Automatic Extraction and Visualization of Subject-Verb-Object (SVO) Triplets

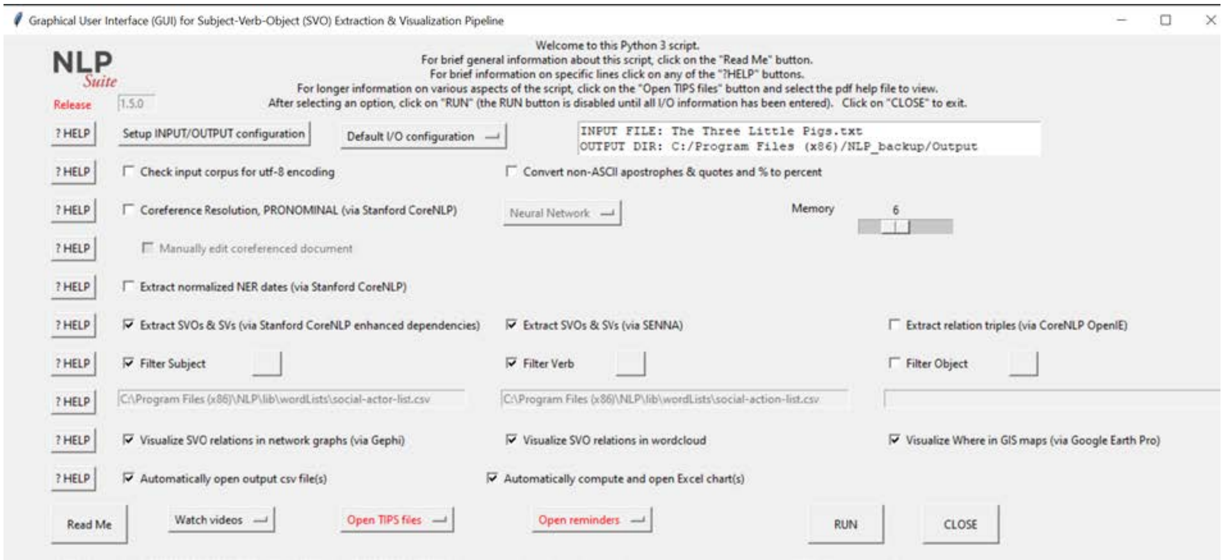
The 5 Ws of narrative: Who, What, Where, When, and Why

The SVO extraction and visualization pipeline is fundamentally about extraction at least some of the 5Ws of narrative: Who, What, Where, When. The SVO are about **Who-What-Whom**, visualized in network graphs (via Gephi). The **Where** is about the locations mentioned in the text and visualized as points on a map (via Google Earth Pro). The **When** is about the times and dates mentioned in the text and visualized as Excel charts.

As for the **Why**… it is beyond the scope of the current pipeline or, more generally, of automatic computational tools.

A convenient GUI for SVO extraction and visualization

The front-end Python 3 scripts provide a convenient and easy-to-use Graphical User Interface (GUI) where the steps required to extract and visualize SVO triplets are clearly laid out.



Graphical User Interface (GUI) for Subject-Verb-Object (SVO) Extraction & Visualization Pipeline

Welcome to this Python 3 script.
For brief general information about this script, click on the "Read Me" button.
For brief information on specific lines click on any of the "?HELP" buttons.
For longer information on various aspects of the script, click on the "Open TIPS files" button and select the pdf help file to view.
After selecting an option, click on "RUN" (the RUN button is disabled until all I/O information has been entered). Click on "CLOSE" to exit.

Release 1.5.0

```
INPUT FILE: The Three Little Pigs.txt
OUTPUT DIR: C:/Program Files (x86)/NLP_backup/Output
```

□ Check input corpus for utf-8 encoding
□ Convert non-ASCII apostrophes & quotes and % to percent
□ Coreference Resolution, PRONOMINAL (via Stanford CoreNLP) — Neural Network — Memory 6
□ Manually edit coreferenced document
□ Extract normalized NER dates (via Stanford CoreNLP)
☑ Extract SVOs & SVs (via Stanford CoreNLP enhanced dependencies)
☑ Extract SVOs & SVs (via SENNA)
□ Extract relation triples (via CoreNLP OpenIE)
☑ Filter Subject — C:\Program Files (x86)\NLP\lib\wordLists\social-actor-list.csv
☑ Filter Verb — C:\Program Files (x86)\NLP\lib\wordLists\social-action-list.csv
□ Filter Object
☑ Visualize SVO relations in network graphs (via Gephi)
☑ Visualize SVO relations in wordcloud
☑ Visualize Where in GIS maps (via Google Earth Pro)
☑ Automatically open output csv file(s)
☑ Automatically compute and open Excel chart(s)

The GUI offers a pipeline for extracting and visualizing SVO triplets conveniently stringing together of a set of Python scripts.

utf-8 compliance

Most CoreNLP algorithms require in input text files that are utf-8 compliant, i.e., texts encoded in utf-8 format (on text encoding, see TIPS_NLP_Text encoding.pdf). Make sure to test your input file(s).

