WordNet

WordNet: What is it?

According to the WordNet website (https://WordNet.princeton.edu/):

> WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (**synsets**), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser. WordNet is also freely and publicly available for download. WordNet's structure makes it a useful tool for computational linguistics and natural language processing.

WordNet superficially resembles a **thesaurus**, in that it groups words together based on their meanings. However, there are some important distinctions. First, WordNet interlinks not just word forms—strings of letters—but specific senses of words. As a result, words that are found in close proximity to one another in the network are semantically disambiguated. Second, WordNet labels the semantic relations among words, whereas the groupings of words in a thesaurus does not follow any explicit pattern other than meaning similarity.

Single words & collocations

The WordNet databases comprises both single words or combinations of two or more words that typically come together with a specific meaning (*collocations*, e.g., coming out, shut down, thumbs up, stand in line, customs duty). Over 80% of terms in the WordNet database are collocations, at least at the time of Miller et al.'s *Introduction to WordNet* manual (1993, p. 2). For the English language (but WordNet is available for some 200 languages) the database contains a very large set of terms. The most **up-to-date numbers of terms** are given in https://WordNet.princeton.edu/documentation/wnstats7wn

Number of words, synsets, and senses

POS	Unique Strings	Synsets	Total Word-Sense Pairs
Noun	117798	82115	146312
Verb	11529	13767	25047
Adjective	21479	18156	30002
Adverb	4481	3621	5580
Totals	155287	117659	206941

Limitations

1. WordNet includes the lexical categories nouns, verbs, adjectives and adverbs but **WordNet ignores prepositions, determiners and other function words**.

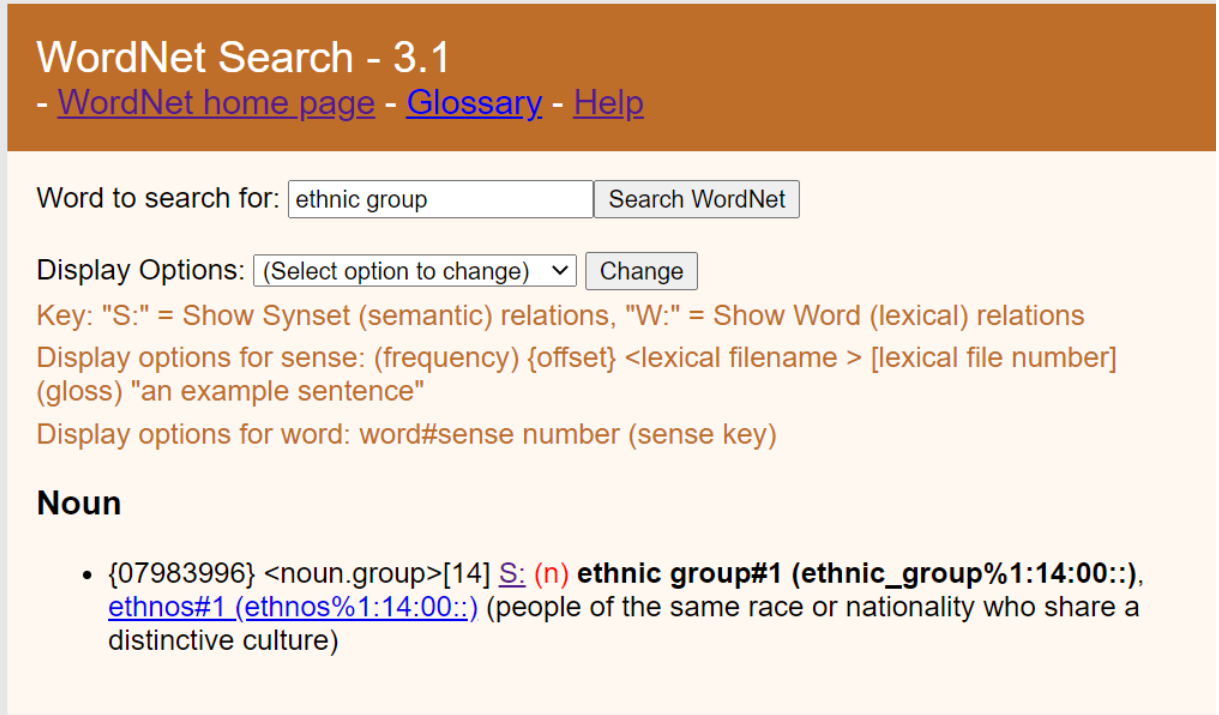
2. **WordNet is better suited to account for concrete concepts than for abstract concepts.** It is much easier to create hyponyms/hypernym relationships between "conifer" as a type of "tree", a "tree" as a type of "plant", and a "plant" as a type of "organism". Not so easy to classify emotions like "fear" or "happiness" into

