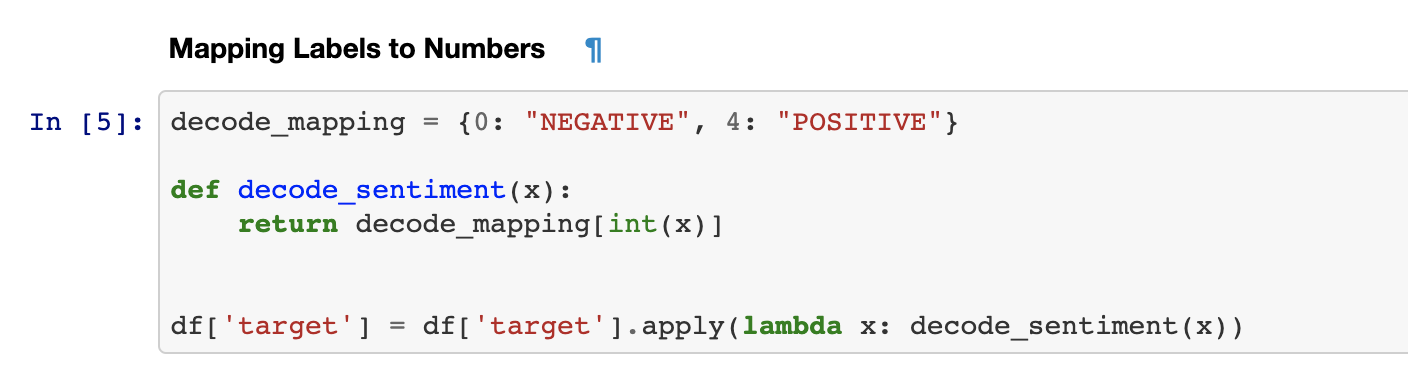
Capstone Project 3 Milestone Report 1

**Sentiment Analysis - Sentiment140 dataset with 1.6 million tweets** ([click here](https://www.kaggle.com/kazanova/sentiment140))

**About the Data:** This is the sentiment140 dataset. It contains 1,600,000 tweets extracted using the twitter api. The tweets have been annotated (0 = negative, 4 = positive) and they can be used to detect sentiment.

