**Topic Clustering**

We have a set of messages published by a Facebook page *NBC News World* and each of the messages is labeled with a list of topics detected in the message. We would like to present these topics to our client. It has been decided that a *word cloud* of individual topics is the best way to visualize the data.

There is a lot of topics, therefore we would like to help the client to orient in the data and present a visualization that is interactive and organize topics into groups.

In the first step, single word cloud will be presented to the client. This word cloud has to contain all topics detected in the messages. Next, client can click on the word cloud and it will split into smaller word clouds. These smaller word clouds should be designed such that each of them will represent a thematic area - a set of topics that are related to each other. A topic can appear in only one thematic area. Client can repeatedly click on the word clouds and see smaller, more narrow thematic areas and this way explore streams of topics discussed on the Facebook page. See example [*topics\_clustering.png*](https://drive.google.com/open?id=1HdVBFbwblsuEKh2z6_9L-aQ3kA4V1Z3T).



A file with input data [*posts\_with\_topics.json*](https://drive.google.com/open?id=10ACW6-m5Xip8hJMOI_Ftf34iwImoL19c) contains a set of messages and their topics in following form:

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

Your task is to load, prepare and process the data and store the result in an appropriate structure as a JSON file. We suggest you to use clustering during the processing phase but it is completely up to you to choose the right approach. Your solution should be implemented in Python or Javascript, you can use libraries as needed. A priority is to deliver the solution on time and document as many details as possible.

We would also like you to describe:

* structure and individual steps of the solution, selection of data representations, methods, etc.
* why did you choose this solution, representation, method (why not something else)
* how would you validate that your solution is correct (algorithm) and functional (implementation)
* strengths and weaknesses of your solution, what should be improved and how
* possible modifications/optimizations (algorithmic and product point of view) that would:
  + work and scale efficiently on large data sets (tens of GBs)
  + low response time in production