What questions can humanists and social scientists find answers to with the computational tools in the NLP Suite?

Contents

Humanists' and social scientists' questions	**Computer scientists' answers and available NLP Suite Tools**
Preparing your documents for NLP algorithms	

How many documents do I need? Most NLP tools will allow you to work with one or several documents in a directory (your *corpus)*. Some NLP tools, however, will only work with an input corpus of many documents (e.g., topic modeling or shape of stories). *What kind of documents do I need?* NLP tools typically take .txt files in input. There are, however, several algorithms in the NLP Suite that will allow you to convert files from one type to the other (e.g., docx or rtf to txt).	*Under Pre-Processing Tools* select the tool that you need among the wide range of available tools.
What's in your documents? A broad view	
Is there a way to get basic **statistics about my *corpus***, i.e., the set of documents you are studying? For example, the number of documents, the number of words per document, the sentence length per document? *Corpus statistics*	*Under CORPUS/DOCUMENT analysis tools select:* • Corpus/document(s) statistics (Sentences, words, lines)
What do your documents (corpus) talk about? What are the topics?	*Under CORPUS analysis tools select:* • Topic modeling (via MALLET & Gensim)
What's in your documents? A closer look	
Syntactic relations between words	
Can computational tools tell me **which words in a text are nouns, verbs, or adjectives**? Yes. There are basic tools called **parsers** that do precisely that, and more and with a high degree of accuracy (over 90%). They can even give the syntactical relation between words, whether verbs are in the infinitive form, gerundive, past, present, passive or active, whether nouns are singular or plural subjects or objects, and, again, active or passive.	The NLP Suite has several different options for parsers: spaCy, Stanford CoreNLP, Stanza, each with slightly different characteristics of accuracy or speed. *Under CORPUS/DOCUMENT analysis tools select:*

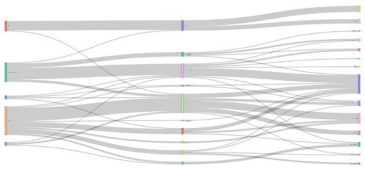
	• Parsers & annotators (BERT, CoreNLP, spaCy, Stanza)
Can they even tell me **which adjectives are used for which nouns**? Like in a study of folktales, which adjectives are used for princesses and princes? *Grammatical analysis of text* Definitely. Parsers produce in output a CoNLL table that contains all the information required to answer those questions.	In the NLP Suite, the CoNLL table analyzer algorithms will allow you to explore these questions. *Under CORPUS/DOCUMENT analysis tools select:* • CoNLL table analyzer - Search the CoNLL table
Who does what, where, when, and why? Can NLP help to identify the **5Ws of stories**? Subject-Verb-Object (SVO) extraction and visualizations (the 5Ws of narrative)	*Under CORPUS/DOCUMENT analysis tools select:* • SVO (Subject-Verb-Object) extractor & visualization
There is a problem in linguistics known as **pronominal anaphora** where in such a sentence as "John hurried home. He said he was ill," where both "he" pronouns (the anaphoric expression) refer to "John" (the antecedent). Can computational tools link pronouns to their nouns (anaphoric expressions to their antecedents)? *Pronominal anaphora and coreference* In NLP, this problem is known as **coreference resolution**. Unfortunately, the accuracy of these algorithms is not as high as for parsers, even for neural network approaches (around 70%?).	There are two points of entry to coreference resolution the NLP Suite and with several different options available. *Under CORPUS/DOCUMENT analysis tools select:* • CoreNLP annotator - coreference (pronominal) • Parsers & annotators (BERT, CoreNLP, spaCy, Stanza)
Semantic relations between words	
Which words come together in texts? Are women/girls mostly associated with adjectives or nouns of physical beauty, family relations, and low-paying occupations in the care industry (waiters, teachers, nurses)? Is the opposite true for men and boys?	*Under CORPUS/DOCUMENT analysis tools select:*