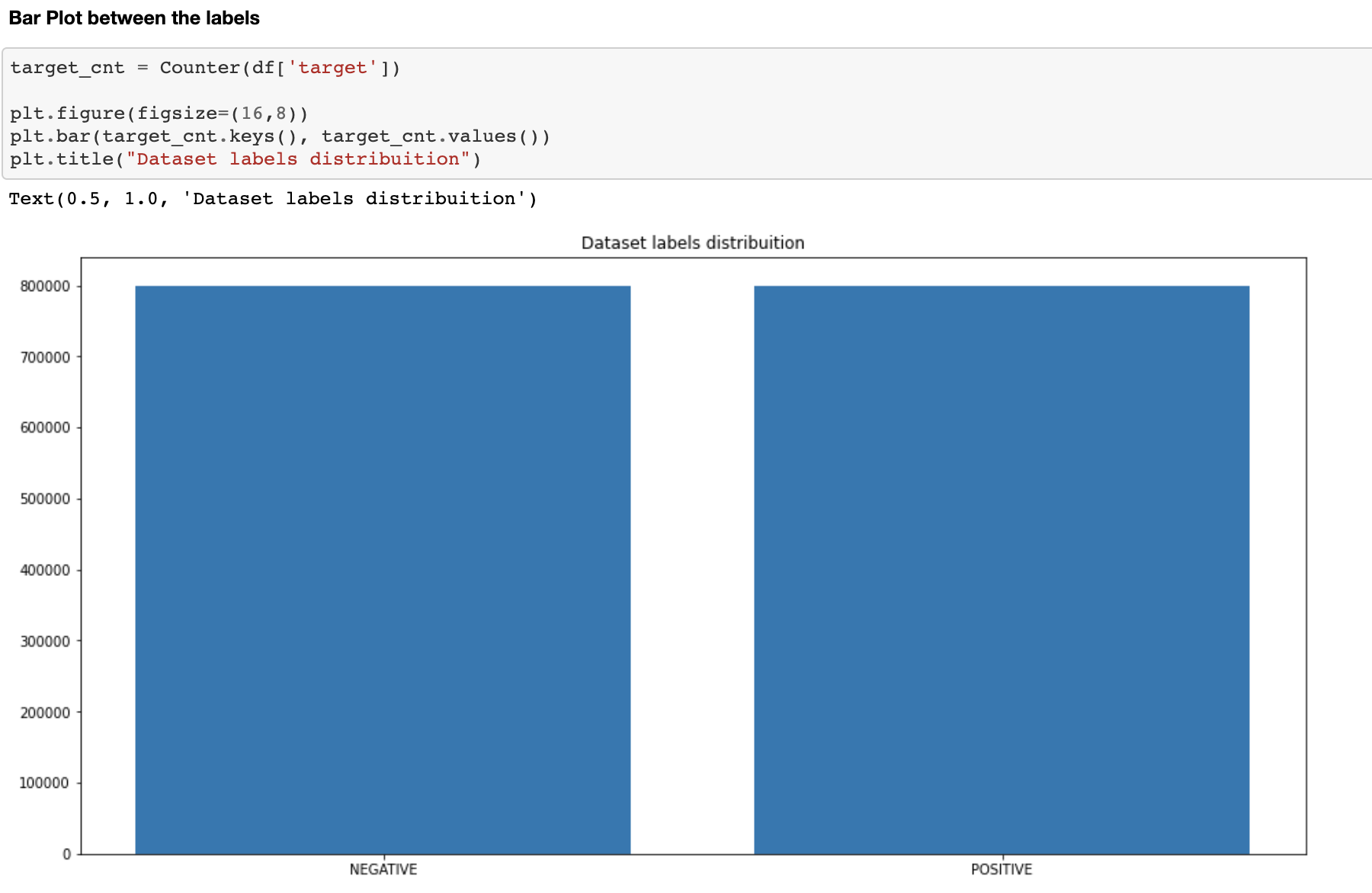
**Problem statement:** Classification of people’s opinion or expressions into different sentiments. Sentiments include Positive, Neutral, and Negative, Review Ratings and Happy, Sad. Sentiment Analysis can be done on different consumer centered industries to analyze people's opinion on a particular product or subject.