Stanford CoreNLP coreference resolution

SVO triplets can represent meaningful sentences in a simplified 3-term structure. In order to improve the quality of SVOs, users can choose to run the pronominal coreference resolution, based on Stanford CoreNLP, before SVO extraction. Basically, coreference would replace pronouns such as *he*, *she*, *they*, *us*, … to their respective nouns. “President Obama said that he would…” would be *resolved* as “President Obama said that Obama would…”. To learn more about coreference resolution read TIPS_NLP_Stanford CoreNLP coreference resolution.pdf

Dealing with an imperfect tool: Manual coreference

Automatic coreference is far from a perfect tool with only about 65% accuracy. You can always manually edit the coreference text. **Unfortunately, this option requires a large memory since it brings in memory both files, original and coreferenced side-by-side. A memory hog! You may not have enough memory on your computer for this option.**

Stanford CoreNLP date extractor

The CoreNLP date extractor extracts any temporal expression mentioned in the text in standardized, uniform ways (“normalized dates”) that can be compared, sorted, and analyzed to bring out a text’s narrative strategy (to learn more about CoreNLP date extractor, normalized dates, and story vs plot narrative strategies, read TIPS_NLP_Stanford CoreNLP date extractor.pdf).

The SVO extractors

At the heart of the pipeline are three different approaches to SVO extraction:

1. Stanford CoreNLP enhanced dependencies parser
2. SENNA
3. Stanford CoreNLP OpenIE (Open Information Extraction)

Stanford CoreNLP Enhanced Dependencies Parser

Please, see the TIPS file **TIPS_NLP_SVO Stanford CoreNLP enhanced dependencies parser.pdf**.

SENNA (Semantic/syntactic Extraction using a Neural Network Architecture)

Please, see the TIPS file **TIPS_NLP_SVO SENNA.pdf**.

Stanford CoreNLP OpenIE (Open Information Extraction)

Please, see the TIPS file **TIPS_NLP_Stanford CoreNLP OpenIE.pdf**.

Filtering SVOs via dictionary files

The SVO extractors produce in output triplets of various types. As a social scientist you may be interested in Who Does What, in social actors and social action. No point, perhaps, knowing that a door is blue or made of wood. The SVO GUI provides a way for you to filter each of the three elements of the SVO triplet, keeping only those values for each contained in a supplied dictionary file.

Where do these dictionary files come from?

The NLP Suite comes with three default dictionary files in csv format and stored in the lib subdirectory: the social-actor-list.csv and social-action-list.csv. These files have been constructed using the WordNet tool (Zoom IN/DOWN option with ‘Person’ as the keyword (synset) starting point. These produce exhausting lists of over 18,000 social actors and 10,000 social actions. For restricted domains these lists may provide both too few and too many entries. For instance, suppose you are working with folk tales. You will find there many talking foxes, flying horses, and many other anthropomorphized animals (e.g., *The Three Little Pigs* English folk tale in our default sample texts). None of the animals will be in the default social-actor-list.csv. You may wish to build your own list.

How do you build your own filter dictionary lists?

You can build your lists, based on your specific corpus, by running the Stanford CoreNLP to produce the CoNLL table, run an SQL query to obtain a frequency distribution of all tokens (found in the field FORM) that have specific POSTAG labels (e.g., nsubj (nominal subject) or specific NER values (e.g., PERSON), then manually inspect the query results keeping social actors only. You can also use the results of these queries in conjunction with the WordNet tool to see which nouns are animals, for instance.

Input

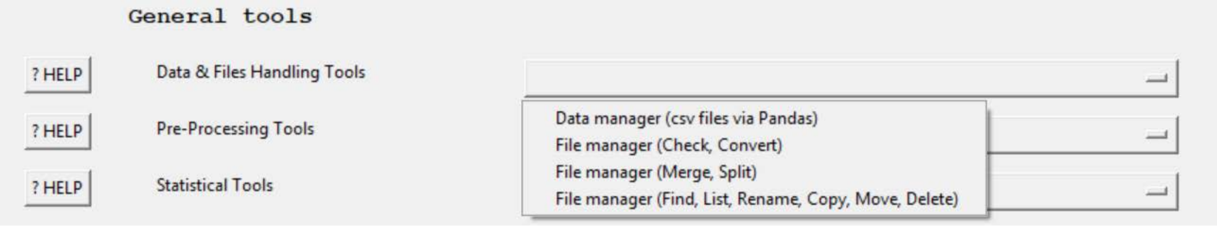
In INPUT the routine expects a text file or a set of text files in an input directory.

