Chris Richardson

Mar 15, 2021

425-426-5481

Github site: https:/github.com/ChrisRichardson/DS_425-426_2021

# ADS 509 Module 3: Group Comparison

Module 3 of Applied Text Mining in Python introduces a group exercise. Groups, ideally, have one knowledgeable spokesperson, a scribe, and a number of other interested participants of customers, sponsors, or different partial parties, work a team formed by different organizations, or students who are all fresh-faced and offered contrasting ideas between groups of like roles, using the Nominal Group Technique.

This assignment asks you to analyze the lyrics and Twitter descriptions for the two artists you selected in Module 1. If the results are interesting, you can do the same for the other artists, you are welcome to use the Spotify Web API and Spotipy for additional lyrics.

- Read in the data, normalize the text, and tokenize it. What are relative your Twitter descriptions, basic hashtags and emoji in use?
- Calculate lexical dispersion to see the rise and fall over and compare the results.
- Pick bigrams of the top nouns. Find the words that are unique to that noun.
- Pick a few tweets to compare.

Each part of the analysis has a section dedicated to it below. Before beginning the analysis there is a section for you to read in the data and do your cleaning (tokenization and normalization).

## Notebook Setup

```
!pip install -U spacy
!python -m spacy download en_core_web_sm
!pip install -U spacy[lookups,transformers]
!pip install -U spacy-lookups-data
!pip install -U spacy-transformers
!pip install -U sentence-splitter
!pip install -U spacy-syllables
```

```
from os import listdir, path, chdir, getcwd, mkdir, rmdir, rename, remove, chmod, walk, sep
from os.path import isfile, isdir, basename, dirname, splitext
from glob import glob, iglob
from shutil import copyfileobj, copytree, rmtree, move
from urllib.request import urlretrieve
from zipfile import ZipFile, ZIP_DEFLATED, BadZipFile, LargeZipFile, zipfile_factory_default, ZipInfo
from tarfile import TarFile, TarInfo, is_tarfile, open ; exclude=[ner,parser]
from spacy.tokenizer import Tokenizer
from spacy.lang.en import English
from re import compile, sub, search, IGNORECASE
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from numpy.random import seed, randint
from numpy import nan, isnan, where, argwhere, count_nonzero, unique, ceil, floor
from pandas import read_csv, DataFrame, Series, concat, pivot_table, cut, qcut
from pickle import dump, load, HIGHEST_PROTOCOL
from datetime import datetime, date, time, timedelta
from sentiment_analysis_spacy.sentiment_analysis_spacy_module.sentiment_analysis_spacy_module import SentimentAnalysis
from spacy import load, prefer_gpu, require_gpu, info
from spacy_syllables import SpaCySyllables
from spacy_lookups_data import load_lookup_from_package
from spacy_transformers import TransformerModel
from sentence_splitter import SentenceSplitter, split_text_into_sentences
from collections import Counter
from itertools import chain
from pprint import pprint
from sys import stdout, path, modules
from os.path import abspath
from re import sub, compile, IGNORECASE
from spacy.tokens.token import Token
from spacy.matcher.phrasematcher import PhraseMatcher
from spacy.matcher.matcher import Matcher
from spacy.lang.en.stop_words import STOP_WORDS
from spacy.lang.en import English
from spacy.tokens import Doc, DocBin
from spacy import displacy
from numpy.random.mtrand import seed
from numpy.lib.arraysetops import unique
from numpy.core.fromnumeric import argmax, argmin
from numpy.core.numeric import where, count_nonzero
from numpy.lib.function_base import vectorize
from numpy.core.numerical_types import nan, isnan
from numpy.core.fromnumeric import ceil, floor
from pandas.core.reshape.api import pivot_table, cut, qcut