hyponyms/hypernym relationships.

3. **The NLP Suite implementation of WordNet does not include adjectives and adverbs.**

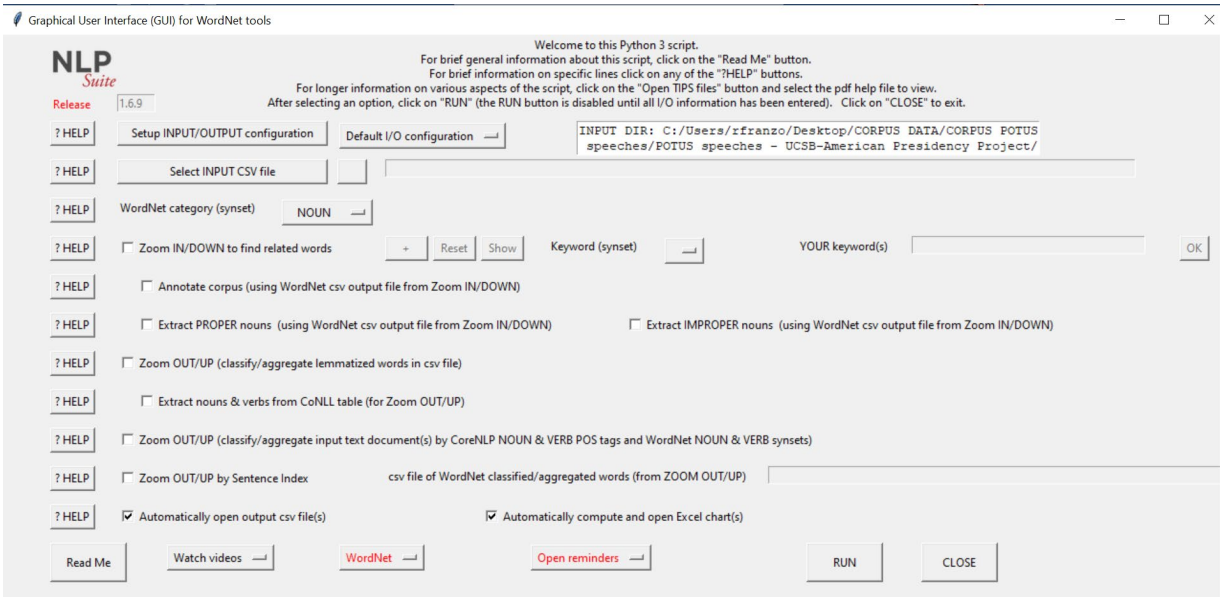
WordNet API

http://WordNetweb.princeton.edu/perl/webwn?s=&sub=Search+WordNet&o2=1&o0=1&o8=1&o1=1&o7=1&o5=1&o9=&o6=1&o3=1&o4=1&h=0



A WordNet GUI

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.



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

Once active, the WordNet GUI provides several options.

WordNet: What can you use it for?

There are two main functions you can perform with the NLP Suite WordNet script: disaggregate and aggregate.