Using a merged file in input

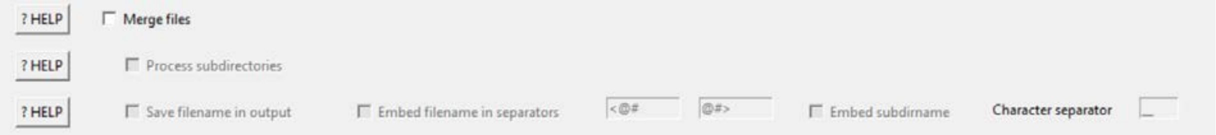
Another analysis strategy for many files, instead of the input directory option where all the text files present in the directory are processed one at a time, is to merge the files into a single file and then process the merged file as a single file.

How do you merge files?

When you run in command line python NLP_main.py, under General tools, Data & File Handling Tools, select the option File manager (Merge, Split) to open the File manager GUI.


General tools
? HELP
Data & Files Handling Tools
? HELP
Pre-Processing Tools
? HELP
Statistical Tools
Data manager (csv files via Pandas)
File manager (Check, Convert)
File manager (Merge, Split)
File manager (Find, List, Rename, Copy, Move, Delete)

You can also run directly in command line python file_manager_merger_splitter_main.py to open that same GUI.



The *Merge files* widget, when ticked, provides many options.

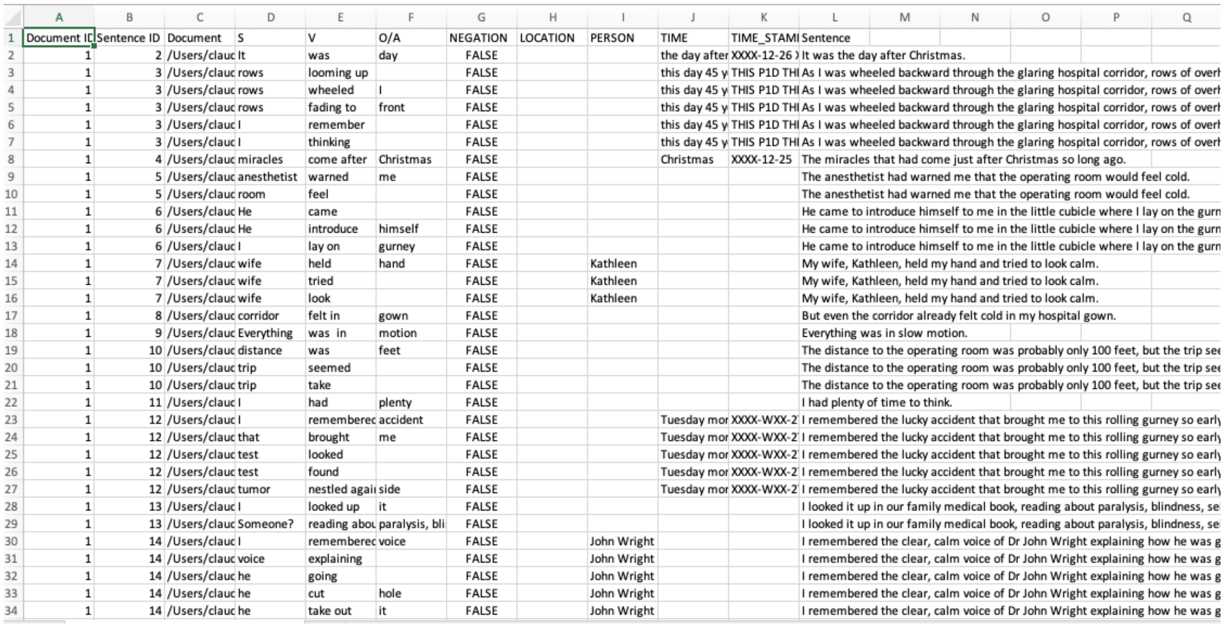
Please, read the TIPS file TIPS_NLP_Files merger.pdf to learn more about this important topic.

Output

In output, the SVO script produces several different files, a csv file, a gexf Gephi file, a kml Google Earth Pro file.

csv files

The main csv file (characterized by the substring _svo in the filename) contains the basic information produced by the SVO extract scripts, perhaps with filtered values. All three SVO extractor scrits produce very similarly formatted output. Here is what that file looks like.