from pandas.core.tools.datetimes import to_datetime
from pandas.core.dtypes.cast.construct_ndarray import construct_1_dimensional_ndarray
from pandas.core.dtypes.dtypes import CategoricalDtype
from pandas.core.dtypes.common import is_categorical_dtype, is_bool_dtype, is_datetime64_any_dtype, is_datetime64tz_dtype, is_timedelta64_dtype, is_sparse, is_scipy_sparse, is_numeric_v_string_like, is_extension_array_dtype, is_object_dtype, is_dtype_equal, is_int64_dtype
from pandas.core.dtypes.cast.infer_dtype_from_scalar import infer_dtype_from_scalar
from pandas.core.dtypes.cast.dtypes_and_scalars.common_dtypes import find_common_type
from pandas.core.dtypes.cast.dtypes_and_scalars.dtypes_and_scalars.dtypes_and_scalars import find_common_type_with_category_has_ordered
from pandas.core.dtypes.cast.dtypes_and_scalars.dtypes_and_scalars.dtypes_and_scalars.dtypes_and_scalars import find_common_type_with_category_has_ordered_with_bool
from pandas.core.dtypes.cast.dtypes_and_scalars.dtypes_and_scalars.dtypes_and_scalars.dtypes_and_scalars.dtypes_and_scalars import find_common_type_with_category_has_ordered_with_bool_or_int64
from pandas.core.dtypes.cast.dtypes_and_scalars.dtypes_and_scalars.dtypes_and_scalars.dtypes_and_scalars.dtypes_and_scalars.dtypes_and_scalars import find_common_type_with_category_has_ordered_with_bool_or_int64_or_datetime64
from pandas.core.dtypes.cast.dtypes_and_scalars.dtypes_and_scalars.dtypes_and_scalars.dtypes_and_scalars.dtypes_and_scalars.dtypes_and_scalars.dtypes_and_scalars import find_common_type_with_category_has_ordered_with_bool_or_int64_or_datetime64_or_timedelta64
from pandas.core.dtypes.cast.dtypes_and_scalars.dtypes_and_scalars.dtypes_and_scalars.dtypes_and_scalars.dtypes_and_scalars.dtypes_and_scalars.dtypes_and_scalars.dtypes_and_scalars import find_common_type_with_category_has_ordered_with_bool_or_int64_or_datetime64_or_timedelta64_or_period
from pandas.core.dtypes.cast.dtypes_and_scalars.dtypes_and_scalars.dtypes_and_scalars.dtypes_and_scalars.dtypes_and_scalars.dtypes_and_scalars import find_common_type_with_category_has_ordered_with_bool_or_int64_or_datetime64_or_timedelta64_or_period_or_sparse
from pandas.core.dtypes.cast.dtypes_and_scalars.dtypes_and_scalars import find_common_type_with_category_has_ordered_with_bool_or_int64_or_datetime64_or_timedelta64_or_period_or_sparse_or_object
from pandas.core.dtypes.cast.dtypes_and_scalars import find_common_type_with_category_has_ordered_with_bool_or_int64_or_datetime64_or_timedelta64_or_period_or_sparse_or_object_or_string
from pandas.core.dtypes.cast import find_common_type_with_category_has_ordered_with_bool_or_int64_or_datetime64_or_timedelta64_or_period_or_sparse_or_object_or_string_or_unicode
from pandas.core.dtypes.cast import find_common_type_with_category_has_ordered_with_bool_or_int64_or_datetime64_or_timedelta64_or_period_or_sparse_or_object_or_string_or_unicode_or_bytes
from pandas.core.dtypes.cast import find_common_type_with_category_has_ordered_with_bool_or_int64_or_datetime64_or_timedelta64_or_period_or_sparse_or_object_or_string_or_unicode_or_bytes_or_bool
from pandas.core.dtypes.cast import find_common_type_with_category_has_ordered_with_bool_or_int64_or_datetime64_or_timedelta64_or_period_or_sparse_or_object_or_string_or_unicode_or_bytes_or_bool_or_int64
from pandas.core.dtypes.cast import find_common_type_with_category_has_ordered_with_bool_or_int64_or_datetime64_or_timedelta64_or_period_or_sparse_or_object_or_string_or_unicode_or_bytes_or_bool_or_int64_or_datetime64
from pandas.core.dtypes.cast import find_common_type_with_category_has_ordered_with_bool_or_int64_or_datetime64_or_timedelta64_or_period_or_sparse_or_object_or_string_or_unicode_or_bytes_or_bool_or_int64_or_datetime64_or_timedelta64
