Lancaster University

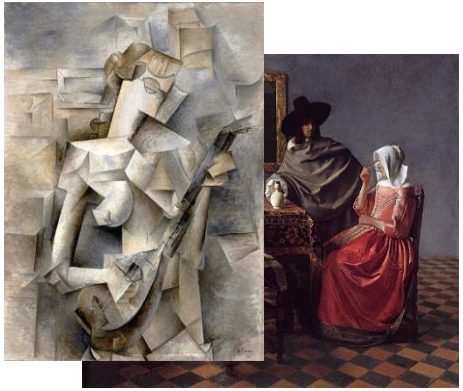
Using Natural Language Processing to Analyse the Shape of Stories

29th May 2020

Luca Davies – B.Sc. (Hons.) Computer Science

Lancaster University

Why?



- Patterns exist in all forms of media
- Art movements, genres of film, book, music, video game
- All of these have patterns
- Sometimes the “template” used is more obvious than others…

How?
Input Text Pre-processing

See Sec. 3

Input media types require different handling

Plain text – additional white space and line breaks removed

XML – VARDed corpora only, tags and extraneous non-text data removed

HTML – Project Gutenberg format, split into chapters where possible based on HTML header tags, then move extraneous, non-text data

How?
Natural Language Processing Methods

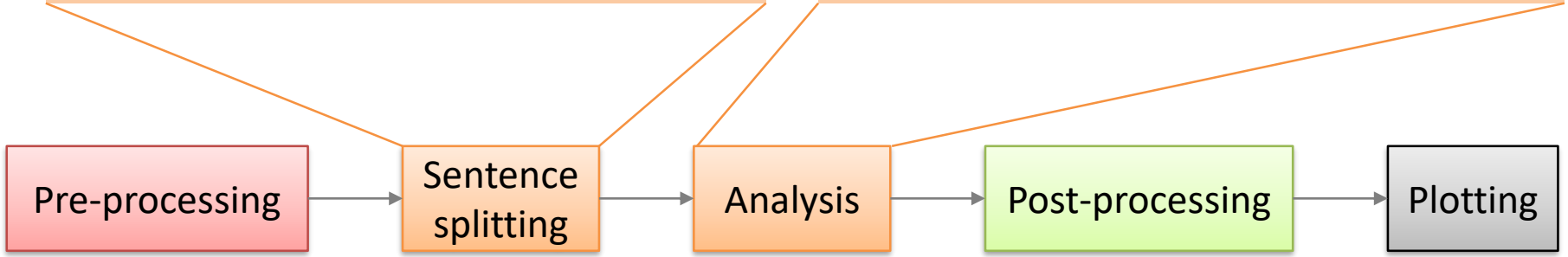
See Sec. 3

Stanford CoreNLP

- Sentence splitting (via “Simple API”)
- Single function call

VADERSharp

- Processing of whole sentences
- Outputting scores per sentence



How?
Collecting VADER's output in a useful way

See Sec. 3

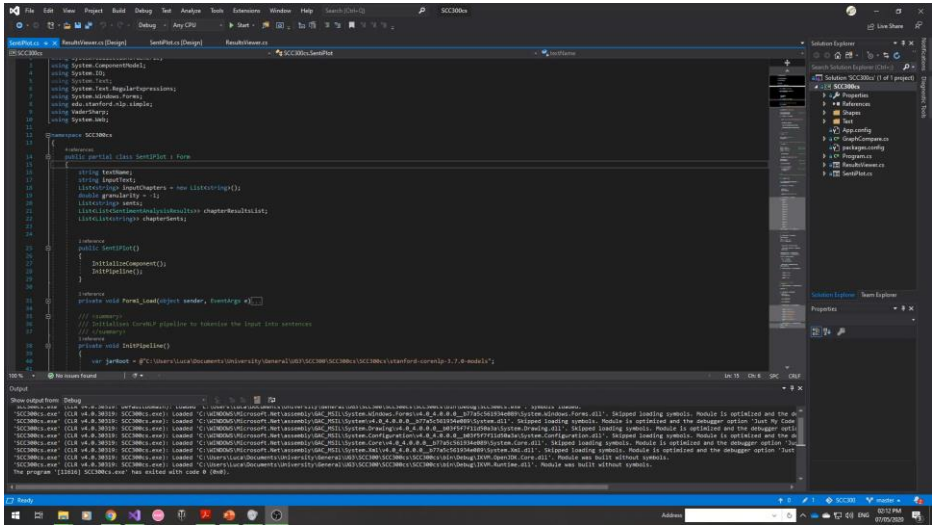
- Collecting the list of analysis results in individual chapters and the whole text
- Generate a new ResultsViewer form and pass the collections through to be plotted

