Analyze Text

Raúl Maximiliano Urrutia Hernández

November 8th, 2019

Read Text

I used the library BeautifulSoup to extract the text "Cupido Motorizado" of the html.

```
17  #Cleaning the text
18  soup = BeautifulSoup(html,"html5lib")
19
20  #Keep only the main text (without the comments)
21  main = soup.find('section', class_="medium-font-size section-single-content")
22
23  text = main.get_text(strip=True)
```

Then tokenize the text with the "findall" method of the "re" library.

```
25  #Tokenize the text
26  tokens = re.findall(r"[\w']+", text)
27  print(tokens)
```

Then I deleted stopwords with help of "stopwords.words" method of the "nltk" library.

```
31  #Delete stopwords
32  sw = stopwords.words('spanish')
33
34  sw.extend(['Y','iba','iban','La','Lo','Si','si','di','sos','Que','Qué','Se',
35             'No','A','Me','Ni','Estoy','El','Yo','Mi','Esa','Ya','va','Sobre',
36             'sobre','ver'])
37
38  for token in tokens:
39      if token in sw:
40          tokens_wsw.remove(token)
```

Word distribution

Word distribution of the text is shown below. The most used word is “chicos ”