Welcome to

instats

The Session Will Begin Shortly

START

Text Analysis Using Python

Session 5: Fundamentals of Supervised Text Analysis

Outline

1. Supervised techniques:
    1. Keyword searching / KWIC
    2. Sentiment analysis

Supervised techniques: keywords

A simple but surprisingly powerful approach to text analysis is to search for keywords in documents.

For example, does a document contain the words “medicine” or “medical”?

Useful for identifying prevalence of terms or topics across a corpus.

Easily extended to phrases e.g., “medical research” or “public health”.

Unsophisticated approach i.e., misses context, synonyms, polysemic words.

Supervised techniques: keywords

An extension is **Key Word in Context (KWIC)**, which finds a keyword and returns the context in which it is found.

For example, "…this charity funds **medical research** in East Africa…"

Useful for understanding the context in which words are used (thus addressing concerns around polysemy, dynamic contexts).

Results can vary depending on context window.

Supervised techniques: sentiment analysis

We can think of documents as belonging to categories or classes (Spirling, 2022). We can known categories or classes and measure the extent to which our documents align with these.

For example, is a product review positive or negative?

To answer we can look at words used in the review and see how they align with a dictionary of words or statements **that are known to be positive or negative.**

Supervised techniques: sentiment analysis

$$Y = X + e$$

Where:

Y = Labelled outcome e.g., 1 = positive review

X = Document features e.g., words and/or frequencies

e = Uncertainty in our prediction

Supervised techniques: sentiment analysis

Basic idea: use a set of pre-defined words with specific connotations to classify our documents automatically, quickly and accurately (Spirling, 2022).

Dictionary-based method.

Aim: measure *whether* and *the extent* to which a document is associated with a given category (sentiment).

$$\sum_{m=1}^{M} \frac{S_m W_{im}}{N_i}$$

Supervised techniques: sentiment analysis

“The Pogues’ Body of an American is a raw, energetic anthem. The driving rhythm, powerful lyrics, and Shane MacGowan’s passionate vocals make it an unforgettable and emotional listening experience.”

- Is this a positive review?
- What words are you using in particular to make your judgement?
- Do you notice any downsides of treating text in this manner?

Supervised techniques: sentiment analysis

“The Pogues’ Body of an American is a raw, energetic anthem. The driving rhythm, powerful lyrics, and Shane MacGowan’s passionate vocals make it an unforgettable and emotional listening experience.”

Overall score = 0.24

Word scores = {"raw": -0.1, "energetic": 0.8, "anthem": 0.5, "driving": 0.3, "rhythm": 0.2, "powerful": 0.9, "lyrics": 0.3, "passionate": 0.8, "vocals": 0.4, "unforgettable": 1.0, "emotional": 0.7, "listening": 0.2, "experience": 0.3}

Supervised techniques: sentiment analysis

Some issues (for this example and in general):

- Loss of context (is unforgettable a positive or negative word on its own?).
- Detect sarcasm?
- Granularity i.e., can this approach really distinguish between categories close together?
- How do we know what sentiment scores to assign?
- How do we know which dictionary is best to use?

STOP