1. **Disaggregate.** You can get listings of individual entries of larger aggregate categories or synsets. Searching WordNet starting with the synset 'person' will export a list of some 19,000 proper and improper names of individuals, groups, and organizations. **You will find that list in the lib subdirectory of the NLP Suite with the name social-actor-list.csv.**
2. **Aggregate.** Conversely, given a list of words, nouns or verbs, you can find out how WordNet would aggregate them in larger categories (e.g., run, walk, amble classified as verbs of *motion*).

The MIT engine behind the GUI

We developed a Java tool based on MIT JWI (Java WordNet Interface) - https://projects.csail.mit.edu/jwi/. We imported MIT JWI and implemented a depth-first search (DFS) in Java to construct comprehensive lists (e.g., of social actors). The ZOOM IN/DOWN and ZOOM OUT/UP connect to the MIT Java script by allowing users to search the WordNet database either by top synsets (25 Nouns and 15 Verbs) or by one or more words (or collocations, i.e., combinations of multiple words[i]) of their choice. If the user selects a top synset, the Java algorithm uses the MIT JWI library to go through every word in the WordNet database and filter out all the words that are under a chosen category. Alternatively, if the user inputs several search words, the Java algorithm uses DFS to recursively find all hyponyms of the search words. In DFS, we first push into a stack all source words (i.e., we insert each source word at the top of the stack). We then pop a word for searching (i.e., we remove the word from the top of the stack) and add all its hyponyms into the stack. Since a word can have

multiple meanings in the WordNet database, including quite rare meanings, we include only the 3 most frequent meanings and the hyponyms under these frequent meanings. We repeat this process until there are no more words in the stack, i.e., we have searched all possible hyponyms. Thus, a search based on the top noun synset 'person' will export a list of some 19,000 proper and improper names of individuals, groups, and organizations. Searching for 'animal' will lead to a similar list of over 14,000 entries. Since proper nouns have an upper-case first letter and improper nouns have a lower-case first letter, we can use these features to export proper nouns or improper nouns selectively.

GUI WordNet options

Let's take a look in detail how these options can be implemented by focusing on the various GUI widgets.

WordNet category (synset): Nouns & Verbs

The NLP Suite implementation of WordNet focuses on two of the four synset types (Noun and Verb, at the exclusion of adjectives and adverbs).

You will need to select the synset type to work with.

There are 25 TOP NOUN synsets in WordNet and 15 TOP VERB synsets.

The 25 top noun synsets are: **act, animal, artifact, attribute, body, cognition, communication, event, feeling, food, group, location, motive, object, person, phenomenon, plant, possession, process, quantity, relation, shape, state, substance, time**.

The 15 top verb synsets are: **body, change, cognition, communication, competition, consumption, contact, creation, emotion, motion, perception, possession, social, stative, weather.**

Both IN/DOWN and OUT/UP scripts are written in Java and rely on the MIT JWI (Java WordNet Interface) to interface with WordNet (https://projects.csail.mit.edu/jwi/).

Zoom IN/DOWN option: Disaggregate

This option allows you to find **disaggregated lists of values**, like using a microscope to zoom in on something.

Keyword (synset)

The script uses the WordNet lexicon database to provide a list of terms associated to a starting keyword (synset) in a lexical hierarchy. From that starting point, the algorithm will navigate down into all the subcategories of a synset.

WordNet uses hyponymy and meronymy to go DOWN the hierarchy to find items in categories.

Hyponym

Hyponym is the specific term used to designate a member of a class. X is a hyponym of Y if X is a (kind of) Y.

Meronym

Meronym is the name of a constituent part of, the substance of, or a member of something. X is a meronym of Y if X is a part of Y.

