**Work with merge\_df**

**Preprocessing:**

* Clean
  + remove quotes and special characters
  + Lower case
  + Merge tech with technology
  + Merge sport with sports
* Drop id and summary
* Keep file\_path (categories) and article
* One-hot encode categories
  + Politics = 0
  + Business = 1
  + Sports = 2
  + Other categories (e.g. accident, architecture) = 3
* Word embeddings for the article
  + Convert the articles into numbers
    - E.g. CountVectorizer, Word2Vec, Doc2Vec

**Split the data:**

* Training and Testing

**Train & Test**

* **Classifier (now)**
  + Neural Network
    - Pre-trained model; don’t train from scratch
    - Fine-tune the pre-trained model (hyper-parameter tuning)
* **Summarization (next)**
  + Bert or Bart\*
  + Pre-trained model;