Ekstrakcja cech oraz wyszukiwanie najbardziej podobnych ogłoszeń

25 października 2016

1 Tematy prezentacji na wykładzie - przydział

2016-12-02	1	Delta Szwadron Super Cool Komando Wilków Alfa	Torch
	2	"Duużyna"β	Theano
	3	Najgorsza	TensorFlow
2016-12-16	1	WK	Deeplearning4j
	2	Error h18	Graph Lab
	3		
2016-12-23	1	ARPA	Accord.net
	2	FUT	Open Ai Gym
	3		

Można zgłaszać modyfikacje swojego przydziału, jeśli nie wpływają one na prezentacje innych grup.

2 Omówienie poprzedniego zadania

3 Upoważnienia

4 Dyskusja na temat zadania konkursowego

- Jakie jest zadanie konkursowe?
- Jak może wyglądać praca nad rozwiązaniem konkursowym?
- Jakie będą elementy potrzebnego *data flow*?
- Jaka jest miara oceny oraz optymalna strategia gdy nic nie wiemy o danym ogłoszeniu?
- Pytania?

5 Ekstrakcja cech ⋆

Treść

Przetwórz dane training.zip http://dataninja.olx.pl/competition/data . Stwórz reprezentację *bag of words* na podstawie tytułów *title* ogłoszeń. Przyjrzyj się posortowanymi słownikowo cechami – wszystkimi słowami występującymi w zbiorze danych. Przyjrzyj się reprezentacji wybranych ogłoszeń w nowej formie i porównaj ją z oryginalnymi tytułami.

Sprawdź ile słów występuje w zbiorze danych – ile jest w nim cech? Sprawdź gęstość zbioru danych – czy wszystkie cechy występują dla każdego przypadku, czy jest on rzadki? Jak należy w związku z tym przechowywać dane?

6 Wyszukiwanie najbliższych sąsiadów ⋆

Treść

Dla pierwszych 10000 ogłoszeń i ich tytułów wyszukaj 4 najbardziej podobne ogłoszenia i ich tytuły, według utworzonej reprezentacji (*bag of words*). Zwróć uwagę na czas obliczeń.

Podaj wyniki dla ogłoszeń o identyfikatorach 23 i 78: wyświetl tytuł, niezerowe cechy i ich wartości, nazwę kategorii przypisanej oraz dystans od ogłoszenia.

W jaki sposób obliczane jest podobieństwo pomiedzy ogłoszeniami? Innymi słowy, jaka metryka została użyta? Jak użyta metryka wpływa na osiągane wyniki? Jaka metryka mogłaby być odpowiednia dla użytej reprezentacji i dlaczego?

Następnie dla ogłoszenia o identyfikatorze 111 wyszukaj 15 najbliższych sąsiadów. Czy jakie kategorie mają przypisane najbardziej podobne ogłoszenia? Czy kategorie te zgadzają się z kategorią ogłoszenia?

7 Najczęstsze kategorie dla każdej cechy ⋆

Treść

Dla każdej cechy wyszukaj kategorie przypisane do ogłoszeń posiadających niezerową wartość tej cechy oraz zlicz liczbę tych ogłoszeń. Dla każdej cechy sprawdź jaka kategoria występuje najczęściej oraz jej częstotliwość wśród ogłoszeń z daną cechą. Wyświetl wyniki dla 50 najczęściej występujących cech.

8 Raport 1 ⋆

Treść

Do wykładu 28 października 2016 na mojej stronie pojawi się opis wymagań dotyczących Raportu 1.