Using the dropdown menu of the set of WordNet synsets for nouns and verbs, you will need to select the starting keyword(s) (synsets) that the script will use to traverse the database in order to provide the list. **Multiple starting synsets are allowed.**

Thus, you can construct a list of **social actors** (i.e., human characters, groups, or organizations) by selecting **person** as starting point. However, if your research deals with **fairy tales** or **folk tales** (e.g., the story of *The Three Little Pigs*), animals may also be characters (e.g., a talking fox), so the starting keyword can be **animal**, with both **person** and **animal** as your combined keywords.

+ RESET SHOW

You will need to press the **+ button** for multiple selections and press the **RESET** button (or ESCape) to delete all values entered and start fresh. The button **SHOW** to display all selected values.

YOUR keyword(s)

You can also enter one or more, comma separated, terms into the **YOUR keyword(s)** field, ignoring the pre-selected keywords. This option is particularly helpful if you want to restrict your search at a lower level, e.g. **ethnic group** instead of **person** to obtain a much shorter list of terms. Press **OK** when finished entering YOUR own values.

Input

No input. All that is required for this option is the starting keywords that you will have selected or entered.

Output

In OUTPUT the script will create 2 csv files, one marked as *verbose*, containing five columns: a list of terms found (column 1), the selected WordNet category (column 2), definitions of the category (column 3), frequency of senses of lemma that are ranked according to their frequency of occurrence in semantic concordance texts (column 4), examples of use (column 5). Each filename contains the synset used for searching the database, e.g., NLP_WordNet_DOWN_verb.social-list.csv when the verb synset *social* is used.

	Term	WordNet Category	Synset Definition	Frequency	Synset Example
2	abdicate	social	give up, such as power, as of monarchs and emperors, or duties and obli	1	"The King abdicated when he married a divorcee"
3	abet	social	assist or encourage, usually in some wrongdoing	2	assist or encourage, usually in some wrongdoing
4	abide_by	social	act in accordance with someone's rules, commands, or wishes	2	"He complied with my instructions"; "You must comply or else!"; "Follow these simple r
5	abolish	social	do away with	2	"Slavery was abolished in the mid-19th century in America and in Russia"
6	abrogate	social	revoke formally	2	revoke formally
7	abstain	social	refrain from voting	3	refrain from voting
8	abuse	social	treat badly	3	"This boss abuses his workers"; "She is always stepping on others to get ahead"
9	aby	social	make amends for	1	"expiate one's sins"
10	abye	social	make amends for	1	"expiate one's sins"
11	accede	social	take on duties or office	2	"accede to the throne"
12	accomplis	social	to gain with effort	3	"she achieved her goal despite setbacks"
13	accredit	social	grant credentials to	2	"The Regents officially recognized the new educational institution"; "recognize an acade
14	ace	social	succeed at easily	1	"She sailed through her exams"; "You will pass with flying colors"; "She nailed her astro
15	achieve	social	to gain with effort	2	"she achieved her goal despite setbacks"
16	acquit	social	behave in a certain manner	2	"She carried herself well"; "he bore himself with dignity"; "They conducted themselves
17	act	social	perform an action, or work out or perform (an action)	6	"think before you act"; "We must move quickly"; "The governor should act on the new e
18	act_on	social	carry further or advance	3	"Can you act on this matter soon?"

The other csv file contains a one-column list of distinct terms in the searched synset.

	Term
2	abdicate
3	abet
4	abide_by
5	abolish
6	abrogate
7	abstain
8	abuse
9	aby
10	abye
11	accede
12	accomplish
13	accredit
14	ace
15	achieve
16	acquit
17	act
18	act_on
19	act_superior
20	act_up
21	act_upon
22	action

Zoom OUT/UP option: Aggregate

The Zoom OUT/UP option takes the opposite direction: the script uses the WordNet lexicon database to aggregate a list of terms (NOUNS or VERBS) contained in an input csv file into top-level synsets (e.g., run, flee, walk, ... aggregated as verbs of motion).

The algorithm uses both ypernymy and holonymy to go UP the hierarchy.