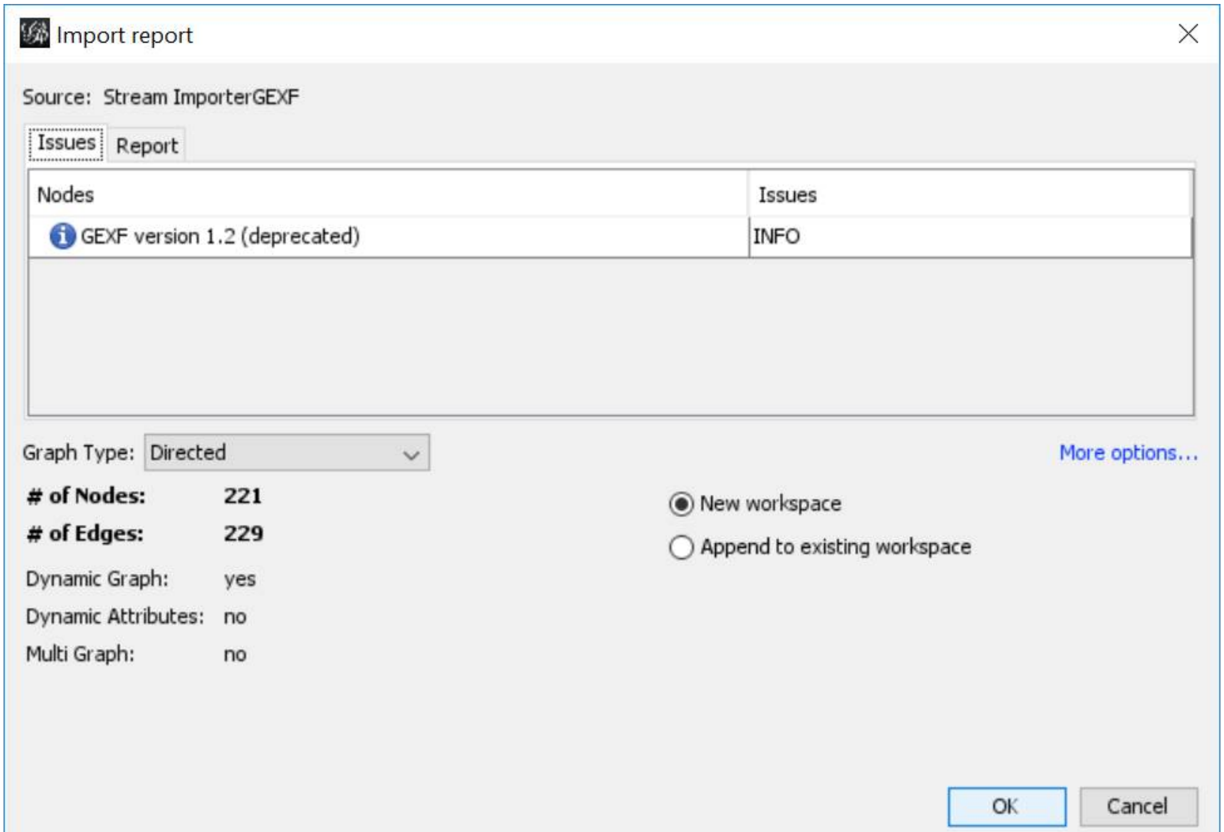
	A	B	C	D	E	F	G	H	I	J	K	L
1	Document ID	Sentence ID	Document	S	V	O/A	NEGATION	LOCATION	PERSON	TIME	TIME_STAMI	Sentence
2	1	2	/Users/clauc	It	was	day	FALSE			the day after	XXXX-12-26)	It was the day after Christmas.
3	1	3	/Users/clauc	rows	looming up		FALSE			this day 45 y	THIS P1D THI	As I was wheeled backward through the glaring hospital corridor, rows of overl
4	1	3	/Users/clauc	rows	wheeled	I	FALSE			this day 45 y	THIS P1D THI	As I was wheeled backward through the glaring hospital corridor, rows of overl
5	1	3	/Users/clauc	rows	fading to	front	FALSE			this day 45 y	THIS P1D THI	As I was wheeled backward through the glaring hospital corridor, rows of overl
6	1	3	/Users/clauc	I	remember		FALSE			this day 45 y	THIS P1D THI	As I was wheeled backward through the glaring hospital corridor, rows of overl
7	1	3	/Users/clauc	I	thinking		FALSE			this day 45 y	THIS P1D THI	As I was wheeled backward through the glaring hospital corridor, rows of overl
8	1	4	/Users/clauc	miracles	come after	Christmas	FALSE			Christmas	XXXX-12-25	The miracles that had come just after Christmas so long ago.
9	1	5	/Users/clauc	anesthetist	warned	me	FALSE					The anesthetist had warned me that the operating room would feel cold.
10	1	5	/Users/clauc	room	feel		FALSE					The anesthetist had warned me that the operating room would feel cold.
11	1	6	/Users/clauc	He	came		FALSE					He came to introduce himself to me in the little cubicle where I lay on the gurn
12	1	6	/Users/clauc	He	introduce	himself	FALSE					He came to introduce himself to me in the little cubicle where I lay on the gurn
13	1	6	/Users/clauc	I	lay on	gurney	FALSE					He came to introduce himself to me in the little cubicle where I lay on the gurn
14	1	7	/Users/clauc	wife	held	hand	FALSE		Kathleen			My wife, Kathleen, held my hand and tried to look calm.
15	1	7	/Users/clauc	wife	tried		FALSE		Kathleen			My wife, Kathleen, held my hand and tried to look calm.
16	1	7	/Users/clauc	wife	look		FALSE		Kathleen			My wife, Kathleen, held my hand and tried to look calm.
17	1	8	/Users/clauc	corridor	felt in	gown	FALSE					But even the corridor already felt cold in my hospital gown.
