**R&Py Training Project 2 - NLP**

**Instructions:**

* You will complete this project using Python.
* Create you own code and compare it with output

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

* **Problem Statement: Overall Objective**

Apply Natural Language Processing to Apple’s news articles.

* **Requirement:**

Connect to webhose, get your API key and do the following, to retrieve the dataset.

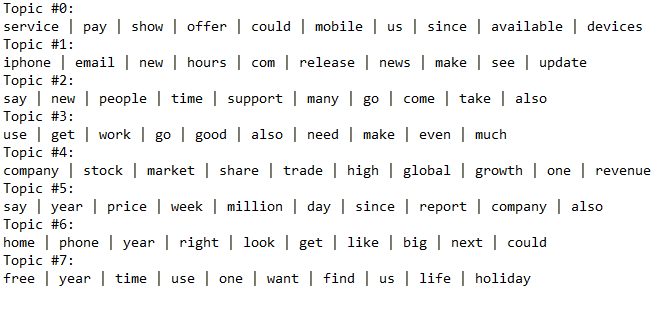


**How to achieve the objective: Step by Step Instructions: Procedure**

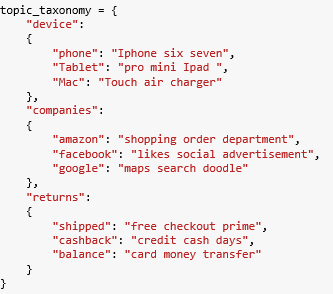
1. Feed the article titles from the dataset to Watson NLU to obtain the list of named entities. The output should be as follows

****

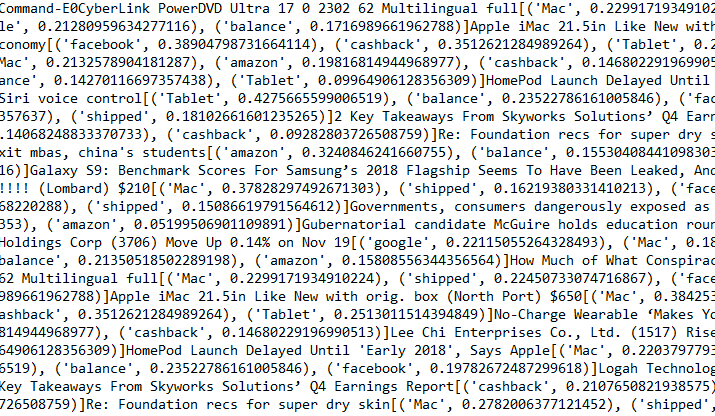
1. Get the raw text from the dataset and perform the following operations
2. Tokenization
3. Lemmatization
4. Vectorization
5. Perform LDA on the article titles. The output should be like what’s below

****

1. Create a topic taxonomy that looks like below, choose the parent topics (device, companies…etc.) and subtopics (shopping, order, department etc.) based on the daughter topics (amazon, facebook…etc.) obtained from the above LDA results. The topic taxonomy should look like below

****

1. Use the word2vec model in conjunction with the topic taxonomy to generate the classification output which looks like

****

* **Final Goal (Desired Output):**

Desired Output (Numbers may vary depending on the training and testing samples):

**Source Code:**

It is highly recommended to try and come up with your own code and solution of above project for best learning. However, if any questions, please use below code for your reference -

<https://github.com/datadream89/Topic-Classification-of-Amazon-Dataset>