Initially we will map binary numerical classifications to labels

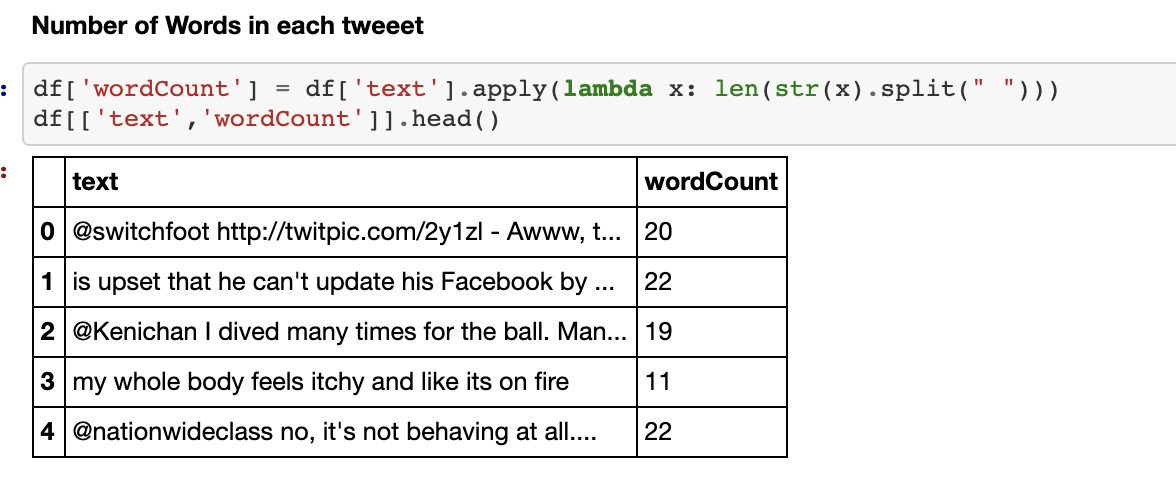


**Analyzing the dataset:**

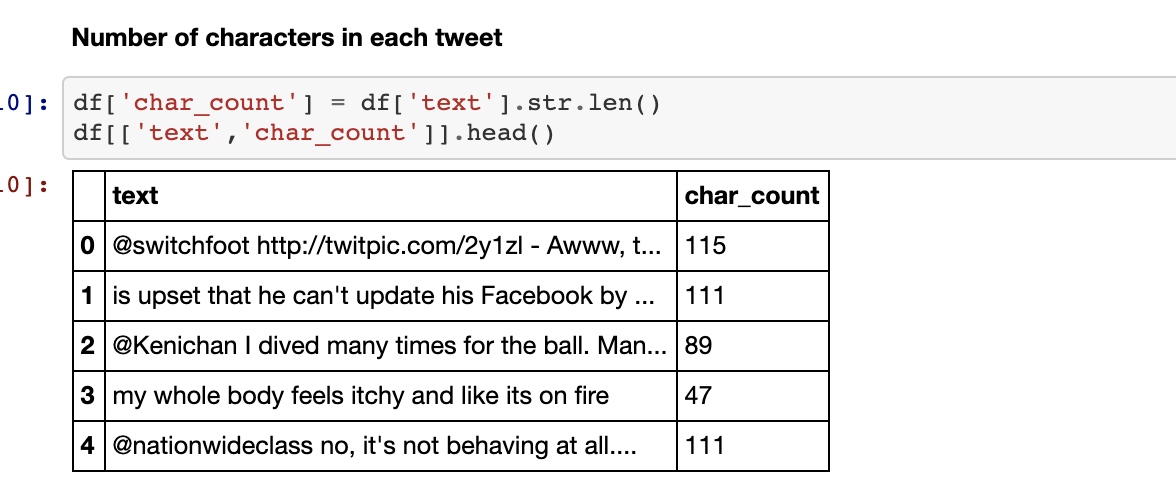