How?
Plotting the data for presentation

See Sec. 3

- ResultsViewer is passed *n* result sets
  - Each contains four result types (combined, positive, negative, neutral), all plotted on whole-text graph, combined only plotted per chapter
- Data points are plotted by averaging *m* values on the sentiment (y) axis and using the final value on the time-progression axis (x)
  - *m* is the number of sentences equating to the percentage of the text selected by the granularity slider

The Sentiplot Tool



Data & Experiment Selection

See Sec. 4/5

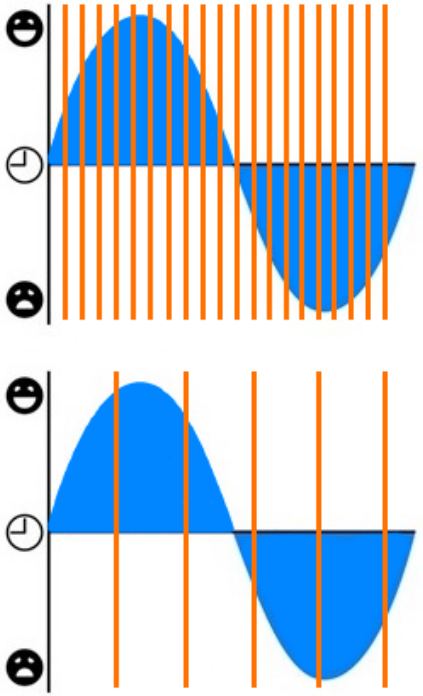
- Texts chosen directly influenced by:
  - Vonnegut's examples
  - My personal literary knowledge

- Experiments chosen by:
  - Vonnegut's rejected thesis (*is there a shape at all?*)
  - Logical further steps arising from previous experiments

Data & Experiment Selection

See Sec. 5

- Analysis Block Size *pg. 17*
- Curve Identification (Whole text / per chapter) *pg. 19*
- Hand Analysis vs. VADER *pg. 20*
- Reader Analysis & Reflection *pg. 22*
- Early Modern Vs. Modern English *pg. 24*



Story graphs from openculture.com article modified to show only a single graph, blank with added lines

Experimentation in a nutshell: Early Modern vs. Modern English

Lancaster University

See Sec. 5.5

This experiment focused exclusively processing Shakespearean plays:

- With original early-modern English
- With spelling-normalised, modern English (as produced via VARD)

Results: Similar pattern, but visibly exaggerated, even some new features entirely. Proof, less that this is better for VADER, more that EModE is poorly recognised by comparison

Conclusion: When attempting to perform sentiment analysis on EModE texts, results may be significantly easier to analyse if the text is first subject to spelling normalisation.

Limitations

- Range of corpora
  - Relatively narrow range of texts examined
  - Small number of texts
- Ability of NLP tools
  - Other tools may exist that are better suited to the analysis carried out
- Inference vs. Statistical Methods
  - Identification of curves *could* be implemented mathematically

Lancaster University

Questions…

29th May 2020

Luca Davies – B.Sc. (Hons.) Computer Science