Mini Project III – Topic Modelling

- Purpose
  - Apply **Unsupervised Learning** to identify **Meaningful Topics**!

1. Choose a **textual unstructured dataset** from any news or social media **API** (i.e. twitter, CNN, etc)
2. **Constrain your analysis to identifying meaningful topics in an unsupervised way (i.e. anomaly detection, clustering, dimensionality reduction)**
3. Follow the steps presented in the course so far,
   a) Extract data using **API**
   b) Perform **data preprocessing** on textual data
   c) Shape unstructured textual data into structured data using **Word Embedding**
   d) Perform **EDA** on the data
   e) Select **features** (i.e. words, phrases, sentences, paragraphs, etc)
   f) Learn **hidden patterns** in data (i.e. unsupervised learning)
   g) Evaluate **how well your models correctly classify** the documents into the right groups!
   h) **Output**: **Slides** (Value Proposition to Business Slides) & **Notebook** (Approach & Methodologies to your fellow Data Scientists).
4. Resources:
   a) Word Embedding Tutorial (Optional)
   b) Anomaly Detection with Auto-Encoders (Optional)
   c) Sample Mini Project III Notebook (i.e.to be used on as reference)
   d) Sample Mini Project III Visualization (i.e.to be used on as reference)
   e) Tableau Software Download (Student Version)
5. **Hint**: Use the resources to reverse engineer learning on concepts (i.e. word embedding, auto-Encoders) you if you need them for the project

© 2020 Institute of Data