Ypernymy

Hypernym is the generic term used to designate a whole class of specific instances. Y is a hypernym of X if X is a (kind of) Y.

Holonym

Holonym is the name of the whole of which the meronym names is a part. Y is a holonym of X if X is a part of Y.

Caveats

1. Unfortunately, there is no easy way to aggregate at levels lower than the top synsets. WordNet is a linked graph where each node is a synset and synsets are interlinked by means of conceptual-semantic and lexical relations. In other words, it is not a simple tree structure: there is no way to tell at which level the synset is located at. For example, the synset "anger" can be traced from top level synset "feeling" and follows the path: feeling -> emotion ->anger. But it can also be traced from top level synset "state" and follows the path: state -> condition -> physiological condition -> arousal -> emotional arousal -> anger. In the first case, "anger" is at level 3 (assuming "feeling" and or other top synsets are level 1). In the second case, "anger" is at level 6. Programmatically, if one gives users more freedom to control the level of aggregating up, it is hard to build a user-friendly communication protocol. If the user wants to aggregate up to level 3 (two levels below the top synset), then should "anger" be considered as a level 3 synset? Does the user want "anger" to be considered as a level 3 synset? Since there is no clear definition of how far away a synset is from the root (top synsets), our algorithm aggregates all the way up to root.
2. For VERBS, the "stative" category includes the auxiliary "be" probably making up the vast majority of stative verbs. Similarly, the category "possession" include the auxiliary "have" (and "get"). You may wish to exclude these auxiliary verbs from frequencies. The NLP Suite provides frequencies for WordNet verb aggregated with and without auxiliaries.

Input

In **INPUT**, the script expects a **one-column csv file** containing a list of NOUNS or VERBS to be aggregated. The question is: where do you get a list of nouns and words that you wish to aggregate? A good start is your CoNLL table, obtained from running the Stanford_CoreNLP_main.py script.

Nouns & verbs from the CoNLL table

NOUNS WOULD HAVE POSTAG VALUES NN* AND VERBS VB*. Below, is the list of verbs obtained this way from Murphy's story of *Miracles Thicker than Fog.*

1	Word
2	abort
3	act
4	add
5	ally
6	alternate
7	apply
8	appreciate
9	approach
10	approach
11	arrange
12	ask
13	ask
14	astonish
15	bear
16	become
17	begin
18	believe
19	believe
20	blanch
21	blow
22	bomb
23	bounce

What kind of verbs are they? Can we aggregate them into a smaller set of verb classes? Indeed. We can do that by using the Zoom UP/OUT option.

Output

In **OUTPUT** the script will create a **two-columns csv file** that contains the original nouns/verbs and their aggregate WordNet values, as shown in the figure below.

	Word	WordNet Category
1	abort	change
2	act	social
3	add	change
4	ally	social
5	alternate	change
6	apply	consumption
7	appreciate	emotion
8	approach	motion
9	approach	motion
10	arrange	contact
11	ask	communication
12	ask	communication
13	astonish	cognition
14	bear	stative
15	become	change
16	begin	change
17	believe	cognition
18	believe	cognition
19	blanch	body
20	blow	body
21	bomb	competition
22	bounce	motion

Limitations and further work

The DPpedia ontology classes (http://mappings.dbpedia.org/server/ontology/classes/) or Schema ontology classes (https://schema.org/docs/full.html) may provide ways to aggregate the data in a way that it is more meaningful for your work.

Verbs

Suppose you are working on a project of gay men personal stories. Many of the verbs may be verbs of violence, violence experienced by the narrators for their gender, verbs such as kick, punch, knife, push, slap. But, hopefully, more verbs will be verbs of caring such as hug, kiss, hold. All of these different types of verbs are classified in WordNet as **contact**. But you would probably want to distinguish between verbs of **violence** and verbs of **physical affection**. Similarly, verbs such as tell, say, mutter, write, telephone are all classified as **communication.** But your research needs may require a diction between **oral communication** and **written communication.**

Nouns

The same is true for nouns. If you are studying stories and you are interested in changing scene (time and space) and the characters who act in the different scenes, such nouns as school, shop, restaurant, church are all classified as **artifact;** but what you need is to recognize these artifacts as locations (home and park are classified as both).

