Assignment 4: Text Data

Data Modelling

- NLTK
- Gensim
- Textblob
- Sklearn (TfidfVectorizer, TfidfTransformer)

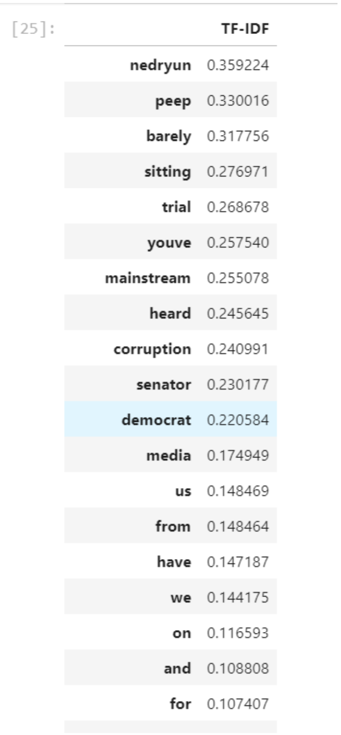
Text Pre-processing:

Hyperlinks, Stop-words, punctuations, emojis, hashtags, numbers, typos, lowercase, other noisy data…

Stored Emojis with description and Hashtags.

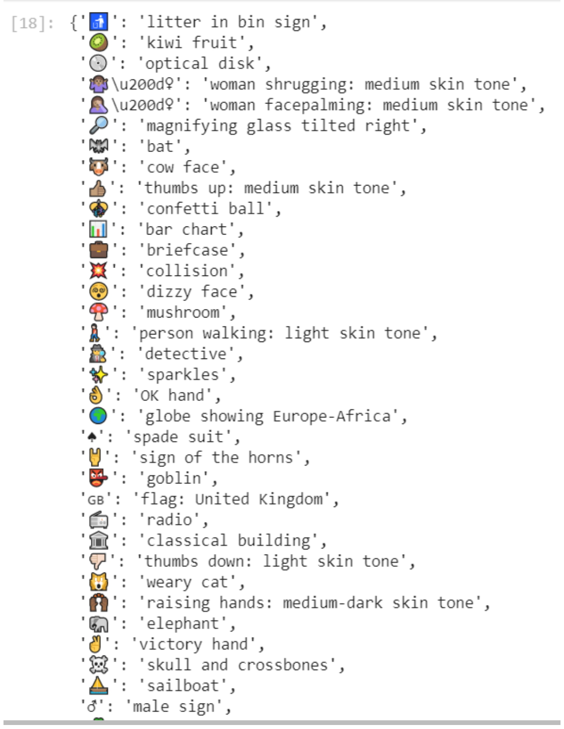
Word Embeddings:

- Word2Vector
- Sentiment Analysis (polarity, subjectivity of a tweet)
- TF-IDF



[25]:

	TF-IDF
nedryun	0.359224
peep	0.330016
barely	0.317756
sitting	0.276971
trial	0.268678
youve	0.257540
mainstream	0.255078
heard	0.245645
corruption	0.240991
senator	0.230177
democrat	0.220584
media	0.174949
us	0.148469
from	0.148464
have	0.147187
we	0.144175
on	0.116593
and	0.108808
for	0.107407



```
[18]: {'🚮': 'litter in bin sign',
       '🥝': 'kiwi fruit',
       '💿': 'optical disk',
       '🤷🏽\u200d♀': 'woman shrugging: medium skin tone',
       '🤦🏽\u200d♀': 'woman facepalming: medium skin tone',
       '🔎': 'magnifying glass tilted right',
       '🦇': 'bat',
       '🐮': 'cow face',
       '👍🏽': 'thumbs up: medium skin tone',
       '🎊': 'confetti ball',
       '📊': 'bar chart',
       '💼': 'briefcase',
       '💥': 'collision',
       '😵': 'dizzy face',
       '🍄': 'mushroom',
       '🚶🏻': 'person walking: light skin tone',
       '🕵': 'detective',
       '✨': 'sparkles',
       '👌': 'OK hand',
       '🌍': 'globe showing Europe-Africa',
       '♠': 'spade suit',
       '🤘': 'sign of the horns',
       '👺': 'goblin',
       'GB': 'flag: United Kingdom',
       '📻': 'radio',
       '🏛': 'classical building',
       '👎🏻': 'thumbs down: light skin tone',
       '🙀': 'weary cat',
       '🙌🏾': 'raising hands: medium-dark skin tone',
       '🐘': 'elephant',
       '✌': 'victory hand',
       '☠': 'skull and crossbones',
       '⛵': 'sailboat',
       '♂': 'male sign',
```