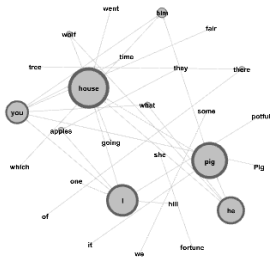
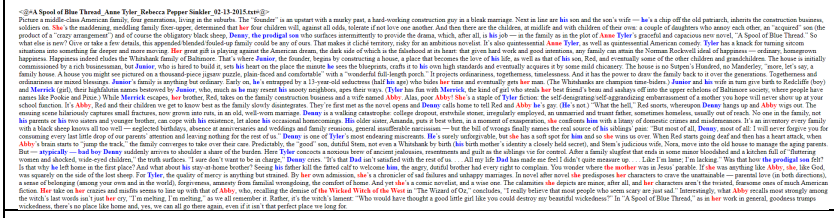
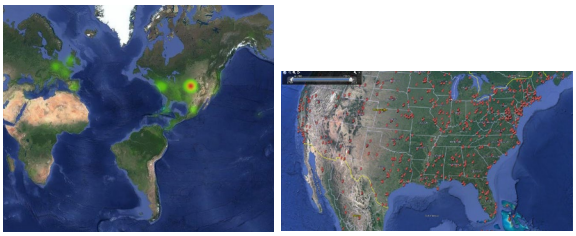
	• Word embeddings (Word2Vec) (via BERT & Gensim)
Do **nouns and verbs cluster** together in specific groups (e.g., verbs of movement or verbs of communication)?	*Under CORPUS/DOCUMENT analysis tools select:* • WordNet
Searching documents	
Besides these powerful NLP tools of syntactic and semantic analysis, are there simpler search tools that would allow me to zoom into a text for specific words or combinations of words? Search words in documents The NLP Suite offers a number of options for searching documents with specific words in mind. Each of these options will allow you to extract the sentence (or a contour of sentences) that contain the search words. And if you work with a large corpus, you can even create a subsample of files that contain the search words for separate analyses. Search the CoNLL table You can search the CoNLL table, finding, perhaps, all the adjectives used to characterize a specific noun (e.g., prince or princess). Search documents for n-grams and co-occurrences The NLP Suite has an n-grams/co-occurrences viewer, similar to Google Ngram Viewer (https://books.google.com/ngrams/) but applied to your own corpus, rather than Google books. The viewer can search for words or combinations of words (e.g., "love letter") and plot them overtime if filenames embed a date (e.g., The New York Times_12-01-1992). Search with knowledge-graph tools: DBpedia, YAGO, WordNet You can search your corpus for words found in what are known as knowledge-graph tools: DBpedia or YAGO or WordNet, the Princeton University semantic database. Do the	*Under CORPUS/DOCUMENT analysis tools select:* • Search (ALL options GUI)

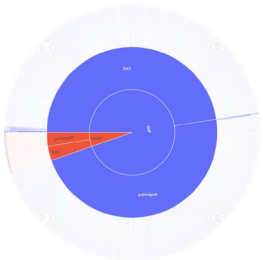
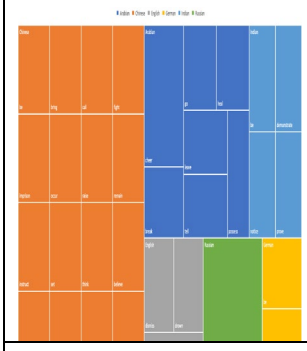
documents contain words in DBpedia or YAGO ontology classes (e.g., emotions, writer) or in WordNet word categories ("synsets", e.g., verbs of movement, such as go, come, run)?	
Getting information from the web about organizations, people, locations	
If documents in your corpus mention proper names (people, organizations, locations) can you find **information on the web** automatically? For instance, when and where someone was born or their occupation?	*Under CORPUS/DOCUMENT analysis tools select:* • Knowledge graphs: DBpedia & YAGO
Sentiments and emotions	
Which **sentiments and emotions** do writers express in their writing? Positive, negative? How do they express these emotions rhetorically and linguistically?	The NLP Suite provides several options to address these questions. To find all the options *Under CORPUS analysis tools select:* • Sentiments/emotions (ALL options GUI) The GUI will give you access to all the various options available in the NLP Suite to find answers to those questions. You can also select • Style analysis (ALL options GUI) Then tick the Vocabulary analysis option and select: Punctuation as figures of pathos (?!)

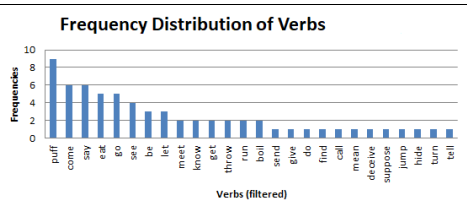
	Or the Repetition tools (repetition is also a figure of pathos). Sentiment analysis (neural network/tensor options: BERT, spaCy, Stanford CoreNLP, Stanza) Sentiment analysis (dictionary options: ANEW, hedonometer, SentiWordNet, VADER) You can also use YAGO's emotion ontology class or WordNet (NOUN feeling VERB emotion) to search for emotion words used in the texts
And when a single story is viewed in the context of hundreds or thousands of other stories, do these stories cluster together in specific **story shapes**, as Kurt Vonnegut would have it? A man in a hole, from rags to riches or riches to rags? Happy ending or a bummer? Comedy or tragedy?	*Under CORPUS analysis tools select:* • Shape of stories
Writing style	
Which **style** do different authors use in their writing? Can the NLP Suite help highlight general characteristics of an author's style?	The NLP Suite provides a wide range of options to find answers to these questions. *Under CORPUS/DOCUMENT analysis tools select:* • Style analysis (ALL options GUI)
Are there words and combinations of words that appear more frequently in texts? How do they different across different authors? N-grams, Co-occurrences, *N-grams/co-occurrences viewer*	*Under CORPUS/DOCUMENT analysis tools select:* • N-grams (word & character)