Recommendation

So… do get a broad distribution of verbs and nouns by WordNet high-level synsets. But once you have your list, you may want to filter it further into lower-level categories better suited for your project. Unfortunately, you may have to do this work yourself.

So… use WordNet by all means to get a first distribution

Zoom OUT/UP by Sentence Index

The option allows you to plot the aggregate WordNet categories for nouns or verbs by sentence index to give you a more in-grained linguistic view of your text.

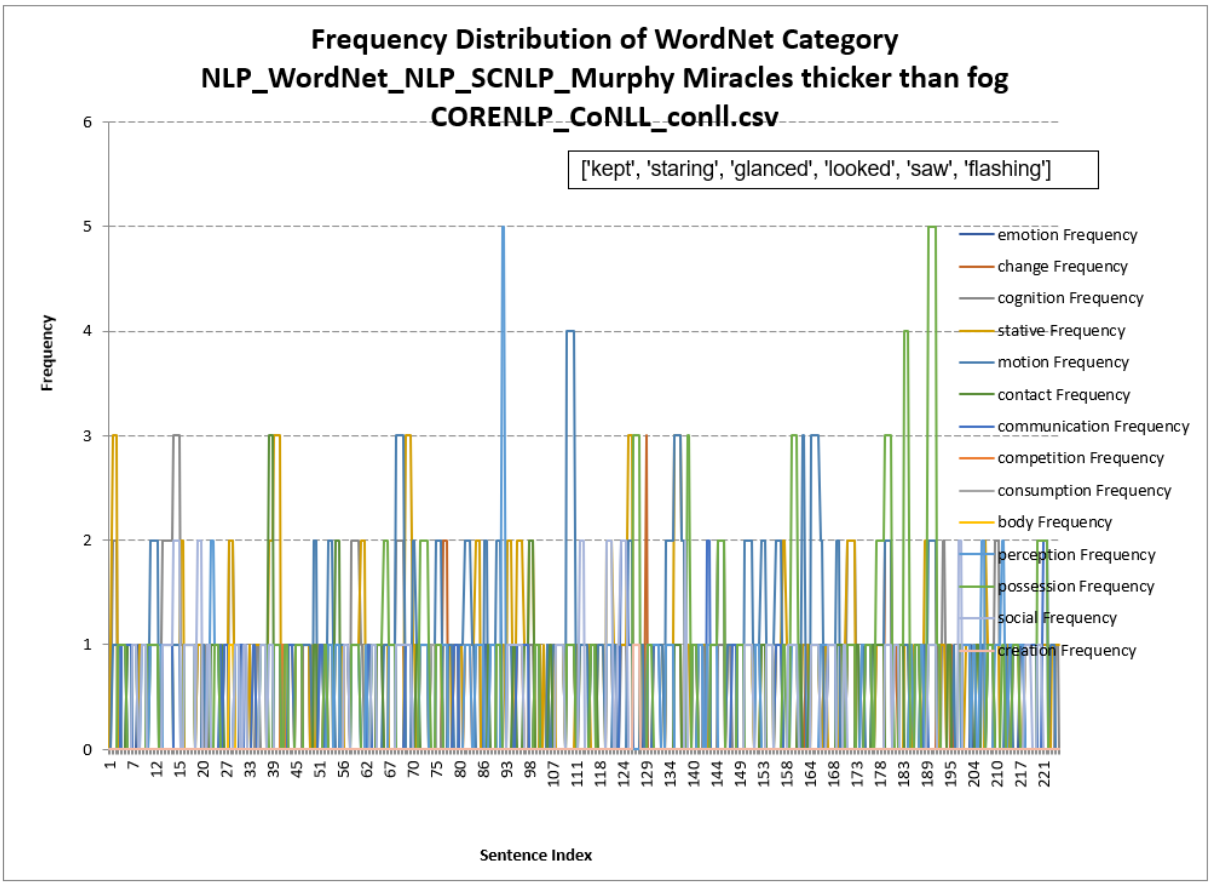
Input

In INPUT, the script expects 2 csv files:
1. a csv CoNLL file; you select this file in the IO widgets at the top of the GUI;
2. a csv dictionary file containing the WordNet classification of words, in lemma form, into higher-level aggregates; you will be prompted to select this csv file when you tick the checkbox.

The 2 csv files (CoNLL, WordNet dictionary) must have been computed from the same input txt file.

Output

In OUTPUT, the script produces a csv file and an Excel line plot of the aggregate WordNet categories by sentence index. Here is an example of the line plot based on Murphy's text.



Proper/improper nouns

WordNet will export values in lowercase or uppercase first letter depending upon they are proper or improper names. Thus, the list of some 19,000 terms you obtain from running the Zoom IN/DOWN option with ‘person’ as synset for nouns looks like this **(you will find that list in the lib subdirectory of the NLP Suite with the name social-actor-list.csv).**

1	Term
2	1st_Baron_Beaverbrook
3	1st_Baron_Verulam
4	1st_Earl_Attlee
5	1st_Earl_Baldwin_of_Bewdley
6	1st_Earl_of_Balfour
7	1st_lieutenant
8	1st_Viscount_Montgomery_of_Alamein
9	2nd_lieutenant
10	A._A._Michelson
11	A._A._Milne
12	A._Conan_Doyle
13	A._E._Burnside
14	A._E._Housman
15	A._E._Kennelly
16	A._E._W._Mason
17	A._Noam_Chomsky
18	A.E.
19	a_Kempis
20	Aalto
21	Aaron
22	Aaron_Burr
23	Aaron_Copland
24	Aaron_Montgomery_Ward
25	abandoned_infant
26	abandoned_person

