Digital Teaching Integration Initiative

Northeastern University
NULab for Texts, Maps, and Networks

ENGL 3370: Summer I
Computational Text Analysis

Important Terminology

- Corpus/corpora: a text or collection of multiple texts that is used for analysis.
  - Example: One could could create corpus of all of Barack obama's presidential speeches to trace his language over the course of his presidency
- nGram: A continuous sequence of n items in a text. For example, in Barack Obama's speeches, a bigram (or 2 continuous words) could be 'United States,' while a tripgram (3 words) could be 'yes we can.'
- Stopwords: commonly used words that are part of natural language usage, but typically do not add meaning to a sentence but can add context. Stopwords are commonly ignored in text analysis, but this depends on the questions being asked.
  - Examples: the, but, this, that

Voyant: https://voyant-tools.org/

Voyant is a powerful web-based text analysis platform "designed to facilitate reading and interpretive practices for digital humanities and scholars as well as for the general public. With Voyant, you can analyze one or more text files and use multiple visualization tools.

Voyant can read .pdf, .doc, but .txt is recommended.

For Voyant documentation/guide see: https://voyant-tools.org/docs/#!/guide/about

Step-by-Step Guide to use Voyant:

1. On the Voyant front page, click on 'upload' and navigate to the chosen corpus to upload text files - then click on 'reveal.'
   a. Alternatively insert URLs or full text into textbox
   b. Advanced options and help can be found in top right of text box.
2. The results are given in five default panes that can be changed.
3. To change a pane:
   a. Hover mouse over one of the panes.
   b. Click the panes button: ⊞
   c. Use the dropdown menu to navigate to different tools.

Alternative Web Tools for Text Analysis

WordCounter: https://databasic.io/en/wordcounter/
WordCounter analyzes a corpus to count words and n-grams.
Word counts, bigram, and trigram data can then be download as a .csv for further analysis.

Step-by-step WordCounter guide:

1. To use your own text, select: paste text, upload a file, or paste a link.
2. Click on count (Note that 'ignore case' and 'ignore stopwords' is selected by default).
3. WordCounter outputs as a word cloud and a list of top words, bigrams, and trigrams.
4. The researcher can output a .csv of top nGrams (n=1, 2, & 3) by scrolling down.

SameDiff: https://databasic.io/en/samediff/
SameDiff compares one corpus or text to another corpus or text and tells the user how similar they are based upon a cosine similarity algorithm.

Step-by-Step SameDiff Guide:

1. Select 'upload files' and upload the two texts you want to compare
2. Click on compare.
3. SameDiff outputs a similarity score, total word counts, and the specific words that are similar and the words that differentiate the two documents.
4. The researcher can output a .csv of word counts accessed by scrolling down

Storybench Textual Analysis: https://storybench.shinyapps.io/textanalysis/
Storybench uses sentiment analysis and nGrams to analyze a .txt or .csv fie. Sentiment analysis uses a predetermined dictionary that measures words as positive or negative.

Step-by-Step Storybench Guide:

1. Merge your corpus into one .txt file
2. Upload your .txt file by clicking "Browse" or simply dragging-and-dropping
3. The researcher can save their results by taking screenshots

Github link: https://www.bit.ly/NUlabDTI Folder: textanalysis > Lewis-IncarcerationArchives

Contact

If you have questions, free to contact us:
Cara Marta Messina: messina.c@husky.neu.edu

Garrett Morrow: morrow.g@husky.neu.edu