18	1	9	/Users/clauc	Everything	was in	motion	FALSE					Everything was in slow motion.
19	1	10	/Users/clauc	distance	was	feet	FALSE					The distance to the operating room was probably only 100 feet, but the trip see
20	1	10	/Users/clauc	trip	seemed		FALSE					The distance to the operating room was probably only 100 feet, but the trip see
21	1	10	/Users/clauc	trip	take		FALSE					The distance to the operating room was probably only 100 feet, but the trip see
22	1	11	/Users/clauc	I	had	plenty	FALSE					I had plenty of time to think.
23	1	12	/Users/clauc	I	remembered	accident	FALSE			Tuesday mor	XXXX-WXX-2	I remembered the lucky accident that brought me to this rolling gurney so early
24	1	12	/Users/clauc	that	brought	me	FALSE			Tuesday mor	XXXX-WXX-2	I remembered the lucky accident that brought me to this rolling gurney so early
25	1	12	/Users/clauc	test	looked		FALSE			Tuesday mor	XXXX-WXX-2	I remembered the lucky accident that brought me to this rolling gurney so early
26	1	12	/Users/clauc	test	found		FALSE			Tuesday mor	XXXX-WXX-2	I remembered the lucky accident that brought me to this rolling gurney so early
27	1	12	/Users/clauc	tumor	nestled agai	side	FALSE			Tuesday mor	XXXX-WXX-2	I remembered the lucky accident that brought me to this rolling gurney so early
28	1	13	/Users/clauc	I	looked up	it	FALSE					I looked it up in our family medical book, reading about paralysis, blindness, se
29	1	13	/Users/clauc	Someone?	reading abou	paralysis, bli	FALSE					I looked it up in our family medical book, reading about paralysis, blindness, se
30	1	14	/Users/clauc	I	remembered	voice	FALSE		John Wright			I remembered the clear, calm voice of Dr John Wright explaining how he was g
31	1	14	/Users/clauc	voice	explaining		FALSE		John Wright			I remembered the clear, calm voice of Dr John Wright explaining how he was g
32	1	14	/Users/clauc	he	going		FALSE		John Wright			I remembered the clear, calm voice of Dr John Wright explaining how he was g
33	1	14	/Users/clauc	he	cut	hole	FALSE		John Wright			I remembered the clear, calm voice of Dr John Wright explaining how he was g
34	1	14	/Users/clauc	he	take out	it	FALSE		John Wright			I remembered the clear, calm voice of Dr John Wright explaining how he was g