1. **Distribution of Labels:**

****

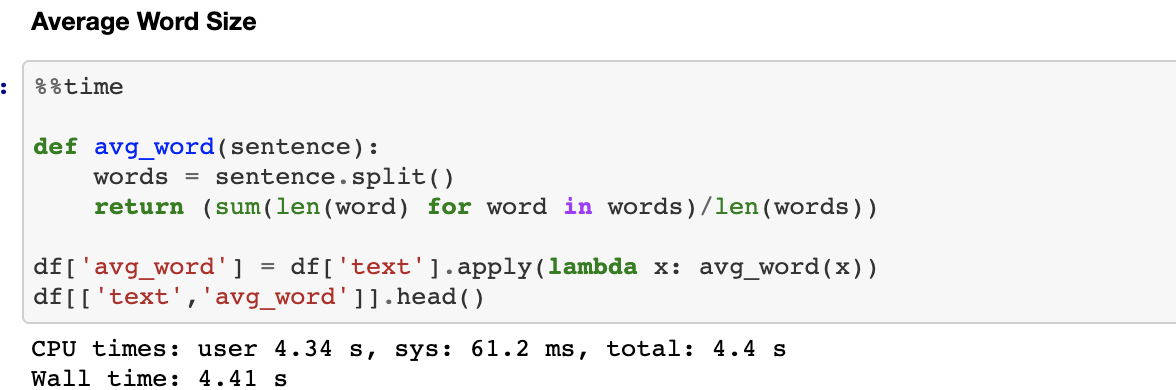
1. **Number of Words in each tweet**

****

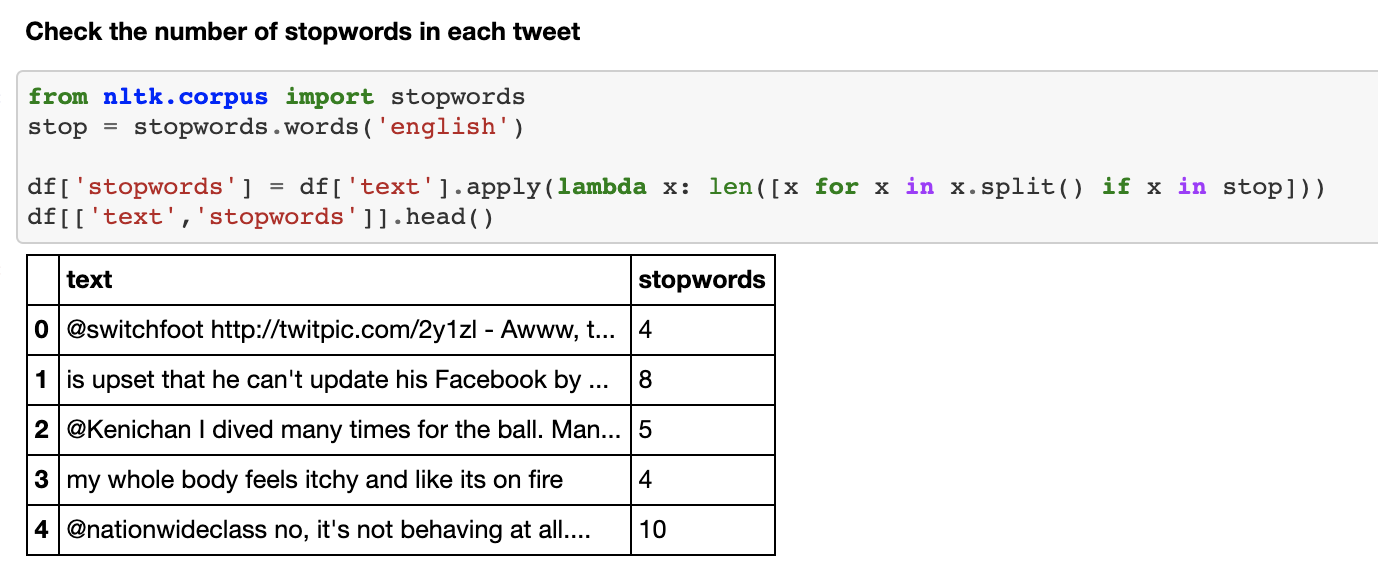
1. **Number of characters in each tweet**

