Zajęcia 12. Podejście TFIDF w algorytmach data science

W których poznanych algorytmach data science (analiza sentymentu - słowniki, klastrowanie k-means, topic modeling - metoda LDA, asocjacje - korelacja Pearsona) zastosowanie podejścia TF-IDF może poprawić jakość wyników? Z kolei w których algorytmach należy unikać TFIDF? - uzasadnij krótko swoją odpowiedź.

Przy klastrowaniu zastosowanie podejścia TF-IDF jest kluczowe, bo pozwala nam zdecydowanie lepiej patrzeć na to, które słowa są kluczowe dla zrozumienia danego tematu czy dokumentu (bo pojawia się głównie w nim), a które są w wielu dokumentach i są mniej ściśle powiązane z innym słowem, które jest obok. Bazowo topic modeling też nie zadziała, oryginalnie potrzeba nam rozkładu Dirichleta, ale przy danych podzielonych już można skorzystać.

Z kolei przy analizie sentymentu należy unikać TFIDF - to, że słowo pojawia się w całym dokumencie nie oznacza, że jest mało ważne. Wręcz przeciwnie, powtarzanie pozytywnego słowa przez wiele dokumentów sprawia, że przekaz wszystkich dokumentów staje się jeszcze bardziej pozytywny. TF-IDF w takim wypadku myliłoby na korzyść pojedynczych odmiennych opinii w paru dokumentach. Podobnie przy asocjacji interpretacja byłaby niemożliwa - bo patrzymy już na start na wszystkie dokumenty, a nam chodzi o surowe wektory.

Generalnie opłaca się korzystać z obu metod, jeżeli mamy możliwości obliczeniowe, ale jeżeli musimy wybierać, to właśnie w taki sposób bym to zorganizował.