Computationally, these objects are known as n-grams: unigram for single words, bigrams for combinations of two words, trigrams for three words, etc. (although it generally makes little sense to go beyond trigrams).	• N-grams/co-occurrences viewer
Is there a consistent use of passive and active verb forms, of present and past verb tenses, of some subjects always portrayed in active or passive forms, acknowledging or denying agency? *Verb modality, tense, and voice* *Nominalization*	*Under CORPUS/DOCUMENT analysis tools select:* • CoNLL table analyzer - Search the CoNLL table • Nominalization
How **readable is a text**? How **complex are the sentences**? *Sentence complexity* *Text readability*	*Under CORPUS/DOCUMENT analysis tools select:* • Sentence/text readability (via textstat) • Sentence complexity
How is the text vocabulary? Are there **unusual words**, **misspelled words**? Are there clear patterns in the **first and last few sentences** of documents? More nouns, more verbs… Are there **repetitions**? How rich is an author's vocabulary? *Vocabulary* Unusual words/misspelled words Short/long words, vowel words Vocabulary richness Abstract/concrete vocabulary Objective/subjective writing In which language are texts written in?	*Under CORPUS/DOCUMENT analysis tools select:* • Style analysis (ALL options GUI) Then tick the Vocabulary analysis option and explore the many options available.
Who wrote the text? Man or woman?	*Under CORPUS/DOCUMENT analysis tools select:* • Who wrote the text? Man or woman? (via Gender guesser)
Is there *dialogue in the text*? And who does the talking? *Is there dialogue in the text?*	Stanford CoreNLP, uniquely, has a special quote annotator

	that provide answers to these questions. *Under CORPUS/DOCUMENT analysis tools select:* • Parsers & annotators (BERT, CoreNLP, spaCy, Stanza)
Visualizing words	
How can I get those pretty pictures of words in different colors and sizes? *Wordclouds* 	This type of image is called a wordcloud, easily done in the NLP Suite and with a wide range of options. *Under CORPUS/DOCUMENT analysis tools select:* • Wordclouds (ALL options GUI)
How can I visualize relations between people or words? *Sankey plot* 	This type of plot is called a Sankey plot. *Under Visualization tools select:* • Sankey flowchart (Plotly) (Open GUI)
How can I visualize relations between people or words (e.g., Subject-Verb-Object)?	This type of plot is called a network graph and you can display one easily in the NLP Suite (and some NLP tools, for instance SVO, export a network graph automatically). *Under Visualization tools select:*

Network graphs 	• Network graphs (Gephi) (Open GUI)
Is it possible to visualize gendered names (male or female) in different colors to have an immediate view of where male/female names occur in the text? *HTML files* 	*Under Visualization tools select:* • HTML annotator - dictionary, gender, DBpedia, YAGO, WordNet - (All options GUI)
Is there a way to visualize locations mentioned in texts as geographic maps? *Pin maps heat maps* 	*Under CORPUS/DOCUMENT analysis tools select:* • Geographic maps: From texts to maps You can also map locations from a list of locations in a csv file: • Geographic maps: Google Earth Pro You will find the same GIS tools under *Visualization tools.*
I have seen presentations with fancy charts like this that move around when you click different parts of the chart. Can it be done easily in the NLP Suite? *Sunburst pie charts*	This type of chart is called sunburst pie chart. It is a powerful interactive visualization tool that you can easily use in the NLP Suite.

	*Under Visualization tools select:* • Sunburst pie chart (Plotly) (Open GUI)
How can I produce a map of this kind (from a comparative analysis of folktales across 6 different countries)? *Treemap* 	This type of map is called a treemap and you can easily create one in the NLP Suite. *Under Visualization tools select:* • Treemap (Plotly) (Open GUI)
Animated time plot	*Under Visualization tools.* • Animated time-dependent bar plot (Plotly) (Open GUI)
 *Box plots*	*Under Visualization tools select.* Box plots (Open GUI)
Can this type of chart be easily produced to visualize the frequency distribution of certain aspects of texts (e.g., verbs like in this case)? *Bar charts, pie charts, line charts*	Every tool in the NLP Suite routinely produces in output a range of plots in either Excel or Plotly, as bar charts or pie charts. You can also access the visualization tools *Under Visualization tools select:*

	• Excel & Plotly charts