MSc in Data Science

Natural Language Analytics

Academic Year: 2017-2018

Exercise 1: Pre-processing

Delivery Date: **17/04/2018**

Question A: (80%)

You are provided with a Web page from Wikipedia, about artificial neural network. The Web page can be found in the following link:

- Artificial neural network: https://en.wikipedia.org/wiki/Artificial_neural_network.

Using this dataset, you are requested to pre-process the Web page, extract its text and answer the following questions:

1. What is the word count and vocabulary of this Web page?
2. How many sentences are contained in the page?
3. What is the lexical diversity of the page?
4. What are the 5 most common lexical categories (parts of speech)?
5. What are the 10 most common unigrams, the 10 most common bigrams? (please exclude stopwords, using the nltk.corpus.stopwords('english') list)
6. How many nouns are in the page?

Resources that can be potentially helpful:

- NLP with Python: https://github.com/DistrictDataLabs/intro-to-nltk.

Question B: (20%)

Provide a Comma Separated File (CSV) with the following requirements:

- It contains a single row, with 3 cells.
  - The first cell must contain the given name of the student in Greek.
  - The second cell must contain the surname of the student in Greek.
  - The third cell must contain the string "Μεταπτυχιακό στην Επιστήμη Δεδομένων", including quotes.
- The provided file must open in Microsoft Excel, only with a single click from the Windows File explorer, without any further action required by the user. The target operating system is Windows 10, with Greek localisation (Appearance and Local settings).