The two checkboxes for proper and improper nouns allow you to filter out the two separately.

Input

In INPUT the function expects a csv file of NOUNs generated by the ZOOM IN/DOWN function (whether simple or verbose).

Output

In OUTPUT, the function saves a csv file with only either proper or improper nouns, as identified by a first letter upper/lower case.

The first column of the dictionary file, whether simple or verbose, will always be used for extracting values.

References

Princeton University “About WordNet.” WordNet. Princeton University. 2010.

Fellbaum, Christiane. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fellbaum, Christiane. 2005. "WordNet and WordNets." In: Brown, Keith et al. (eds.), *Encyclopedia of Language and Linguistics*, Second Edition, Oxford: Elsevier, 665-670.

Miller, George Armitage, Richard Beckwith, Christiane Fellbaum, Derek Gross, Katherine J. Miller. 1990. "Introduction to WordNet: An On-line Lexical Database." *International Journal of Lexicography*, Vol. 3, pp. 235-244. Revised August 1993.

Miller, George A. 1995. "WordNet: A Lexical Database for English." *Communications of the ACM* Vol. 38, No. 11: 39-41.

See also:

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. 2010. "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." In *Proceedings of the 7th Conference on Language Resources and Evaluation (LREC'10)*, Valletta, MT, 2010, pp. 2200–2204.

---

[i] The WordNet databases comprises both single words or combinations of two or more words that typically come together with a specific meaning (collocations, e.g., coming out, shut down, thumbs up, stand in line, customs duty). Over 80% of terms in the WordNet database are collocations, at least at the time of Miller et al.'s Introduction to WordNet manual (1993, p. 2). For the English language (but WordNet is available for some 200 languages) the database contains a very large set of terms. The most up-to-date numbers of terms are given in https://WordNet.princeton.edu/documentation/wnstats7wn.