****

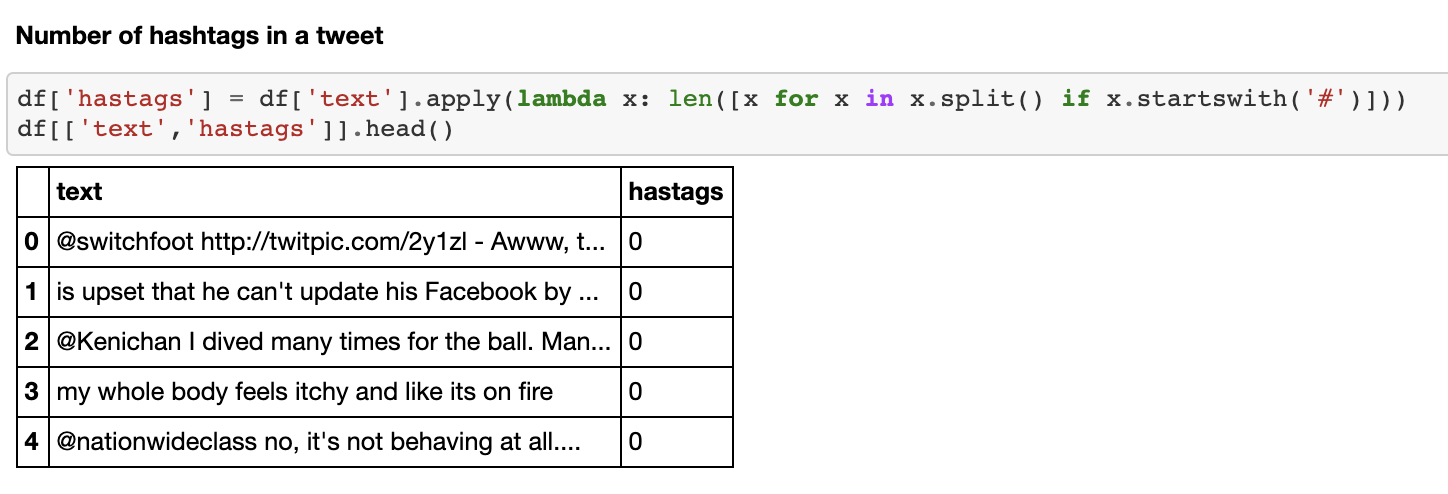
1. **Average Word Size**

****

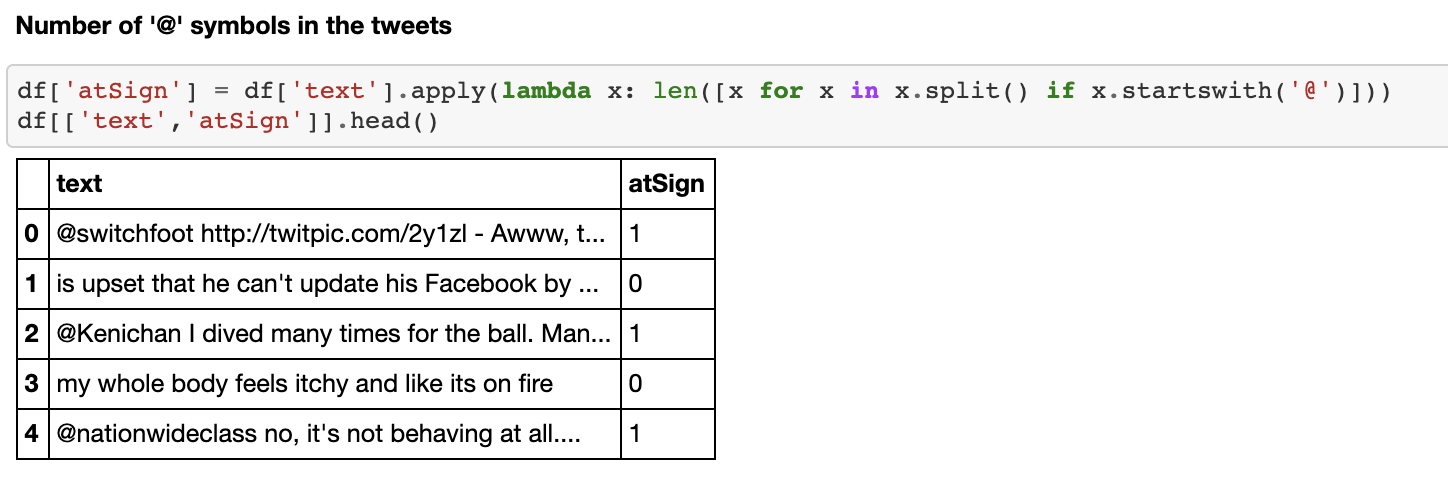
1. **Check the number of stopwords in each tweet**

