące użytej technologii

Zbiór danych

Sto trzydzieści tysięcy rekordów zebranych przez National UFO Reporting Center (NUFORC) w latach 1906 - 2014.

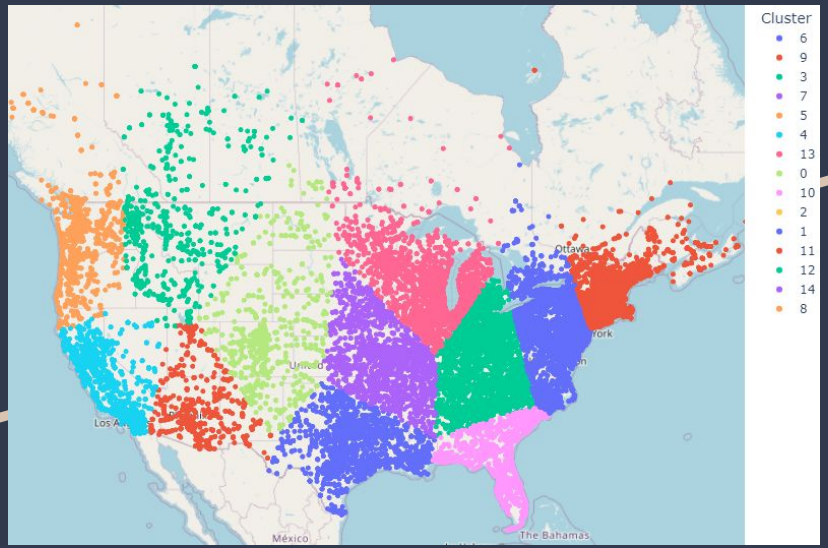
Atrybuty:

- czas zaobserwowania zjawiska z dokładnością do daty i godziny;
- miejsce (miasta i regiony);
- stan;
- kraj;
- długość i szerokość geograficzny;
- kształt UFO;
- czas trwania zjawiska w sekundach;
- czas trwania zjawiska w godzinach;
- komentarz osób, które zaobserwowały zjawisko;

Metody analizy

- Parsowanie danych
- Wyszukiwanie epicentrów wystąpień:
  - dla 5, 10 i 15 klastrów
- Częstość wystąpienia UFO w poszczególnych miejscach
- Relacja długości komentarza do czasu obserwacji
- Częstotliwość wystąpień UFO w ujęciu sezonowym
- Korelacja między wystąpieniem UFO a kształtem UFO

Wnioski dotyczące danych



Modele (Naive Bayes i Regresja logistyczna) nie dały najlepszych wyników, ponieważ nie zostały użyte modele neuronalne. Nawet jednak modele neuronalne mogłyby mieć problemy, ponieważ było ponad 27 klas.

Jeśli chodzi o eksplorację danych, to można wysnuć następujące wnioski o zaobserwowanych wystąpieniach UFO:

- skoncentrowane są na zachodzie USA;
- występują najczęściej w okresie letnim;
- długość opisu jest proporcjonalna do czasu obserwacji UFO

Wnioski dotyczące Azure Databricks

Zalety:

- Lepszy niż AWS
- Przejrzysty interfejs
- Możliwość dynamicznego zmieniania klastrów
- Możliwość usuwania klastrów
- Automatyczna regulacja pracy klastra
- Łatwy dostęp do rozbudowanych statystyk użycia klastra

Wady:

- Brak możliwości zmiany liczby workerów
- Problemy z podpięciem grafany (skomplikowana procedura)

Dziękujemy za uwagę

Patryk Krukowski

Marcin Świątkowski