From Research papers

An Exploratory Study of Word-Scale Graphics:

SparkClouds:

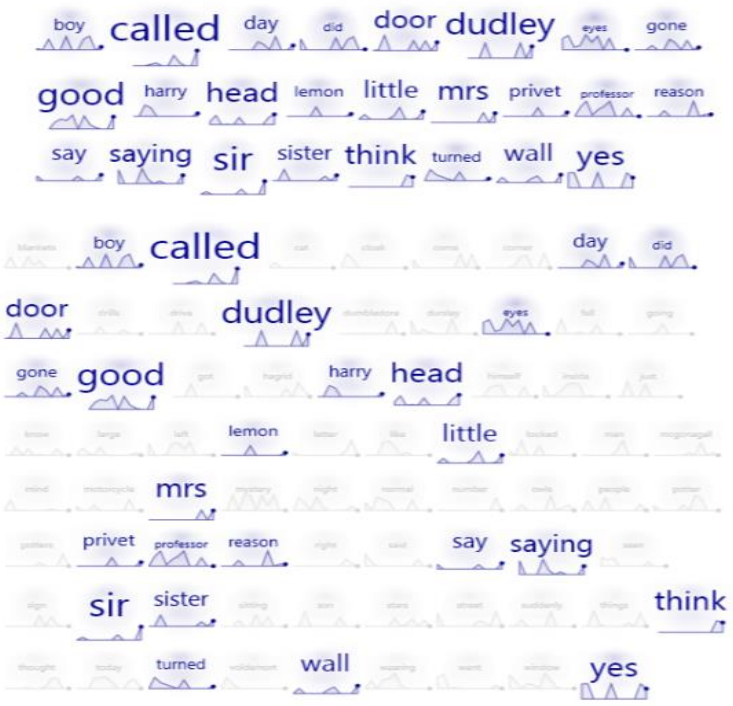


Fig. 1. SparkClouds showing the top 25 words for the last time point (12th) in a series. 50 additional words that are in the top 25 for the other time points can be (top) filtered out or (bottom) shown in gray at a smaller fixed-size font. (bottom) is used in the study.

ParallelCloud:

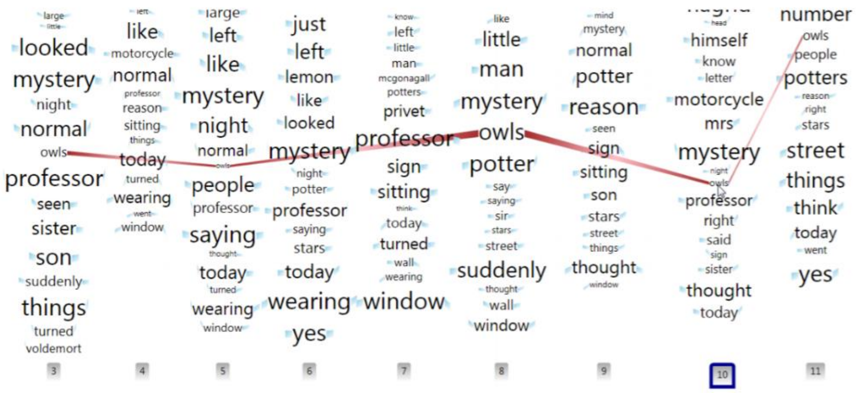


Fig. 4. ParallelCloud displays a gradient line that links the same word occurring in multiple tag clouds when people move the cursor over a word.

Pictures taken from https://www.microsoft.com/en-us/research/wp-content/uploads/2010/01/sparkclouds_infovis2010.pdf