

Text Analytics – Week 4

Program: *IM-KMBI*

Teacher: *R. Henriques, F. Peleja, J. Jagusch*

Date: *Friday, March 9, 2018*

First things first...



I WANT YOU TO TELL ME YOUR NAME!

We happened last week...

- Introduction to GitHub, Anaconda, Jupyter Notebook
- Introduction to the essentials of Python:
 - Data structures
 - Conditional constructs
 - Functional programming
 - Classes
 - Working with text
 - The Trump Tweet program

Let's get things up and running...

- Clone GitHub:

<https://github.com/jbj2505/text-analytics-lecture>

- Download NLTK corpus (in Anaconda prompt):

- python
- import nltk
- nltk.download('all')

- Download Stanford Core NLP:

<https://stanfordnlp.github.io/CoreNLP>

Our agenda for today...

- Normalizing text
 - Expanding contractions
 - Remove special characters
 - Removing stop words
 - Correct spelling
- Tokenizing text
 - Sentence tokenization
 - Word tokenization
- Stemming & Lemmatization
- Parsing
 - Shallow parsing
 - Dependency parsing
 - Constituency parsing