Other csv files provide various information produced, for instance, by the GIS scripts.

SVO relations: Network graphs in Gephi

Relations between S and O via V can be visualized as network graphs using the freeware Gephi tool. The NLP Suite wrapper for Gephi will automatically construct a dynamic network graph, i.e., a graph that changes overtime and where the sentence index of each SVO is taken as proxy for time. This way, you can see the movement of SVOs across the document.

This is what Gephi looks like when it opens up on the SVO output generated using one of the default documents, Murphy's *Six miracles* story.



Import report

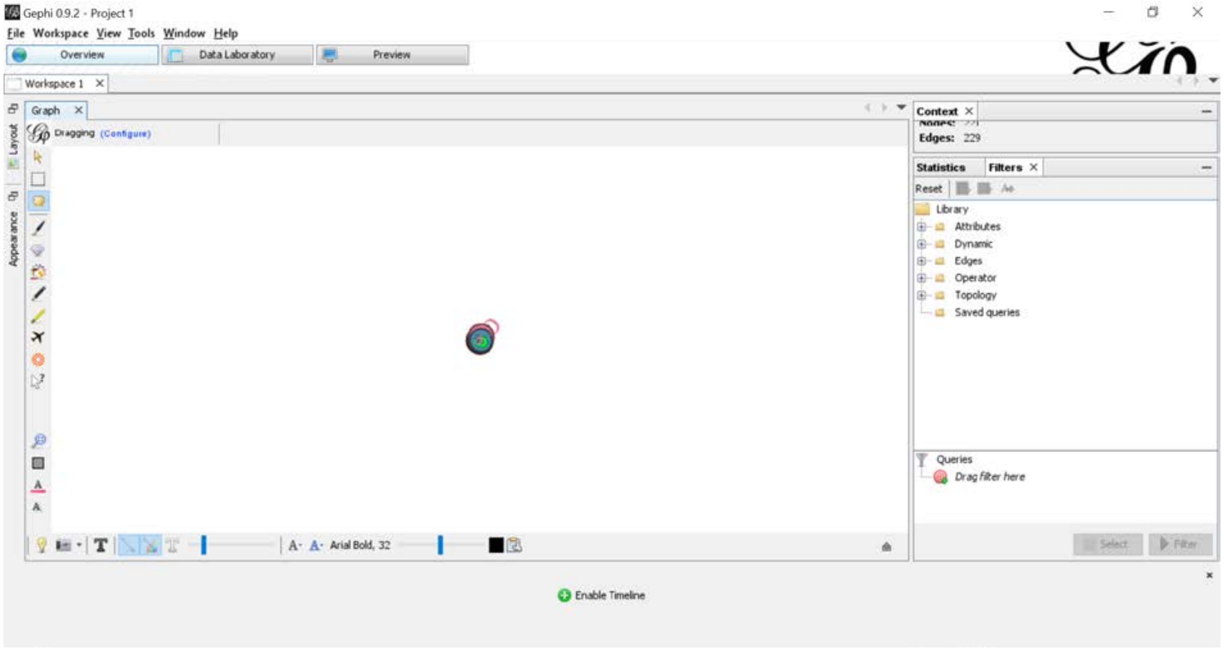
Source: Stream ImporterGEXF

Nodes	Issues
GEXF version 1.2 (deprecated)	INFO

Graph Type: Directed

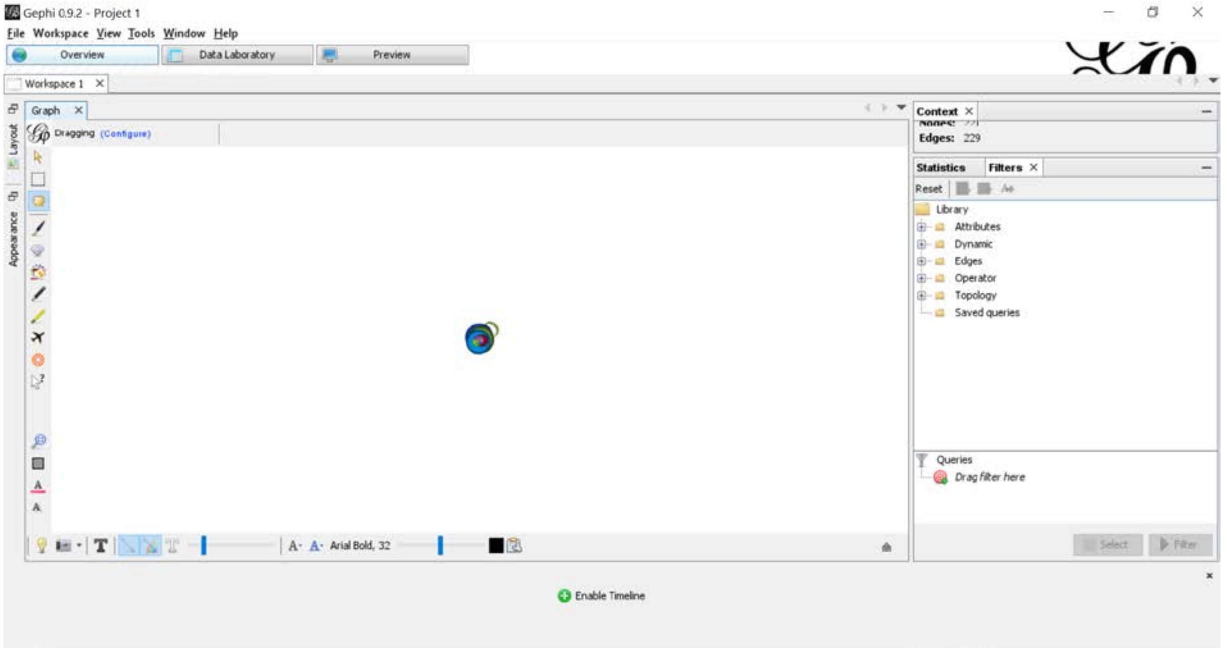
# of Nodes: 221
# of Edges: 229
Dynamic Graph: yes
Dynamic Attributes: no
Multi Graph: no