from pandas.core.dtypes.cast import find_common_type_with_category_has_ordered_with_bool_or_int64_or_datetime64_or_timedelta64_or_period_or_sparse_or_object_or_string_or_unicode_or_bytes_or_bool_or_int64_or_datetime64_or_timedelta64_or_period
from pandas.core.dtypes.cast import find_common_type_with_category_has_ordered_with_bool_or_int64_or_datetime64_or_timedelta64_or_period_or_sparse_or_object_or_string_or_unicode_or_bytes_or_bool_or_int64_or_datetime64_or_timedelta64_or_period_or_sparse
from pandas.core.dtypes.cast import find_common_type_with_category_has_ordered_with_bool_or_int64_or_datetime64_or_timedelta64_or_period_or_sparse_or_object_or_string_or_unicode_or_bytes_or_bool_or_int64_or_datetime64_or_timedelta64_or_period_or_sparse_or_object
from pandas.core.dtypes.cast import find_common_type_with_category_has_ordered_with_bool_or_int64_or_datetime64_or_timedelta64_or_period_or_sparse_or_object_or_string_or_unicode_or_bytes_or_bool_or_int64_or_datetime64_or_timedelta64_or_period_or_sparse_or_object_or_string
from pandas.core.dtypes.cast import find_common_type_with_category_has_ordered_with_bool_or_int64_or_datetime64_or_timedelta64_or_period_or_sparse_or_object_or_string_or_unicode_or_bytes_or_bool_or_int64_or_datetime64_or_timedelta64_or_period_or_sparse_or_object_or_string_or_unicode
from pandas.core.dtypes.cast import find_common_type_with_category_has_ordered_with_bool_or_int64_or_datetime64_or_timedelta64_or_period_or_sparse_or_object_or_string_or_unicode_or_bytes_or_bool_or_int64_or_datetime64_or_timedelta64_or_period_or_sparse_or_object_or_string_or_unicode_or_bytes
from pandas.core.dtypes.cast import find_common_type_with_category_has_ordered_with_bool_or_int64_or_datetime64_or_timedelta64_or_period_or_sparse_or_object_or_string_or_unicode_or_bytes_or_bool_or_int64_or_datetime64_or_timedelta64_or_period_or_sparse_or_object_or_string_or_unicode_or_bytes_or_bool
from pandas.core.dtypes.cast import find_common_type_with_category_has_ordered_with_bool_or_int64_or_datetime64_or_timedelta64_or_period_or_sparse_or_object_or_string_or_unicode_or_bytes_or_bool_or_int64_or_datetime64_or_timedelta64_or_period_or_sparse_or_object_or_string_or_unicode_or_bytes_or_bool_or_int64
from pandas.core.dtypes.cast import find_common_type_with_category_has_ordered_with_bool_or_int64_or_datetime64_or_timedelta64_or_period_or_sparse_or_object_or_string_or_unicode_or_bytes_or_bool_or_int64_or_datetime64_or_timedelta64_or_period_or_sparse_or_object_or_string_or_unicode_or_bytes_or_bool_or_int64_or_datetime64
from pandas.core.dtypes.cast import find_common_type_with_category_has_ordered_with_bool_or_int64_or_datetime64_or_timedelta64_or_period_or_sparse_or_object_or_string_or_unicode_or_bytes_or_bool_or_int64_or_datetime64_or_timedelta64_or_period_or_sparse_or_object_or_string_or_unicode_or_bytes_or_bool_or_int64_or_datetime64_or_timedelta64
from pandas.core.dtypes.cast import find_common_type_with_category_has_ordered_with_bool_or_int64_or_datetime64_or_timedelta64_or_period_or_sparse_or_object_or_string_or_unicode_or_bytes_or
```

## Rezzy Word Clouds

### Twitter Word Clouds

```
In [1]: %load_ext rezzywordcloud
        %rezzy --twitter --twitter_user_id 1234567 --stopwordsfilter
```

### Lyrics Word Clouds

```
In [2]: %load_ext rezzywordcloud
        %rezzy --lyrics --lyrics_artist Taylor Swift --stopwordsfilter
```




Q: What observations do you have about these (relatively straightforward) word clouds?

Although we ran the Twitter Cloud for Shenee and artist from the, like, music, and time. What separates the two Twitter clouds is that Shenee's cloud contains other words such as the following:

- year
- month
- number
- wife
- boyfriend
- home
- ex

Most of Shenee's appear to be a younger crowd utilizing words such as:

- today
- tomorrow
- ig
- ur
- u
- instagram
- tweet