**Topic Clustering**

Máme k dispozici sadu příspěvků vypublikovaných stránkou *NBC News World* a každý příspěvek je označen seznamem témat, která byla v jeho obsahu detekována. Chceme klientovi prezentovat témata, o kterých stránka píše. Jako forma vizualizace byl zvolen *word cloud* jednotlivých témat.

Protože je témat mnoho, rádi bychom klientovi usnadnili orientaci a prezentovali mu vizualizaci, která je interaktivní a navíc témata organizuje do skupin.

Klientovi se představí jeden word cloud (všech detekovaných témat), na tento word cloud může klient kliknout myší a on se rozpadne na dva menší. Vzniklé menší word cloudy představují tématické oblasti. Každá tématická oblast obsahuje množinu témat, která jsou si podobná. Každé téma se vyskytuje pouze v jedné tématické oblasti. Klient může opakovaně klikat na word cloudy a vytvářet tak menší, úžeji zaměřené tématické oblasti. Vizte ukázku [*topics\_clustering.png*](https://drive.google.com/open?id=1HdVBFbwblsuEKh2z6_9L-aQ3kA4V1Z3T).



Soubor se vstupními daty [*posts\_with\_topics.json*](https://drive.google.com/open?id=10ACW6-m5Xip8hJMOI_Ftf34iwImoL19c) obsahuje sadu příspěvků a jejich témat ve formátu:

[{post\_id: “123\_456”, topics:[“t1”, “t2”, “t3”]}, {post\_id: “789\_123”, topics:[“t1”, “t2”, “t3”]},..].

Vaším úkolem je načíst a připravit data, data zpracovat a výsledek ve vhodné struktuře uložit jako JSON. Jako přístup ke zpracování dat navrhujeme clustering, ale můžete použít jiný vhodnější přístup. Řešení naimplementujte ideálně v jazyku Python nebo Javascript, při řešení můžete použít libovolné knihovny. Prioritou je dodat kompletní řešení včas, zdokumentujte případné ústupky.

Kromě samotné implementace navíc stručně popište:

* jednotlivé kroky řešení, datové reprezentace, výběr metody
* proč jste zvolil(a) právě takové řešení, reprezentaci, metodu
* jak byste ověřil(a), že je vaše řešení správné a funkční
* jaké jsou silné a slabé stránky vašeho řešení, co by se na vaší práci dalo vylepšit a jak
* navrhněte modifikace/optimalizace vaší práce tak, aby
  + řešení efektivně fungovalo/škálovalo nad velkým data setem (řádově desítky GB)
  + řešení fungovalo v produkci velmi rychle