****

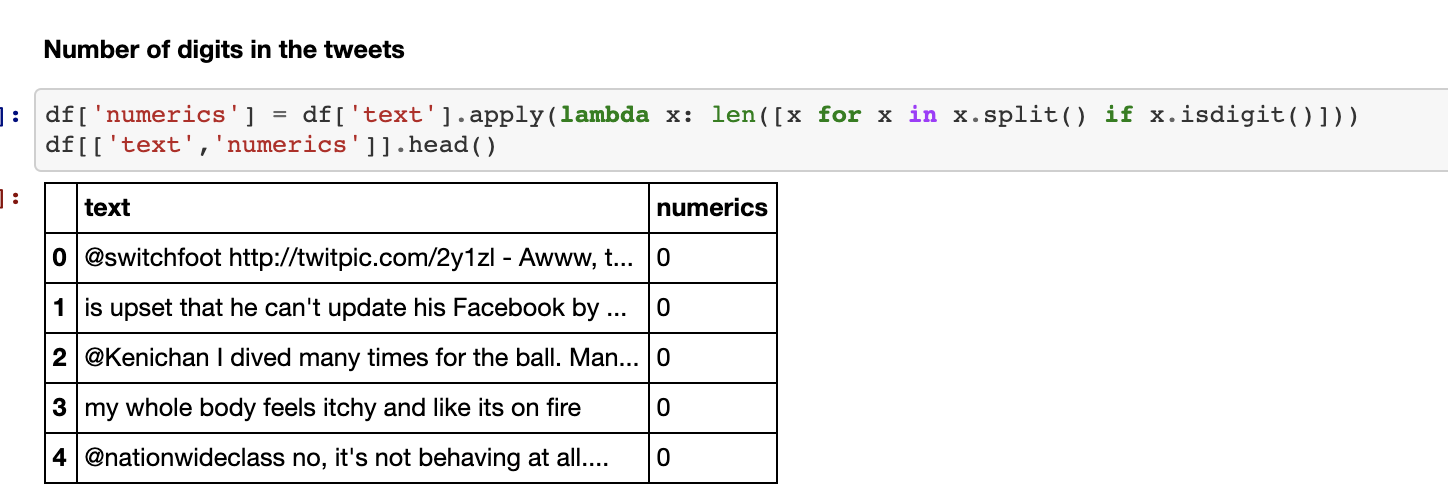
1. **Number of hashtags in a tweet**

****

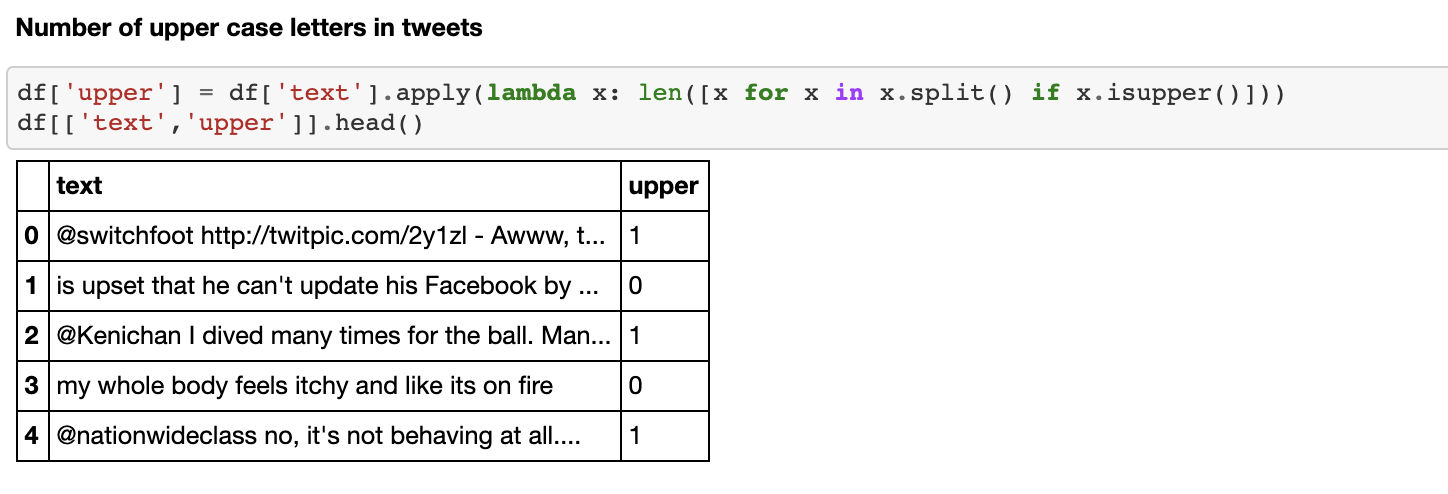
1. **Number of '@' symbols in the tweets**

****

1. **Number of digits in the tweets**

****

1. **Number of upper case letters in tweets**

****