● New workspace
○ Append to existing workspace



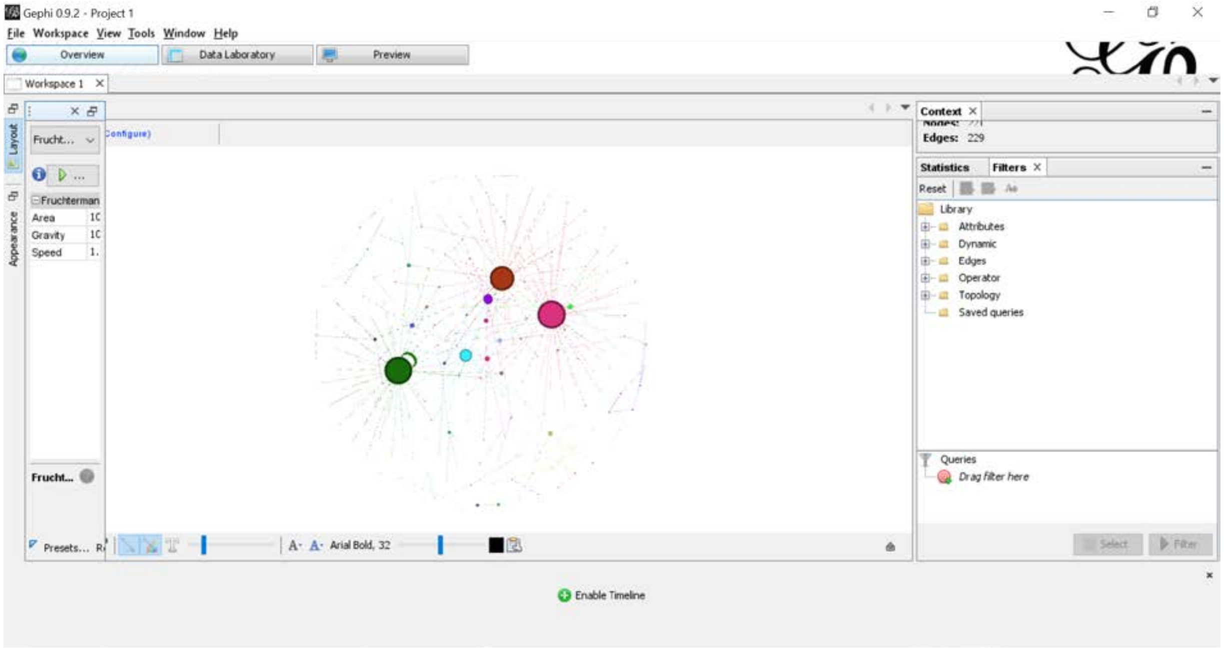
Edges: 229

You click OK and you get this



Edges: 229

What is that blob? Not a helpful visual! Don’t be turned off. Easy to bring to life this blob, including visualizing the network as a dynamic graph by enabling the timeline… please, read all the TIPS files on how to get going with Gephi.



Area 1C
Gravity 1C
Speed 1.

