BTS | Barcelona Technology School

eurecat
Centre Tecnològic de Catalunya

Data Science Foundations

Master in Big Data Solutions 2017-2018

Liana Napalkova
liana.napalkova@bts.tech

Ludovico Boratto
ludovico.boratto@bts.tech

Francisco Guitierres
francisco.gutierres@bts.tech

BTS | Barcelona Technology School

eurecat Centre Tecnològic de Catalunya

Session4

Introduction to Text Mining

Assignments

Assignments

1. How many words are there in `text2`? How many distinct words are there?
2. Compare the lexical diversity scores for humor and romance fiction in the Brown Corpus (`from nltk.corpus import brown`). Which genre is more lexically diverse?
3. Produce a dispersion plot of the four main protagonists in *Sense and Sensibility*: Elinor, Marianne, Edward, and Willoughby. What can you observe about the different roles played by the males and females in this novel? Can you identify the couples?

Assignments

1. Find the collocations in `text5`.
2. Consider the following Python expression: len(set(text4)). State the purpose of this expression. Describe the two steps involved in performing this computation.
3. Many words, like *ski* and *race*, can be used as nouns or verbs with no difference in pronunciation. Can you think of others? Hint: think of a commonplace object and try to put the word to before it to see if it can also be a verb, or think of an action and try to put the before it to see if it can also be a noun. Now make up a sentence with both uses of this word, and run the POS-tagger on this sentence.