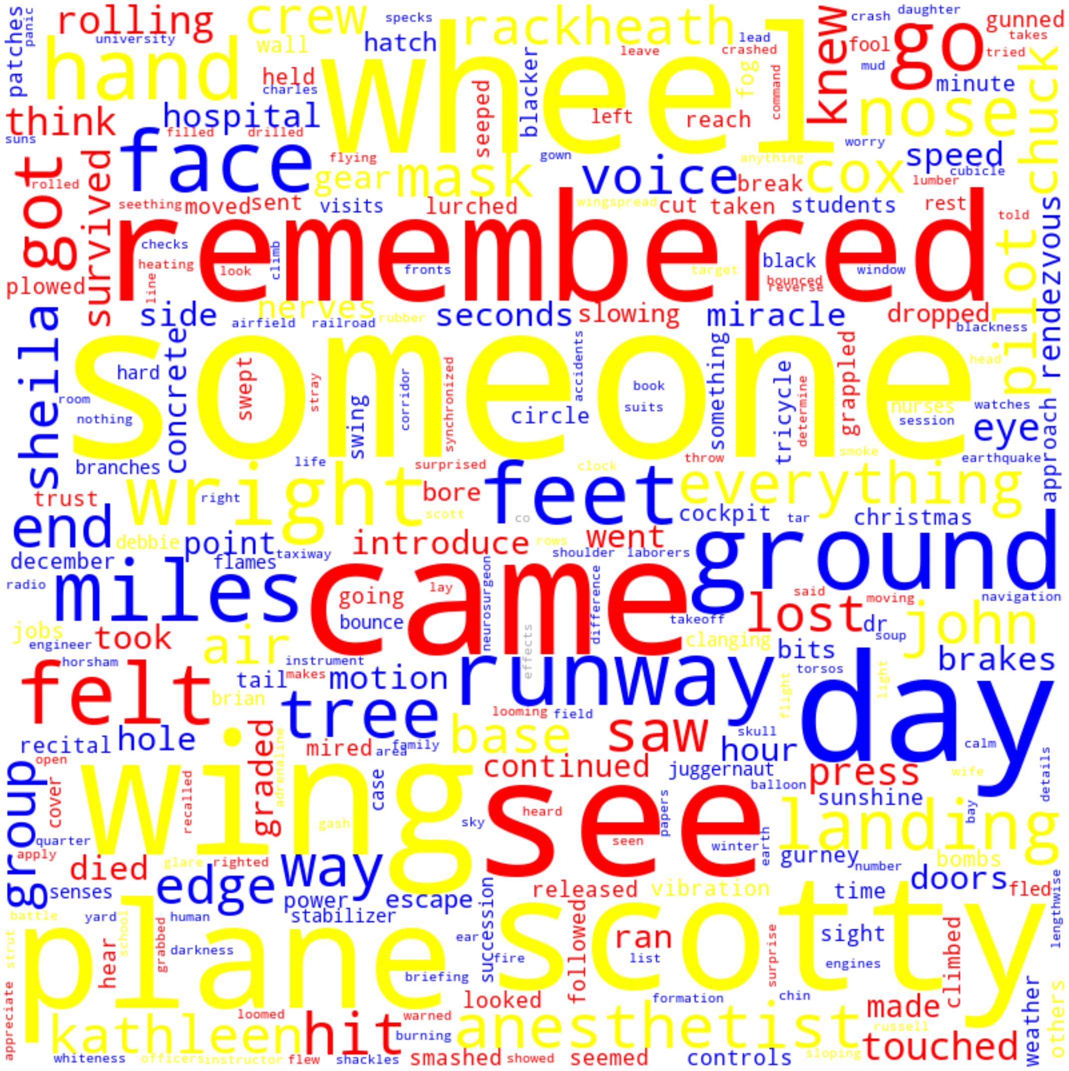
Nodes: 221
Edges: 229

SVO relations: A wordcloud of Subjects, Verbs, and Objects

Another way to visualize relations between SVOs is to map S, V, and O values in different colors in a wordcloud.

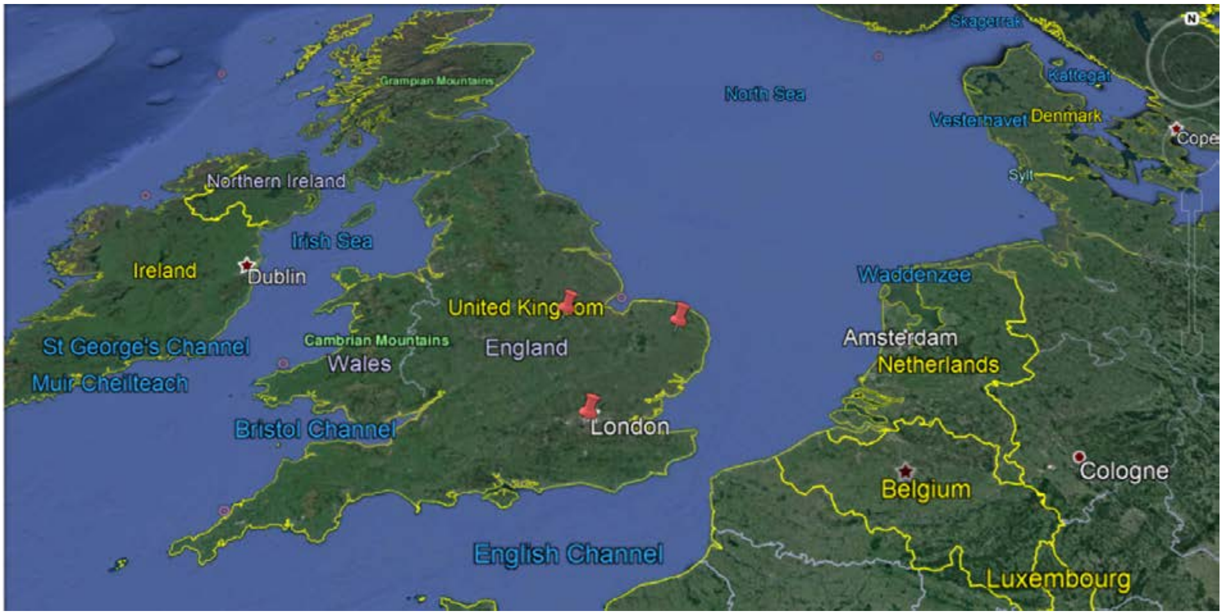
Word size and color

The combination of word sizes and colors may hopefully give you a visual sense of what is going on, where **word size** is related to word frequency; and the **colors** indicate S, or V, or O: yellow for S, red for V, blue for O.



Mapping the WHERE: GIS maps in Google Earth Pro

When the Visualize Where option is ticked, the SVO script will automatically geocode any location found in the text and map in using Google Earth Pro. When running SVO using one of the default documents, Murphy’s *Six miracles* story, this is what the map will look like.



Excel charts

The SVO pipeline also produces several kinds of Excel bar and line charts to help the user to make sense of the data.

References

TIPS_NLP_Java download install run.pdf
TIPS_NLP_Stanford CoreNLP parser.pdf
TIPS_NLP_Stanford CoNLL table.pdf
TIPS_NLP_Text encoding.pdf
TIPS_NLP_Files merger.pdf
TIPS_NLP_Stanford CoreNLP coreference resolution.pdf
TIPS_NLP_Stanford CoreNLP date extractor.pdf
TIPS_NLP_SVO Stanford CoreNLP enhanced dependencies parser.pdf
TIPS_NLP_SVO SENNA.pdf
TIPS_NLP_Stanford CoreNLP OpenIE.pdf
TIPS_NLP_Geocoding.pdf
TIPS_NLP_Google Earth Pro.pdf
TIPS_NLP_Google Earth Pro From KML to Excel.pdf